

Language, Cognition, and Mind

Henk Zeevat
Hans-Christian Schmitz
Editors

Bayesian Natural Language Semantics and Pragmatics

 Springer

Language, Cognition, and Mind

Volume 2

Series editor

Chungmin Lee, Seoul National University, Seoul, South Korea

Editorial board members

Tecumseh Fitch, University of Vienna, Vienna, Austria

Peter Gaerdenfors, Lund University, Lund, Sweden

Bart Geurts, Radboud University, Nijmegen, The Netherlands

Noah D. Goodman, Stanford University, Stanford, USA

Robert Ladd, University of Edinburgh, Edinburgh, UK

Dan Lassiter, Stanford University, Stanford, USA

Edouard Machery, Pittsburgh University, Pittsburgh, USA

This series takes the current thinking on topics in linguistics from the theoretical level to validation through empirical and experimental research. The volumes published offer insights on research that combines linguistic perspectives from recently emerging experimental semantics and pragmatics as well as experimental syntax, phonology, and cross-linguistic psycholinguistics with cognitive science perspectives on linguistics, psychology, philosophy, artificial intelligence and neuroscience, and research into the mind, using all the various technical and critical methods available. The series also publishes cross-linguistic, cross-cultural studies that focus on finding variations and universals with cognitive validity. The peer reviewed edited volumes and monographs in this series inform the reader of the advances made through empirical and experimental research in the language-related cognitive science disciplines.

More information about this series at <http://www.springer.com/series/13376>

Henk Zeevat · Hans-Christian Schmitz
Editors

Bayesian Natural Language Semantics and Pragmatics

 Springer

Editors

Henk Zeevat
ILLC
University of Amsterdam
Amsterdam
The Netherlands

Hans-Christian Schmitz
Fraunhofer Institute for Communication
Information Processing and Ergonomics
FKIE
Wachtberg
Germany

ISSN 2364-4109

Language, Cognition, and Mind

ISBN 978-3-319-17063-3

DOI 10.1007/978-3-319-17064-0

ISSN 2364-4117 (electronic)

ISBN 978-3-319-17064-0 (eBook)

Library of Congress Control Number: 2015939421

Springer Cham Heidelberg New York Dordrecht London

© Springer International Publishing Switzerland 2015

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media
(www.springer.com)

Preface

Natural language interpretation (NLI) can be modelled analogously to Bayesian signal processing: the most probable message M (corresponding to the speaker's intention) conveyed by a signal S (a word, a sentence, turn or text) is found by two models, namely the prior probability of the message and the production probability of the signal. From these models and Bayes' theorem, the most probable message given the signal can be derived. Although the general capacity of Bayesian models has been proven in disciplines like artificial intelligence, cognitive science, computational linguistics and signal processing, they are not yet common in NLI.

Bayesian NLI gives three departures from standard assumptions. First, it can be seen as a defence of linguistic semantics as a production system that maps meanings into forms as was assumed in generative semantics, but also in systemic grammar, functional grammar and optimality theoretic syntax. This brings with it a more relaxed view of the relation between syntactic and semantic structures; the mapping from meanings to forms should be efficient (linear) and the prior strong enough to find the inversion from the cues in the utterance.

The second departure is that the prior is also the source for what is not said in the utterance but part of the pragmatic enrichment of the utterance: what is part of the speaker intention but not of the literal meaning. There is no principled difference between inferring in perception that the man who is running in the direction of the bus stop as the bus is approaching is trying to catch the bus and inferring in conversation that the man who states that he is out of petrol is asking for help with his problem.

The third departure is thus that interpretation is viewed as a stochastic and holistic process leading from stochastic data to a symbolic representation or a probability distribution over such representations that can be equated with the conversational contribution of the utterance.

Models relevant to the prior (the probability of the message M) include Bayesian networks for causality, association between concepts and (common ground) expectations. It is tempting to see a division in logic: classical logic for expressing the message, the logic of uncertainty for finding out what those messages are. Radical Bayesian interpretation can be described as the view that not just the

identification of the message requires Bayesian methods, but also the message itself and the contextual update have to be interpreted with reference to Bayesian belief revision, Bayesian networks or conceptual association.

The papers in this volume, which is one of the first on Bayesian NLI, approach the topic from diverse angles. The following gives some minimal guidance with respect to the content of the papers.

- Henk Zeevat: “[Perspectives on Bayesian Natural Language Semantics and Pragmatics](#)” Zeevat gives an overview of the different concepts of Bayesian interpretation and some possible applications and open issues. This paper can be read as an introduction to Bayesian NL Interpretation.
- Anton Benz: “[Causal Bayesian Networks, Signalling Games and Implicature of ‘More Than n’](#)”. Benz applies Causal Bayesian Nets and signalling games to explain the empirical data on implicatures arising from ‘more than n’ by modelling the speaker with these nets.
- Satoru Suzuki: “[Measurement-Theoretic Foundations of Logic for Better Questions and Answers](#)” The paper is concerned with finding a qualitative model of reasoning about optimal questions and makes a promising proposal. It is part of a wider programme to find qualitative models of other reasoning tasks that are normally approached by numerical equations like the stochastic reasoning in Bayesian interpretation.
- Stefan Kaufmann: “[Conditionals, Conditional Probabilities, and Conditionalization](#)” Kaufman gives a logical analysis of the relation between the probability of a conditional and the corresponding conditional probability, proposing Bayes’ theorem as the link.
- Christian Wurm: “[On the Probabilistic Notion of Causality: Models and Metalanguages](#)” Wurm addresses the well-known problem of the reversibility of Bayesian nets. Nets can be turned into equally other nets by reversing all the arrows.
- Mathias Winther Madsen: “[Shannon Versus Chomsky: Brain Potentials and the Syntax-Semantics Distinction](#)” Based on a large number of existing experimental results, Madsen argues for a simple information theoretic hypothesis about the correlates of N400 and P600 effects in which an N400 is the sign of a temporary loss of hypotheses and a P600 the sign of too many hypotheses. This information theoretic approach has a strong relation with incremental Bayesian interpretation.
- Jacques Jayez: “[Orthogonality and Presuppositions: A Bayesian Perspective](#)” Jayez gives a direct application of Bayesian interpretation to the differential behaviour of various presupposition triggers in allowing presupposition suspension.
- Grégoire Winterstein: “[Layered Meanings and Bayesian Argumentation: The Case of Exclusives](#)” Winterstein applies a Bayesian theory of argumentation to the analysis of exclusive particles like “only”.
- Ciyang Qing and Michael Franke: “[Variations on a Bayesian Theme: Comparing Bayesian Models of Referential Reasoning](#)” Inspired by game-theoretical

pragmatics, Qing and Franke propose a series of improvements to the RSA model of Goodman and Frank with the aim of improving its predictive power.

- Peter R. Sutton: “[Towards a Probabilistic Semantics for Vague Adjectives](#)” Sutton formalises and defends a nominalist approach to vague predicates in situation theory in which Bayesian learning is directly responsible for learning the use and interpretation of such predicates without an intervening logical representation.

The volume grew out of a workshop on Bayesian Natural Language Semantics and Pragmatics, held at the European Summer School of Logic, Language and Information (ESLLI) 2013 in Düsseldorf. We would like to thank all the workshop participants for their contributions to the success of the workshop and the quality of the papers in this volume. Furthermore, we thank the members of the programme committee, Anton Benz, Graeme Forbes, Fritz Hamm, Jerry Hobbs, Noah Goodman, Gerhard Jäger, Jacques Jayez, Stefan Kaufmann, Uwe Kirschenmann, Ewan Klein, Daniel Lassiter, Jacob Rosenthal, Remko Scha, David Schlangen, Markus Schrenk, Bernhard Schröder, Robert van Rooij, Grégoire Winterstein and Thomas Ede Zimmermann, for their help in selecting suitable contributions. We also owe thanks to the series editor Chungmin Lee, to Helen van der Stelt of Springer and to the anonymous reviewers who helped out in the reviewing and selection process. Last but not least, we thank the German Society for Computational Linguistics & Language Technology (GSCL) for its financial support of the workshop.

Henk Zeevat
Hans-Christian Schmitz

Contents

Perspectives on Bayesian Natural Language Semantics and Pragmatics	1
Henk Zeevat	
Causal Bayesian Networks, Signalling Games and Implicature of ‘<i>More Than n</i>’	25
Anton Benz	
Measurement-Theoretic Foundations of Logic for Better Questions and Answers	43
Satoru Suzuki	
Conditionals, Conditional Probabilities, and Conditionalization	71
Stefan Kaufmann	
On the Probabilistic Notion of Causality: Models and Metalanguages	95
Christian Wurm	
Shannon Versus Chomsky: Brain Potentials and the Syntax-Semantics Distinction	117
Mathias Winther Madsen	
Orthogonality and Presuppositions: A Bayesian Perspective	145
Jacques Jayez	
Layered Meanings and Bayesian Argumentation: The Case of Exclusives	179
Grégoire Winterstein	

Variations on a Bayesian Theme: Comparing Bayesian Models of Referential Reasoning 201
Ciyang Qing and Michael Franke

Towards a Probabilistic Semantics for Vague Adjectives 221
Peter R. Sutton

Contributors

Anton Benz Centre for General Linguistics, ZAS, Berlin, Germany

Michael Franke Linguistics, University of Tübingen, Tübingen, Germany

Jacques Jayez ENS de Lyon and L2C2, CNRS, Lyon, France

Stefan Kaufmann University of Connecticut, Mansfield, USA

Mathias Winther Madsen Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, The Netherlands

Ciyang Qing Linguistics, Stanford University, Stanford, CA, USA

Peter R. Sutton Institute for Language and Information, Heinrich Heine University, Düsseldorf, Germany

Satoru Suzuki Faculty of Arts and Sciences, Komazawa University, Setagaya-ku, Tokyo, Japan

Grégoire Winterstein Aix Marseille Université, Laboratoire Parole Et Langage and Nanyang Technological University, Singapore, Singapore

Christian Wurm Universität Düsseldorf, Dusseldorf, Germany

Henk Zeevat ILLC, University of Amsterdam, Amsterdam, The Netherlands

Perspectives on Bayesian Natural Language Semantics and Pragmatics

Henk Zeevat

Abstract Bayesian interpretation is a technique in signal processing and its application to natural language semantics and pragmatics (BNLSP from here on and BNLI if there is no particular emphasis on semantics and pragmatics) is basically an engineering decision. It is a cognitive science hypothesis that humans emulate BNLSP. That hypothesis offers a new perspective on the logic of interpretation and the recognition of other people's intentions in inter-human communication. The hypothesis also has the potential of changing linguistic theory, because the mapping from meaning to form becomes the central one to capture in accounts of phonology, morphology and syntax. Semantics is essentially read off from this mapping and pragmatics is essentially reduced to probability maximation within Grice's intention recognition. Finally, the stochastic models used can be causal, thus incorporating new ideas on the analysis of causality using Bayesian nets. The paper explores and connects these different ways of being committed to BNLSP.

Keywords Bayesian interpretation · Semantics · Pragmatics · Stochastic natural language interpretation · Presupposition · Implicature

Overview

This paper explains six ways in which one can understand the notion of Bayesian Natural Language Semantics and Pragmatics and argues for each of them that they merit consideration since they seem to allow progress in understanding human language interpretation and with respect to problems in the area of NL semantics and pragmatics. The six ways can be listed as in (1).

H. Zeevat (✉)

ILLC, University of Amsterdam, Amsterdam, The Netherlands
e-mail: henk.zeevat@uva.nl

© Springer International Publishing Switzerland 2015
H. Zeevat and H.-C. Schmitz (eds.), *Bayesian Natural Language Semantics and Pragmatics*, Language, Cognition, and Mind 2,
DOI 10.1007/978-3-319-17064-0_1

- (1)
 - a. the proposal to compute the most probable NL interpretations by taking the product of prior and likelihood
 - b. the hypothesis that the human brain has implemented the proposal in a.
 - c. the Bayesian logic of interpretation: analysing rational interpretation as finding the maximum product of prior and likelihood
 - d. coordination on meaning using Bayesian interpretation
 - e. analysing intention recognition by Bayesian interpretation
 - f. switching to causal models for determining prior and likelihood.

The paper follows (1) trying in each case to explain the concept and its merits, before embarking on some controversial consequences and some urgent issues. What is the best way to implement NL interpretation is the best way to implement NL interpretation and the brain is also what it is. Therefore (a) and (b) will have to be settled by future research. (c) and (d) however can be seen as a thesis about the nature of NL interpretation. (c) and (d) commit one to a Bayesian analysis of Grice's views on meaning and the repercussions of those analysis for the notion of rational NL interpretation. (a) and (b), and especially the characterisation they give of the problem of NL interpretation as finding one's way through massive ambiguity would give important support to this thesis. The more speculative (e) and (f) deal with a counterargument to the thesis, i.e. the argument that it cannot deal with the causal component of the thesis.

1 Computational BNLSP

Bayesian Interpretation is an important technique in the AI discipline of signal processing, especially in speech recognition and computer vision. Its basis is Bayes' theorem, stated in (2). The theorem analyses the posterior probability of hypothesis H given a signal S , $p(H|S)$ as the product of the prior probability of the hypothesis $p(H)$ and the likelihood $p(S|H)$ divided by the probability of the signal $p(S)$. It essentially gives an alternative way to find $p(H|S)$ if $p(H|S)$ cannot be reliably measured in a direct way.

$$(2) \quad p(H|S) = \frac{p(H)p(S|H)}{p(S)}.$$

This is simply derived as follows:

$$p(S \& H) = p(H \& S) = p(H)p(S|H) = p(S)p(H|S)$$

$$\text{Therefore } p(H)p(S|H) = p(S)p(H|S)$$

$$\text{and } p(H|S) = \frac{p(H)p(S|H)}{p(S)}$$

If S is given, the most probable H for a given S written $\text{argmax}_H p(H|S)$ can be computed in a different way given Bayes' theorem. $\text{argmax}_H p(H|S) = \text{argmax}_H (p(H)p(S|H))$, since $p(S) = 1$ if S is given. If one is searching for the most probable interpretation H of a signal S , it will therefore suffice to have a good estimation of the probability of H (called the prior, the probability assigned to H before the signal occurred) and of the probability $p(S|H)$ (the likelihood of the

signal given the interpretation). This way of finding the most probable H is useful and often unavoidable, since it can be difficult or even impossible to directly estimate $p(H|S)$ (the posterior probability of the interpretation H , after the occurrence of the signal S).

The difficulty for direct estimation is especially clear when the set of signals is infinite or continuous, as in speech perception or in vision. For example, in speech perception, one measures the distance between the perceived signal and the predicted signal to estimate likelihood and uses the prediction for the next phoneme from the preceding phoneme sequence obtained from a large body of written text to estimate the prior.

The definition just given makes BNLI a special way of obtaining the most probable interpretation of an utterance and BNLI is thereby a branch of stochastic signal interpretation.

One needs a good argument for using Bayesian interpretation rather than the direct method of estimating the posterior. And one needs to accept that natural language is ambiguous to go stochastic in the first place.

The method by itself does not entail that both likelihood and prior need to be estimated stochastically. The prior can be partly computed by symbolic techniques, e.g. by a theorem prover showing that the prior is not 0 by showing the consistency of the interpretation with the context. Similarly, one can use symbolic techniques (e.g. a text generator in NLI—these are standardly symbolic) to estimate likelihood.

Doing NLI or NLSP with this technique makes BNLSPP a branch of stochastic NLSP. Stochastic NLSP obviously helps in obtaining good results in interpretation, since NL utterances severely underdetermine their interpretations in context and other techniques for disambiguation are limited in scope. It is also easy to see that very often disambiguation cannot be explained by an appeal to hard facts and logic, but needs to appeal to what is probable in the world. For example, if I tell you that I took the bus, you will understand me as having used the public transport system and not as having lifted an autobus with my crane. Another basic example is (3).

(3) Tweety flies.

If you are like me, Tweety is the bird from AI discussions about default reasoning and “flies” means “can fly using his wings”. There must be many other Tweeties however and there is no logical reason why the verbal form could not mean “uses aeroplanes” or “can pilot aeroplanes”.

The underdetermination of meaning by form can be argued by looking at any branch of linguistic description. The most obvious and most traditional place is the lexicon. Non-ambiguous words are rare to non-existent. Chomsky’s proposals for syntax brought syntactic ambiguity to the centre of attention. Formal treatments of phonology and morphology have also led to phonological and morphological ambiguity, while the semantics of morphemes makes them much more ambiguous than proper lexical items. Montague grammar initially concentrated on quantifier scope ambiguities. But this is just another instance of semantic relations that remain unmarked in syntax. Famous examples in English are noun-noun combinations and

the semantic relation to the main clause as marked by a participial construction. Areas of pragmatics that can be described as “ambiguous” are pronoun resolution, the recognition of rhetorical relations, temporal reference resolution, exhaustivity and stereotypicality effects and the treatment of presuppositions.

It follows that success in NLSP can only be expected by viable approaches to selection of interpretations, an aspect of the form-meaning relation which is still neglected with respect to the composition of meaning within semantics and pragmatics.

Ambiguity adds up to an argument for stochastic NLI. But not necessarily to an argument for Bayesian NLI. Within the confines of stochastic NLI there is another hypothesis possible, the direct method in which a system of hierarchically organised cues directly leads to the most probable meaning. It is not directly obvious that this cannot work for NLI. In stochastic parsing for example there have been successful models based on direct estimation.

We will look at some evidence about human NLI and will argue that human BNLI is the most plausible hypothesis explaining that evidence. From that hypothesis, an argument can be constructed against the direct method in computational NLI.

2 Human BNLS

In cognitive science, BNLS can be seen as the claim that humans emulate BNLS when they are interpreting language utterances.

There are general grounds for believing that this is the case. Some (Oaksford and Chater 2007; the contributions in Doya et al. 2007; Kilner et al. 2007) speak of an emerging consensus that Bayesian reasoning is the logic of the brain—even though there is also considerable opposition to this view (e.g. Bowers and Davis 2012). Such an emerging consensus builds on successes of Bayesian analyses in major areas of cognitive science such as inference, vision and learning. It would follow from the importance of Bayesian reasoning in the brain that it is unlikely that human NLI would employ a different logic.

The same point can also be made by developing the analogy between vision and NLI. Visual cues are highly ambiguous and the standard procedure in automatic vision is based on the availability of a powerful model of likelihood: the translation of the hypothesis into properties of the signal by means of the “mental camera”. Together with powerful priors about what can be perceived this gives a way to find the most probable hypotheses among the ones that are made available by cueing.

Human language interpreters have a device much like the mental camera in their ability to translate communicative intentions into natural language utterances. In addition, most of the priors for vision are directly usable in NLI. It follows that they can find the most probable interpretation for an NL utterance as reliably as they can find the most probably perceived scene for a visual signal. In sum, there is no obstacle for assuming that human NLI has the same architecture as human vision and a strong evolutionary argument for assuming that the architecture of human vision

and the priors involved were exapted for the emerging human NLI.¹ This would be the simplest way in which the problem of NLI could be solved by the brain. Given the short time frame in which NL and NLI emerged, a completely different architecture like symbolic parsing would be difficult to achieve in evolution.

If we accept the conclusion, there is an aspect of interpretation that is not predicted by the direct cueing model: the role of the priors. The priors shared with vision would be generalisations about what is going on in the world to the degree that it is accessible to normal perception. Such priors can be seen as causal prediction, topographical information and conceptual knowledge. Thereby they concern primarily the highest levels of linguistic analysis: semantics and pragmatics.

For a pure hierarchical cueing model, the direct way of estimating the posterior, having the best priors on the highest level is a problem. Those priors can only help selection at the final levels of analysis and would be relatively unimportant at the lower levels, leading to a large number of possibilities that need to be checked before the final level. In a Bayesian model, the high-level priors can directly predict down to the lowest level. The most likely prior hypothesis that is cued by the input can be used for simulated production that can be checked against initial segments of the input.

It would appear therefore that a causal model of production combined with priors as for vision is the better architecture for early selection between interpretational possibilities.

There is also a body of empirical evidence that makes sense under a Bayesian architecture for natural language interpretation. Pickering and Garrod (2007) gives an overview of evidence for the simulated production in interpretation that is predicted by Bayesian interpretation. Clark (1996) discusses conversational moves that end with the hearer co-producing the final segment of the utterance, something that is hard to understand without assuming simulated production. As noted by Kilner et al. (2007), the activation of mirror neurons in perception of the behaviour of others makes good Bayesian sense: connecting to one's own way of carrying out the same behaviour will give access to more fine-grained prediction from the hypothesis that can be matched against the visual signal.

So there is a lot of evidence in favour of a Bayesian architecture of human language interpretation. Human BNLI also supports the idea that efforts should be made to achieve computational BNLI. If one seems to understand how humans solve a problem, that gives a strategy for solving the same problem on a computer. There may certainly be good computational reasons to depart from the human solution. But in the face of the serious imperfections of current NLI and the absence of sufficient investment in BNLSP, a computational case against BNLSP is just not available at this moment.

Computational NLSP using the direct method seems to be forced into the pipeline model, due to the different data-sets needed for the different kinds of processing. A pipeline model using the corpus-based methods of current stochastic parsing would

¹Symbolic parsing with visual grammars becomes intractable as soon as the grammar becomes interesting (Marriott and Meyer 1997). It follows that visual recognition must be stochastic. To profit fully from the mental camera, it must also be Bayesian.

be an alternative model for human NLI. It would involve finding the most probable syntactic tree, followed by finding the most probable interpretation of the words in that tree, followed by the best proto-logical form, followed by the best fully resolved logical form and its most probable integration in the context. The problem is simply that a 0.8 success score on each of the levels (roughly the result for accuracy in stochastic parsing) translates into a $0.8^5 = 0.328$ score for the full pipeline, very much below human performance.

Bayesian models escape from this problem by assuming a causal model of utterance production that works in conjunction with a heuristic cueing system. If the high-level priors are strong enough and the causal model is accurate and sufficiently deterministic, the combination has a much higher chance of reaching the most probable interpretation. The direct method is not eliminated in this set-up since it is needed for defining the cueing system and also because it could count as a baseline, that is improved by the causal model. The causal model of production seems more adequate however for evaluating linguistic knowledge in the process.

3 BNLSP as the Logic of Interpretation

The logic of interpretation in Montague grammar is just classical logic augmented by a theory of abstraction and function application. It does not deal with uncertainty. Boole developed his Boolean logic from the broader perspective of a logic that combines classical reasoning with reasoning about probabilities. In making BNLSP the logic of interpretation, one restores the Boolean view of logic to deal with the uncertainty about what a particular word or phrase means on a particular occasion.

If one takes the selection problem seriously and accepts the rationality of trying to solve it by probability maximation, BNLSP offers a promising starting point for the analysis of semantics and pragmatics. There is little choice about taking the selection problem seriously. As noted above, modern linguistics can be read as an overwhelming case for the underdetermination of meaning by form and a good deal of the evidence comes precisely from the philosophical exploration of presupposition, thematic roles, implicature and discourse relations. To doubt the rationality of probabilistic disambiguation is a way of doubting rationality as such, given that stochastically based disambiguation is needed for the interpretation of nearly any utterance.

BNLSP motivates some simple principles. For a good interpretation, prior probability must be properly higher than zero and likelihood must be properly higher than zero. If another interpretation has the same likelihood, it must have a lower prior. If another interpretation has the same prior, it must have a lower likelihood. These are powerful principles that can explain important aspects of interpretation. The following are some examples of their application.

Gricean Maxims

The simple BNLSP principles have a different status from Gricean pragmatics, since they are merely consequences of stochastic disambiguation. A reduction of Gricean pragmatics to these principles is thereby a step forward since it eliminates the normative account of production that is the cooperativity principle. The reduction is however a question of reducing the Gricean implicatures to the principles, not a reduction of the Gricean maxims to these principles.

The principles cannot motivate the Gricean maxims or the cooperativity principle on which they are based, though there are important relations. For example, anything that increases the relevance of the utterance interpretation is automatically something that increases the likelihood of the interpretation. If manner is defined as choosing the most probable way of expressing whatever the speaker wants to express, it is also directly related to likelihood: the speaker then boosts the likelihood of the interpretation. Transgressions of manner reduce likelihood and may be the trigger of additional inferences to bring it up again. Quality is related to prior maximization: overt falsehoods do not qualify for literal interpretation. Underinformative contributions are misleading since they will lead—by the increase in likelihood due to extra relevance—to unwarranted exhaustivity effects. Overinformative contributions will merely force accommodation of extra issues and are not harmful.

The lack of direct fit is due to the interpretational perspective of Bayesian pragmatics, while Grice gives a production perspective in the form of norms the speaker is expected to follow. The interpretational perspective merely searches for the most probable explanation of the speaker's utterances.

Given its simplicity and its pure rationality, it is surprising that pragmatics can be reduced to Bayesian pragmatics. The following runs through some effects of the Gricean maxims. The explanations are not new or surprising. It is surprising however that so few assumptions are needed once one operates under Bayesian interpretation.

(4) (Quality) The cat is on the mat but I don't believe it.

The assumption that comes with assertions that the speaker believes the content leads to contradiction and thereby to a zero prior. A norm that speakers must speak the truth is not necessary.

(5) (Manner) Mrs. T. produced a series of sounds closely resembling the score of "Home Sweet Home".

The likelihood of the speaker reporting that Mrs. T. sang "Home Sweet Home" in this particular way is very low, since the verb "sing" would be the most probable choice for expressing the report. A lack of appreciation for the singing and the strategy of innuendo by prolix expression are the assumptions under which likelihood is driven up by adding the speaker's lack of appreciation for the singing to the interpretation. The innuendo-by-prolix-expression strategy is an extra assumption, but one that like the strategy of irony is needed in a description of speaker culture in standard English. It is not to be expected that these strategies are universally available.

Grice illustrates the flouting of the relevance maxim by (6), in the context of a direct question asking for the academic qualities of the person in question.

(6) (Relevance) He has a beautiful handwriting.

As an answer to the direct question the purported answer has a likelihood of zero: it just does not answer the question. The implicature that the person is not very good arises from an explanation of why the speaker does not answer the question. The speaker has nothing positive to say, but feels he cannot say that directly.

These explanations do not appeal to the maxims and just use likelihood and some facts about the example, e.g. that having a beautiful handwriting is not an academic quality or that one cannot use oneself as an authority knowing that the cat is on the mat if one is known not to believe that.

Parsing, Consistency and Informativity

Bos (2003) implemented a combination of parsing and theorem proving designed to capture the account of Van der Sandt (1992) of presuppositions. The system parses English sentences, resolves pronouns and resolves or accommodates the presuppositions induced by presupposition triggers in the sentence and proves consistency with the context and informativity in the context of the interpretation. The effect is contextual incrementation with consistent and informative interpretations obtained from linguistic input.

The system can be motivated fully from BNLSP. To show that the sentence parses with a compositional interpretation that can be resolved and enriched by accommodation to a contextual increment shows that it has a non-zero likelihood: the sentence can be produced for the interpretation and the anaphora and presupposition triggers are licensed with respect to the context. To show that the interpretation is consistent with the context, shows that the increment has a non-zero prior. Informativity finally shows that the likelihood is non-zero for another reason, by showing that the sentence makes sense as an attempt to give the hearer new information. An assertion that would not give new information is not rational and as such has zero likelihood.

The extra premises needed are grammar, licensing conditions for presuppositions and anaphora and the assumption that asserting is trying to make the hearer believe something. The system also captures part of disambiguation as a side effect of consistency and informativity checking.

The system can be described as assigning a uniform distribution of prior probability over interpretations consistent with the context and a uniform distribution of likelihood over interpretations of an utterance U that follow DRS-induction and are informative. The system can be made more sophisticated by refining these priors and likelihoods. The first could be used to decide between the two readings of “Mann beißt Hund” (ambiguous in German between “man bites dog” and “dog bites man”) in favour of the dog as the biter, the second would prefer the more frequent SVO order for the interpretation and seems to be the clear winner in this particular case.

The point is that informativity and consistency would follow from BNLSP without needing a basis in the analysis of assertion of Stalnaker (1979), the starting point of Van der Sandt (1992).

Lexical Interpretation

Hogeweg (2009) formulates a simple theory of lexical interpretation based on Smolensky (1991). In Smolensky's account experience of the use of lexical items associates lexical items with a number of binary semantic micro-features that are observable in the utterance situation and connected with the use of the word. New uses of the word activate all those micro-features and in this way may overspecify the lexical meaning: there may be more features than can consistently be assumed to hold in any context.

An example is the particle "already" that expresses both "perfectivity" and "surprise". (7) expresses both that Bill has gone (perfectivity) and that this was earlier than expected (surprise).

(7) Bill already went home.

But in the use with an infinitive (Singapore English and colloquial British English) as in (8), surprise has been removed.

(8) You eat already?

Hogeweg turns lexical interpretation into an optimality-theoretic constraint system: $FIT > MAX$ where FIT keeps the local reading consistent and MAX tries to incorporate as many semantic features as possible. The theory has been successfully applied to particles and prepositions and is promising for other categories of meanings. E.g. it captures the observations of Dalrymple et al. (1998) about *each other* as in (9).²

(9) a. The lines cross each other. ($\forall x, y \in L \text{ cross}(x, y)$)
 b. The chairs are on top of each other. ($\forall x \in C \exists y \in C (on(x, y) \vee on(y, x))$).

The point is however that FIT is nothing more than the demand to have a non-zero prior for the interpretation and that any feature projected by MAX will be an additional possible explanation for the use of the lexical item, thereby making it more likely. The OT constraints and their ordering are merely a restatement of BNLS in one particular setting.

The Cumulative Hypothesis

The cumulative hypothesis in presupposition projection (Langendoen and Savin 1971; Gazdar 1979) is the idea that presuppositions triggered in a context project by default.

The hypothesis finds a lot of empirical confirmation: the best empirical results are obtained by Gazdar (1979), Van der Sandt (1992), Heim (1983) that incorporate aspects of the cumulative hypothesis. Rejection of the hypothesis leads to the generalised proviso problem in which cumulative effects need to be explained by other means.

²The choice of the right features is not a trivial matter.

In Bayesian interpretation the cumulative hypothesis seems an automatic consequence.

The basic assumption in any account of presupposition triggers must be that these devices need their presupposition in order to do their semantic work of referring (e.g. definite description) or predicating (lexical presuppositions) or their pragmatic work of relating the host utterance to the context (particles). It follows that when the context does not already make the presupposition available, a bridging inference explaining the use of the trigger is needed to make it available.

The possibilities for these bridging inferences are given by the operators under which the trigger appears and by contextual knowledge. Operators like $\lambda qp \rightarrow q$ or quantifiers $Q(P, R)$ allow explanations of the trigger use in which the common ground extended by p or Pa (for an arbitrary a) provides the presupposition, if trigger T of presupposition p occurs in the scope of the operator. Likewise, an attitude verb V allows the beliefs of the subject of V together with parts of the common ground that are consistent with these beliefs to provide the presupposition, if T occurs in the complement of V . A possibility operator introducing a possibility P finally allows consistent extensions of the possibility with parts of the common ground to provide the presupposition, if the trigger is part of an addition to the same possibility P (as in: A wolf might come in. It would eat you first.). The simplest explanation of the trigger is however always the assumption that the speaker believes the presupposition.

All of these possible bridging inferences will be inhibited by evidence against them: this brings down the prior of the interpretation. Each bridging inference will however contribute to the overall likelihood of the trigger.

The different explanations are not independent. Assuming that the speaker believes the presupposition puts it into the common ground and extensions thereof and makes it more probable that the subject of an attitude also believes the presupposition. The gain in likelihood due to the subordinate inferences is therefore already—completely or partly—realised by assuming that the speaker believes the presupposition. It also makes sense to assume that the priors of the different explanations are completely or partly dependent on each other, with the priors of the presupposition in more subordinate contexts going up if the presupposition is part of a higher context.

What these considerations taken together make plausible is the principle (10). Each extra context with the presupposition increases the likelihood of the trigger. With each addition, the cost in terms of decreasing priors becomes smaller.

- (10) the probability of the interpretation increases *ceteris paribus* if the interpretation I of S contains the presupposition of a trigger T in S in a context c of T .

The *ceteris paribus* is essential. If there is evidence against the presupposition being in a context, that suffices for blocking the inference unless the evidence is inconclusive and there is no other way to give the trigger a non-zero likelihood.

The principle motivates a preference for the assumption of the speaker belief in the presupposition. That preference can be undone if the prior for that explanation is lower than the prior for the alternative explanations.

That preference is the cumulative hypothesis. The principle is slightly stronger³ which seemed a preference for assuming all the explanations of the trigger. This seems to be the intuition in a sentence like (11), uttered in a context in which it has not been established that John has a niece or that he believes he has a niece.

(11) John believes that his niece will visit him next week.

The example allows readings where John believes that some woman who is not his niece but whom John thinks is his niece to visit him and readings where John believes that a woman—who unbeknownst to John is in fact his niece—will visit him, but these require special priors. The default reading is that John believes he will be visited by a person who he believes to be his niece and who is in fact his niece.

Satisfaction

But it would be wrong to assume that the logic of interpretation is purely an implementation of the cumulative hypothesis like Gazdar (1979) since it also predicts the satisfaction theory as in (12).

(12) if the interpretation does not make the presupposition available to the trigger, the trigger cannot do its semantic or pragmatic job and the sentence containing the trigger has likelihood 0.

It is precisely the interaction between the cumulative hypothesis and the satisfaction theory that explains presupposition projection. This is a restatement of the conclusion of Soames (1982) and explains why Soames and both Heim (1983) and Van der Sandt (1992) have such good empirical results. The Bayesian underpinning for both the cumulative hypothesis and the satisfaction theory is however new to the current paper. And it is different, because projection is not a side effect of (12), but a question of bridging inferences that increase the likelihood of the trigger.

The ancient counterexample (13) to a satisfaction-theoretic account of projection (having grandchildren entails having children, so it should not project that John has children) should be a case where the gain in likelihood outranks the loss in prior due to the assumption that John has children.

(13) If John has grandchildren, his children will be happy.

The cumulative hypothesis makes the use of a presupposition trigger a weak signal that the speaker believes in the presupposition. It is a weak signal, because it can be easily overridden, but it does not stop being a weak signal if the presupposition is

³The principle gives the preference for global accommodation in terms of Heim (1983) and Van der Sandt (1992) in terms of BNLSP: nothing else would do. It gives however what I defend in Zeevat (1992), a preference for multiple accommodation, something which does not make sense on the versions of the satisfaction theory in these two references.

locally entailed. This is in some sense unfortunate, since in many cases the satisfaction theory correctly predicts non-projection, and Gazdar's explanation of non-projection in terms of clausal implicatures is problematic.⁴ These matters will be taken up in more detail in Zeevat (2014a) that gives a more complete Bayesian account of presupposition. In that paper non-projection is explained by a Bayesian account of causal inference between a given but non-entailed cause and the presupposition. Projection of the presupposition would then force the projection of the non-entailed cause as well. Examples of the phenomenon are in (14).

- (14) a. If John is a diver, he will bring his wetsuit to the party.
 b. If John is a diver and wants to impress his girlfriend, he will bring his wetsuit to the party.
 c. If John wants to impress his girlfriend, he will bring his wetsuit to the party.
 d. John dreamt that he was a baker and that everybody knew he had been in jail.
 e. John dreamt that he was a baker and that everybody knew his bagels were the best in town.

In (a) and (b), the causal inference: John's being a diver causes his having a wetsuit prevents projection, as in (c). In (e), it is inferred that John being a baker is the cause of his bagels being as good as they are in his dream and prevents the projection of the complement of *knew* as in (d).

The exception (15) is explained by the fact that the two possible causal inferences (Mary has to leave the States because she has no green card, Mary has to leave the States because she has committed a crime) block each other: neither explanation will beat the probability of all other possible explanations taken together.

- (15) If Mary has no green card, she will be sorry to leave the States, but if she has committed a crime, she will not be sorry to leave.

The Hermeneutic Circle

This is the idea in philology, history, philosophy and literary criticism that there is a never-ending improvement of understanding by a process of rereading and reinterpretation.

In a Bayesian view, every increase in understanding will lead to better priors. The increased understanding can be understood as a better approximation to the world of the text as given by the text. This would translate into a better prediction from the pre-context to any point in the text either by eliminating uncertainty or by becoming closer

⁴Garcia Odon (2012) is an admirable discussion of the empirical facts around projection in Karttunen filters. Her analysis however rests on conditional perfection. Causal inferences seem a much better way of deriving the crucial pattern: $p \rightarrow q$, where p is the presupposition and q the non-entailed cause of p . If it is given that p and that q is the cause of p then q must be the case as well. The non-projection in (15d) also shows that the causal explanation also works outside Karttunen filters.

to the prediction available in the context of writing. For especially older texts, the improvement in the prior also leads to a better base for understanding the likelihood, that interpreters typically know less well than contemporary language use.

Can the prior become stable? There is no Bayesian reason why it could not eventually stabilise. The problem for many of the enterprises in which the hermeneutic circle plays a role is that there is insufficient evidence available to eliminate the uncertainty in the prior to such an extent that the text loses its poly-interpretability. Convergence to a final interpretation is then unlikely.

4 BNLSP and Coordination

Alvin Liberman's Motor Theory of Speech Perception states that speech perception is the distal perception of the motor gestures of the speaker's articulatory system (Liberman et al. 1967).

Distal perception is the perception of events that are not purely perceptual. For example, one sees that John is taking the elevator by seeing that John is pushing a button and one sees the latter by seeing John's finger touch the button. The intention of pushing the button and the further intention of using the elevator are the aspects that cannot be directly seen. The motor gestures of the various movements of the tongue, the vocal chords, the lips, the lungs and the mouth that make up vocalisation can not be directly perceived in the same way, but in perceiving that someone said the word "computer" one would be committed to perceiving the articulatory gestures making up the pronunciation of that word. Liberman formulated the theory in response to the discovery of coarticulation that makes it hard to see the acoustic signal as a series of discrete sound events.

As a psychological hypothesis the theory did not do well until the discovery of mirror neurons and was criticised repeatedly because its predictions did not find confirmation. Liberman retreated to a perhaps more interesting thesis, stating that the motor theory is one of a range of hypotheses that would answer the parity problem. Speaking and hearing are very different processes and there must be a level of representation at which the analysis of a speech signal and the corresponding articulation process receive an identical representation. It is precisely here that mirror neurons are interesting. Under the hypothesis that these code motor intentions (Brozzo 2013; Rizzolatti and Craighero 2004), the observation that they fire both in production and perception makes this a suitable candidate for a level of representation at which parity is reached.

The effect on the hearer can be seen as a simulation of the speaker in carrying out the vocalisation.

From the perspective of Bayesian interpretation, the intention to articulate some utterance adds something to the quality of speech perception. It makes it possible to use the prediction about what one's own articulation would sound like as a check on the correctness of the hypothesis. The connection with the production is thereby a way to improve the quality of the perception and contributes directly to the probability

that there will be identity between the intended vocalisation and the perceived one, i.e. to the probability of coordination.

Grice's account of non-natural meaning is closely related. According to Grice, the non-natural meaning of a signal is the intention to reach some effect on the hearer by means of the signal through the hearer's recognition of that intention.

From the hearer perspective, understanding an utterance is therefore recognising the intention of the speaker in producing the utterance. Again, this is a case of distal perception in which one arrives at a simulation of the speaker's plan in producing the utterance.

Again, the simulation allows a better prediction about the utterance that can be checked with respect to the perceived utterance, thereby enhancing the quality of the interpretation.

The situation is much like the standard example in the mirror neuron literature, that of grasping a glass of water by an experimenter. If the neuron fires in perception, the subject simulates the motor intention of the experimenter in its pre-motor cortex.

On the basis of this evidence, it would seem that the brain has exapted parts of its motor skills in making sense of the behaviour of similar organisms in order to enhance its prediction of their behaviour. One recognises the behaviour as possible behaviour of one's own, in recognising a particular vocalisation, in recognising the speaker intention of some utterance or in recognising the grasping of a glass in the behaviour of a similar organism.

5 Recognition and BNLSP

The examples above depend essentially on recognition. Recognition however is not the same as disambiguation by means of probability maximation. It is not enough that the interpretation is a probabilistic maximum among the different possible interpretations of the utterance, but it should also come with the certainty that it is the right interpretation. There should not be two rational probability distributions P and P' such that $P(x) < P'(y)$, even if x is the winner in any P and P' .⁵

⁵There are other ways of formulating the inequality, e.g. in all rational P $P(x) < P(y)$. Rational probability assignments can be formalised straightforwardly on n -bounded models ("worlds") for a finite first order language L . An n -bounded model has a domain with cardinality smaller than n . An n -bounded model can be seen—under isomorphism—as a row in a truth-table for propositional logic and can interpret a given set of connections by probabilities given directly by the frequencies for the connection in the model. This will assign a probability $p_w(e)$ to an event e , deriving from the presence of other events in w and connections of them to e . Assume a probability distribution P over W , the set of worlds. P can be updated by bayesian update on the basis of observations e and the distributions p_w over events. The probability of a formula of L can now be equated with the sum of the probabilities of the worlds on which it is true. A rational probability assignment is an assignment P over W which respects the beliefs of the subject (the beliefs receive probability 1) and the experience of the subject, in the double sense that the experiences themselves are treated as beliefs and receive probability 1 and that the probability assignments to the worlds are updated by Bayes' rule on the basis of those experiences. In this way, a world acquires more probability

This would make “to recognise x ” be the conjunction of (16.1) and (16.2). In (16.1) and (16.2) $\varphi \gg \psi$ stands for the property that for all rational probability assignments P and P' $P(\varphi) > P'(\psi)$. (16) then defines recognition.

- (16) 1. x causes a signal s which evokes x in turn
- 2. $cause(x, s) \gg cause(y, s)$ for all other y evoked by s .

Recognition must be distinguished from the non-factive “seem to recognise x ” that can be defined as in (17).

- (17) 1. s evokes x
- 2. $cause(x, s) \gg cause(y, s)$ for all other y evoked by s .

Assuming (16) understanding an utterance u is to recognise the intention leading to u and the process of reaching understanding the construction of an intention that could cause the utterance and meets the probabilistic criterion for recognition.

An analysis of recognition that does not deal with the problem of certainty has to face the problem that there is no rational basis for settling for the particular recognised object and ruling out its competitors.

A visual example. Suppose I see the bus ride on the other side of the street. Of course I know that some movie people could very well be manipulating a cardboard copy of the bus at the right spot, but since there are no cameras, movie plans that I am aware of, good movie plans involving cardboard bus models being carried around, the cardboard bus has a very low prior probability compared with the every-day and every-fifteen-minutes event of the bus coming by.

A linguistic example. An utterance of the word “bun” can on a particular occasion sound more like “bunk”. Nothing goes wrong if the priors cooperate and to make an occurrence of “bunk” highly unlikely. But without the priors intervening “bunk” would have a slight advantage on “bun”, in every probability assignment. This should be an undecided instead of a win for “bunk”, and there is no recognition of the word bunk.

A pragmatic example is the example (18) from the presupposition literature.

- (18) If Mary does not have a Green card, she will be sorry that she has to leave the States, but if she has committed a crime, she will be happy that she has to leave.

Two different causes for Mary having to leave the States are both presented as possible, making it impossible that either of them wins for all rational probability assignments as the cause of her leaving the States: they can both be the cause and have probability 1. This predicts the projection of the fact that she has to leave the States, the intuitively correct prediction. If either cause were recognised as such and

(Footnote 5 continued)

to the degree that its own frequencies predict the frequencies in the experience. The information state of the subject can be defined as the combination of the set of worlds it allows and the rational probability assignments over those worlds. Updates eliminate worlds and reassign probability by experience. It is possible to remove the restriction to n-bounded worlds, by using probability density functions and Bayesian updates over those, but that is not as straightforward.

be only possible, the effect would also be only possible, preventing projection. In the absence of a recognised cause that is only possible, the presupposition will project, in line with the cumulative hypothesis.

Some other consequences of BNLSP are the following. First, it would follow that the speaker has to design her utterance so that intention recognition is possible. It should also hold that the utterance must in its context naturally associate with the intended interpretation.

Second, uncertainty elimination—it is not given that it is always successful—may be directly connected to the enhancement of perception by simulation. Accurately predicting the behaviour of conspecifics is vital and faulty prediction dangerous. The simulation enhancement would be necessary to be able to reach recognition.

6 Causality

In a loose sense, the use of Bayesian nets in NL semantics and pragmatics as a use of Bayesian methods in NL semantics and pragmatics also belongs to BNLSP. In the stricter sense used in this chapter this is not so: BNLSP proper is doing semantic and pragmatic interpretation by finding the maxima for the product of likelihood and prior.

This section explores the stronger view that Bayesian nets are crucially involved in explaining likelihood and prior, by postulating an intrinsic connection between probability and causality. This seems unavoidable for stochastic models of likelihood: the process from the interpretation to the form is clearly a causal process involving the “laws” of language production under which the speaker tries to achieve an effect on his audience by appropriate planning. If likelihood is given by a stochastic model and not by e.g. an algorithm, it still needs to learn the causal connections between the variables that are involved in the production process.

The role of causality is not as clear for the prior. In principle, one could have the view that while recognising the potential of Bayesian nets for capturing causality and helping with the causal aspects of NL semantics, the rest of BNLSP should be carried out without a commitment to this further restriction on probability distributions.

It seems however uncontroversial to equate the probability that p —even if it is subjective probability—with the sum of the products of the distinct possible causes of p and their prior probability. The model of the prior can then be a set of causal dependencies between stochastic variables with an estimation of the strength of these dependencies. This becomes a dynamic model if one adds a mechanism that takes in facts by putting the probability of the variables representing them at 1 and a second mechanism that adapts the strength of the dependencies on the basis of learning data for the dependencies. Further mechanisms could be added for updating the dependency structure itself. On the level of a dynamic model of the prior, there is no need to enforce conditions that make the system computational. One can have

loops and uncertainty about the dependencies. This just means that having proper predictions is not always guaranteed, but that seems a realistic feature. Computation is not possible in certain cases and sometimes several possibilities need to be computed for prediction.

A dynamic causal model of the prior is however bound to contain submodels that are proper structural causal models in the sense of Pearl (2009). These are standardly used for explaining and modeling causality and the stochastic version is equivalent with proper Bayesian nets. Proper structural causal structures are computable and their dependency relation is free from loops and finite.⁶

A structural causal model is a triple $\langle U, V, F \rangle$ where U and V are two sets of stochastic variables, the exogeneous variables U and the endogeneous variables V . F is a function that extends a valuation u for U to the whole net by specifying for each endogeneous variable V_i a function $f_i : U_i \times PA_i \rightarrow \text{range}(V_i)$ where $U_i \subseteq U$, $PA_i \subseteq V$ and $\text{range}(V_i)$ the possible values of V . Dependencies (“arrows”) can be read off from $PA_i \cup U_i$: if $W \in PA_i \cup U_i$, V_i depends on W , i.e. there is an arrow from W to V_i . The resulting graph is acyclic and directed if F is indeed a function that extends a valuation to U to $U \cup V$.

The exogeneous variables capture the state of the world, the endogeneous ones receive their values from each other and from the exogeneous variables, by means of functions f_i for each endogeneous variable V_i .

Let $Y(u)$ be the subfunction of F that assigns values y to $Y \subseteq V$ for values u for U .

The interesting property of structural causal models is that they allow the computation of the effect of setting some endogeneous variables X to values x , thus recomputing the values of the variables Y dependent on X , called the intervention $do(X) = x$.

This gives a submodel M_x of M where $X_x = x$ and for other endogeneous variables V_i (V_i) is computed by the original f_i , possibly using the new values for x for X .

This gives an intuitive approximation of counterfactuals. If X were x , then Y would be y , that makes sense only if Y depends on X . Other uses are the definitions in (19) taken from Pearl (2009).

⁶Bayesian nets are popular technology and introductions to Bayesian nets, their theory and their applications abound, as well as software environments for defining these nets and computing with them. Any attempt to give an example in these pages could not compete with what the reader can find immediately on the net. It is however unclear to the author how important Bayesian nets are for modeling the dynamic causal dependencies that are needed for BNLSP, though they definitely are an important source of inspiration.

- (19) X is a cause of Y if there exist two values x and x' of X and a value of u such that $Y_x(u) \neq Y_{x'}(u)$

X is a cause of Y in context $Z = z$ if there exist two values x and x' of X and a value of u such that $Y_{xz}(u) \neq Y_{x'z}(u)$

X is a direct cause of Y if there exist two values x and x' of X and a value u of U such that $Y_{xr}(u) \neq Y_{x'r}(u)$ for some realization r of $V \setminus (X \cup Y)$

$X = x$ always causes $Y = y$ if $Y_x(u) = y$ for all u , and there are realizations u' of U and x' of X such that $Y_{x'}(u') \neq y$
(pp. 222–223)

$X = x$ is an actual cause of $Y = y$ in a world $U = u$ if:

- (i) $X(u) = x$ and $Y(u) = y$ (i.e. both events are realized)
- (ii) There is a partition of V in Z and W , with $X \subseteq Z$, and there are values x' of X and w of W such that if $Z(u) = z$, then $Y_{x'}w \neq y$ and $Y_{xw'}z' = y$ for all w' realizations of $W' \subseteq W$ and z' of $Z' \subseteq Z$ that correspond to w in z in overlapping variables
- (iii) X is the smallest set of variables satisfying these two conditions.
(p. 330).

Structural causal models can be made probabilistic by assuming a probability distribution over U giving the probability $p(u)$ for possible values of U . In that case, the probability that $Y = y$ can be stated as (20).

$$(20) \quad p(Y = y) = \sum_{u:Y_u=y} p(u).$$

This causal probabilistic perspective on reality is inclusive. It provides us with the necessary probabilities for carrying out all of the other degrees of commitment to BNLS, but adds to that a semantic view on causality, counterfactuals and conditionals. Of special interest are the applications in pragmatics. E.g. the inference that Bill fell because John pushed him, for (21) can be formalised using the treatment of causality introduced in this section.⁷

- (21) Bill fell. John pushed him.

Section 3 and Zeevat (2014a) also make the case that it is causal inferences of this kind and not entailment of the presupposition by the antecedent that is responsible for non-projection in the examples that led Karttunen to the satisfaction theory of presupposition.

⁷The probability of the effect given the possible cause should outweigh its probability if the possible cause did not obtain.

7 Controversial Consequences of BNLSP in Linguistics

BNLSP is anything but neutral on quite a number of issues. With respect to many people's prejudices, it is properly contentious.

Syntax

The received view about the role of grammar in interpretation is the view from symbolic parsing. The first step in interpretation is to obtain an analysis tree from the input by a parsing algorithm that uses the grammar.

In a Bayesian architecture this is not necessary and not even a good idea. If one has a way to obtain probable hypotheses about what the input means the grammar can be used to check if these hypotheses explain the input. While it is possible to use symbolic parsing to get such hypotheses, the prediction from the context and the cues provided by the words seem the better and more road to such hypotheses.

The grammar formalisms that are most suitable for checking whether a hypothesis leads to the signal are formalisms that produce surface forms from semantic input by nearly deterministic methods. Such formalisms are quite common in the linguistic tradition. The oldest one is Panini's grammar for Sanskrit, early generative grammar (Chomsky 1957), generative semantics and systemic grammar (Mellish 1988) are good examples. A recent version is optimality theoretic production syntax that has the additional advantage of working on semantic fragments (Zeevat 2014b).

Such grammars are traditional in the sense that they answer the question of how to speak—given that one knows what one wants to say. They are also precursors of BNLSP in that they can be seen as adequate treatments of interpretation under the assumption that the selection problem is not important, an assumption that made sense until the current consensus about the ambiguity problem that emerged in the 1990s. Production-oriented syntax is also supported by an important linguistic argument: blocking by the elsewhere principle. This principle that goes back to Panini (Kiparsky 1973) states that a general rule of production can have exceptions due to a more specific rule. Examples include those in (22). The principle does not work unless one accepts the reality of production rules. While blocking effects have been found in interpretation, it does not seem possible to attribute them to a similar competition between interpretation rules.

- (22) *gooses* cannot mean *geese* (it has no meaning)
How late is it? cannot mean *What time is it?* as its Dutch or German counterparts would.

The grammar needs to be supplemented with a theory that predicts what people want to say in a given context. This aspect has been studied to some extent in natural language generation under the heading of text planning (Reiter and Dale 2000). Typically hearers do not have all the information that speakers have available. But what they may know still follows the priors about what may be the case. These priors together with text planning and grammar add up to a probability distribution over what the speaker will say, a probability distribution that will quickly adapt to the incoming cues.

Inverse Semantics

Semantic structure is standardly read off from syntactic structure, mostly by a version of the rule by rule hypothesis according to which a syntactic structure $f(x, y)$ means $f'(x', y')$ where f' is a semantic function interpreting f and x' and y' are the meanings of x and y respectively.

That there is something wrong with this picture can be seen from blocking as discussed just now. English is sufficiently like Dutch and German in syntax and lexical structure that the compositional mechanism that assigns the question “what time is it?” to the Dutch and German versions will also assign the same semantics to the English structure “How late is it?” But that is a wrong prediction. The English structure is clearly well-formed and meaningful (One can ask for the amount of delay of the bus with it for example) but it cannot be used for asking what time it is.

If production syntax is the right way to go for describing particular languages, it would also be the right way for deciding what meanings a syntactic structure can have by consulting the inverse F^{-1} of the function (or nearly functional relation) F that maps intentions in a context to their surface forms. Blocking can be part of the definition of that function F .

Inverse semantics is specifying semantics by specifying what surface form to use for a given semantic input. Blocking is also a good argument for inverse semantics.

The speaker is well-advised not to express the input by any utterance that is grammatical for it, since the input may not be the most probable interpretation of that utterance in the context in which case the hearer will get it wrong. Natural languages abound in optional markers that can save the day. These are particles, optional morphology, variation in lexical choice and word order variation. For a full discussion of these cases of self-monitoring, see Zeevat (2014b). It follows that grammar will be wrong in predicting an interpretation unless it is also the most probable interpretation.

This leads to the methodological principle (23) that can be called Jacobson’s principle.

No marked formal feature of the form can remain uninterpreted.

The principle does not follow from compositionality which allows any feature to have a trivial semantics and has difficulties with double expressions of the same semantic feature. It can however play a similar role in enforcing good practice in linguistic semantics: proper attention to all formal features of the linguistic structure in interpretation.

Pragmatics as Probability Maximation

The third area of linguistics in which Bayesian interpretation seems to point in a direction that is far removed from many accepted views is in pragmatics. This is often seen as an inference process that starts from the semantic interpretation and the context to give an enriched representation.

The simplest BNLSP view is that pragmatics is exhausted by probability maximation. This is not a new view. Jerry Hobbs has for many years now been defending “Interpretation by Abduction”. In this view an utterance is a fact to be explained and explanations are provided by a database of weighted abduction axioms. The best explanation is the proof of the utterance with the lowest weight. Weights should be inversely related to the probability of the axiom and Bayesian statistics is used to compute these weights from experience (Hobbs et al. 1990). That means that the abductive systems deals with pragmatics in the course of finding the most probable explanation of the utterance. Hobbs is able to reduce an impressive number of pragmatic phenomena to this mechanism and would be committed to the claim that the whole of pragmatics can be treated in this way.

BNLSP already makes probability maximation be the central concern of interpretation because it focuses on the selection task (integration or composition is needed to get likelihood maximation going). That means that many pragmatic effects are integrated under only minimal assumptions.

Some such assumptions are listed in (23).

- (23) a. pronouns refer to an object (likelihood will check agreement and pronoun class)
 b. presupposition triggers need their presupposition in order to do their semantic or pragmatic job
 c. a cardinal or a cardinality adjective answer a how-many question
 d. parameters such as topic, time and place are subject to self-monitoring.

(23a) forces the reconstruction of the speaker intention to contain a referent for the pronoun. If there is a choice, it should be the one that is the most probable.⁸ (23b) brings as discussed above both the satisfaction theory and the cumulative hypothesis. (23c) forces the construction of a how-many question and the assumption that its answer is as indicated. (23d) on topic forces conservativity in rhetorical relations, since resetting the parameter must be marked in the utterance under self-monitoring. This makes restatements and elaborations the unmarked rhetorical relations (the pivot topic is maintained) followed by lists (the local topic changes) and finally by contrastive pairs (the topic is reset). On time and place, it also gives rise to the typical changes of reference time and reference place in narrations.

BNLSP maximises probability and assumptions about the preconditions of lexical items and constructions and about the features for self-monitoring should do the rest.

8 Two Issues for BNLSP

It is correct to refer to vision for the priors that are needed, but that does not lead to a usable model of these priors. It is more likely that NLI will come up with such priors than that computer vision will come up with them. There are quite a number

⁸Zeevat (2010) gives a more detailed discussion.

of computational linguistic techniques that clearly contribute to the determination of the priors, as listed in (24).

- (24) probabilistic selection restrictions
 prediction of speech act/discourse relation
 prediction of intention
 world knowledge
 contextual knowledge
 conceptual knowledge
 causal modelling
 theorem proving
 simulation techniques (coherent visual scene, coherent plan).

The first three of these can be learnt from corpora and emerging automatic interpreters can be used to determine contextual knowledge. Theorem provers can be employed for borderline cases like the elimination of inconsistent interpretations (the prior is zero) or uninformative interpretations (the likelihood is very low). World knowledge can presumably be dealt with using techniques from information retrieval. That leaves one with the two most psychological aspects: simulation and conceptual knowledge. Benotti and Blackburn (2011) is a solid first step for the simulation of coherent plans and there are techniques from computer vision that can be exapted to visual scene coherence. Conceptual knowledge has to wait until more and better decompositional techniques are available in lexical semantics. Causal knowledge can in principle be learnt.

A major issue is that these different resources would have to work together for good results and that it is far from clear how to let the combination result in a single number for prior probability.

Another issue for working automatic BNLSP are emerging regularities. It happens to be the case—and not just by accident—that the objects that fill the agent thematic role in clauses are mostly animate. That means that a weak inference from being animate to being the subject of the clause is possible. This inference is an explanation for marking strategies that prevent the inference, e.g. passivising the sentence so that the inanimate subject is now marked as oblique. The empirical effects of the inference shows that animacy is in fact a cue for being the subject of the sentence.

It does not follow however that any stochastic regularity is a cue or that the inference as formulated above is the correct one. Should it not be merely concluded from animacy that the referent is relatively agentive in terms of Dowty (1990) proto-agent properties?

It is basically unclear in terms of which properties the brain does its grammar and learns the cues that it uses. Grammatical, lexical and typological generalisations are only imperfect windows on these matters, but they seem to be the only windows that there are.

As this paper hopes to have shown, there are many reasons to think that notwithstanding the existence of quite a number of well-entrenched views that need to be overcome before it will be the consensus view, BNLSP is highly promising for the study of language.

Liberman or Grice made the connection with production and simulation without realising that the selection problem is central for natural communication and without realising that it would be nearly intractable without simulated production.

References

- Benotti, L., & Blackburn, P. (2011). Causal implicatures and classical planning. *Lecture Notes in Artificial Intelligence (LNAI) 6967* (pp. 26–39). Springer.
- Bos, J. (2003). Implementing the binding and accommodation theory for anaphora resolution and presupposition projection. *Computational Linguistics*, 29(2), 179–210.
- Bowers, J., & Davis, C. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, 138(3), 389–414.
- Brozzo, C. (2013). Motor intentions: Connecting intentions with actions. Ph.D. thesis. Università degli Studi di Milano.
- Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.
- Clark, H. (1996). *Using language*. Cambridge: CUP.
- Dalrymple, M., Kanazawa, M., Kim, Y., Mchombo, S., & Peters, S. (1998). Reciprocal expressions and the concept of reciprocity. *Linguistics and Philosophy*, 21(2), 159–210.
- Dowty, D. (1990). Thematic proto-roles and argument selection. *Language*, 67(3), 547–919.
- Doya, K., Ishii, S., Pouget, A., & Rao, R. P. N. (Eds.). (2007). *Bayesian Brain: Probabilistic approaches to neural coding*. Cambridge: MIT Press.
- Garcia Odon, A. (2012). *Presupposition projection and entailment relations*. Ph.D. thesis, Università Pompeu Fabra.
- Gazdar, G. (1979). *Pragmatics: Implicature presupposition and logical form*. New York: Academic Press.
- Heim, I. (1983). On the projection problem for presuppositions. In M. Barlow, D. Flickinger, & M. Westcoat (Eds.), *Second Annual West Coast Conference on Formal Linguistics* (pp. 114–126). Stanford University.
- Hobbs, J., Stickel, M., Appelt, D., Martin, P. (1990). Interpretation as abduction. Technical Report 499, SRI International, Menlo Park, California.
- Hogeweg, L. (2009). *Word in Process. On the interpretation, acquisition and production of words*. Ph.D. thesis, Radboud University Nijmegen.
- Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: an account of the mirror neuron system. *Cognitive processing*, 8(3), 159–166.
- Kiparsky, P. (1973). “Elsewhere” in phonology. In S. R. Anderson & P. Kiparsky (Eds.), *A Festschrift for Morris Halle* (pp. 93–107). New York: Holt.
- Langendoen, D. T., & Savin, H. (1971). The projection problem for presuppositions. In C. Fillmore & D. T. Langendoen (Eds.), *Studies in Linguistic Semantics* (pp. 373–388). New York: Holt.
- Liberman, A., Cooper, F., Shankweiler, D., & Studdert-Kennedy, M. (1967). Perception of speech code. *Psychological Review*, 74, 431–461.
- Marrion, K., & Meyer, B. (1997). On the classification of visual languages by grammar hierarchies. *Journal of Visual Languages and Computing*, 8(4), 375–402.
- Mellish, C. (1988). Implementing systemic classification by unification. *Computational Linguistics*, 14(1), 40–51.
- Oaksford, M., & Chater, N. (2007). *Bayesian rationality*. Oxford: Oxford University Press.
- Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd ed., Vol. 29). Cambridge: MIT press.
- Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, 11(3), 105–110.

- Reiter, E., & Dale, R. (2000). *Building natural-language generation systems*. Cambridge: Cambridge University Press.
- Rizzolatti, G., & Craighero, G. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, 27, 169–192.
- Smolensky, P. (1991). Connectionism, constituency and the language of thought. In B. M. Loewer & G. Rey (Eds.), *Meaning in mind: Fodor and his critics* (pp. 286–306). Oxford: Blackwell.
- Soames, S. (1982). How presuppositions are inherited: A solution to the projection problem. *Linguistic Inquiry*, 13, 483–545.
- Stalnaker, R. (1979). Assertion. In P. Cole (Ed.), *Syntax and semantics* (Vol. 9, pp. 315–332). London: Academic Press.
- Van der Sandt, R. (1992). Presupposition projection as anaphora resolution. *Journal of Semantics*, 9, 333–377.
- Zeevat, H. (1992). Presupposition and accommodation in update semantics. *Journal of Semantics*, 9, 379–412.
- Zeevat, H. (2010). Production and interpretation of anaphora and ellipsis. *International Review of Pragmatics*, 2(2), 169–190.
- Zeevat, H. (2014a). Bayesian presupposition. MS, Amsterdam University.
- Zeevat, H. (2014b). *Language production and interpretation. Linguistics meets cognition* (CRISPI). Leiden: Jacob Brill.

Causal Bayesian Networks, Signalling Games and Implicature of ‘*More Than n*’

Anton Benz

Abstract We use causal Bayesian networks for modelling experimental data on scalar implicatures of numerals modified by the comparative quantifier ‘*more than*’. The data are taken from the study of Cummins et al. (2012). They showed that subjects infer upper bounds for the true number n when ‘*more than n*’ has been used. They also showed that the difference between upper and lower bounds differs with the roundness of n . We show how the data can be explained by a production model of the speaker. We are interested in the architecture of such a model, and argue that causal Bayesian networks provide a natural and promising approach.

Keywords Quantity implicature · Roundness · Modified numerals · Sequence rule · Game theoretic pragmatics · Bayesian production model

1 Introduction

Causal Bayesian networks (Pearl 2000) provide a natural framework for high-level descriptions of the interaction of different (cognitive) modules involved in communication. In the simplest case, a Bayesian network is a sequence of random variables in which the value of each variable conditionally depends on the value of the preceding variable. For example, the speaker’s intended meaning depends on his world knowledge, the signal which he produces depends on the intended meaning, the hearer’s interpretation depends on the signal she receives, and subsequent actions she chooses

This work was supported by Bundesministerium für Bildung und Forschung (BMBF) (Grant Nr. 01UG1411). It greatly profited from discussions at the ESSLLI workshop on *Bayesian Natural Language Semantics and Pragmatics* organised by Hans-Christian Schmitz and Henk Zeevat in 2013. In particular, I thank Chris Cummins, Manfred Krifka, Stephanie Solt, and Clemens Mayr for their insightful comments, and, last but not least, the two anonymous reviewers, as well as the editors of this volume.

A. Benz (✉)

Centre for General Linguistics, ZAS, Berlin
e-mail: benz@zas.gwz-berlin.de

may depend on the information gained from the signal. The state of the world, the intended message, the signal, its interpretation, and the subsequent action can all be seen as values of random variables which are causally dependent on each other in a linear order. The Bayesian representation of the communication process can be turned into a game by assuming that the random variables are independent players which have to solve the coordination problem of successful communication. Games of this type have been introduced under the name of *error models* in Benz (2012). This model provides the background for the causal models which will be applied to experimental data on the comparative quantifier ‘*more than n*’ (Cummins et al. 2012).

A key prerequisite for drawing the quantity implicature from ‘*three*’ to ‘*not more than three*’ in e.g. ‘*Nigel has three children*’ is the assumption that the speaker knows the exact number of Nigel’s children. This led to the assumption that ‘*more than n*’ constructions generate no implicatures as the comparative quantifier ‘*more than*’ signals that the speaker lacks sufficient knowledge for making a more precise statement (Krifka 1999). However, experimental results from Cummins et al. (2012) show that scalar implicatures are available from these constructions. For example, the size of the interval defined by the estimated lower and upper bounds for the number of people getting married is much higher in (1b) than in (1a):

- (1) (a) More than 90 people got married today.
 (b) More than 100 people got married today.

Cummins (2013) proposes an optimality theoretic model of the experimental findings. It is a one-sided production model with free constraint ranking. The input–output pairs are the speaker’s intended meaning and its linguistic realisation. Implicatures are calculated by inferring the speaker’s meaning from output. The model has no theory of information states. We show that this, together with the ‘*the winner takes it all*’ rule of optimality theory, means that it cannot fully account for ‘*more than*’ implicatures with round numerals.

As did Cummins, we derive implicatures from a production model. However, we represent it in the form of a causal Bayesian network: The first random variable represents the state of the world, in example (1) the true number of people getting married, the second variable represents the speaker’s information state, in our example an interval in which the speaker expects the true number to lie, then the third variable describes the possible linguistic output, and the fourth variable the hearer’s interpretation of the output. The key step is the definition of the possible speaker information states. We apply here the unmodified sequence rule of Jansen and Pollmann (2001) such that an open interval (n, m) is a possible information state if m immediately follows n in a sequence of natural numbers defined by Jansen and Pollmann’s rule.¹ Although this use of the sequence rule is not unproblematic from a cognitive perspective, we show that the model predicts the experimental results of Cummins et al. (2012) about implicatures from ‘*more than n*’ with round n .

¹It defines sequences of *round* numbers in different scales, e.g. ‘5, 10, 15, . . .’, ‘10, 20, 30, . . .’ or ‘50, 100, 150, . . .’ The rule says that numbers appearing in approximation contexts like ‘*between n or m*’ are always adjacent elements of such sequences, e.g. (20, 30), (50, 100), or (100, 150). See the definition in (10) and examples in (11).

In the experiments of Cummins, Sauerland and Solt, subjects were asked to estimate the lower and upper bounds of the interval in which the true number of people has to lie. As the use of a modified numeral ‘*more than n*’ does not allow for an unambiguous reconstruction of the upper bound of the speaker’s information state, the game theoretic model proposed in (Benz 2012) implies that the upper bound cannot be part of the intended speaker meaning. This raises the question, how these implicatures about upper bounds are related to scalar implicature. Based on our production model, we argue that inferences about upper bounds are not communicated non-natural information in the sense of Grice (1989, Chap. 14).

2 Scalar Implicature

Normally, an utterance of (2) will communicate that *some but not all* of the students failed:

(2) Some of the students failed

Grice (1989, Chap. 2) distinguished between *what is said* (*some, possibly all*), i.e. the semantic meaning of the utterance, and what is *implicated* (*not all*). Both aspects are part of the intended speaker meaning. The implicated meaning is inferred from the assumption that the speaker adheres to the so-called cooperative principle, and a number of *maxims*. In short, they ask the speaker to be cooperative, and to say as much (Quantity) as he knows (Quality) as long as it is relevant (Relation). In addition, his presentation should be short, orderly, perspicuous, etc. (Manner). For Grice, an implicature is a proposition which one can infer the speaker to believe in order to maintain the assumption of cooperativity and adherence to the maxims. Implicatures are, therefore, inferred after semantic meaning, hence pragmatic reasoning follows semantic reasoning, except in cases of disambiguation and saturation of anaphors and indexicals, which may involve pragmatic reasoning and precede semantic interpretation. That implicatures are not part of semantic meaning can be seen from their *cancellability*, i.e. they can be explicitly negated by the speaker as in ‘*Some, in fact, all of the students failed*’.

The implicature from ‘*some*’ to ‘*not all*’ is explained by adherence to the maxim of quantity: if the speaker had known that all failed, he would have said so (Quantity). So, he cannot know that all failed. If it is assumed that he knows whether all failed (Expert), then the implicature that not all failed follows.

This reasoning has been systematised by so-called *neo-Griceans*.² In the previous example, we find a *scale* (*all, some*) and a sentence frame $A(x) \equiv$ ‘*x of the students failed*’ such that $A(\textit{all})$ semantically implies $A(\textit{some})$ but not the other way round. As we have seen, an utterance of the weaker $A(\textit{some})$ implicates that $\neg A(\textit{all})$. This generalises as follows: given a *scale* $\langle e_1, \dots, e_n \rangle$ of expressions of similar complexity which can be filled into a sentence frame $A(\cdot)$ such that for $i < j$

²See Levinson (1983, Chap. 3) for a summary.

$A(e_i) \rightarrow A(e_j)$ and $A(e_j) \not\rightarrow A(e_i)$, then an utterance of a weak alternative $A(e_j)$ implicates that all stronger alternatives $A(e_i)$, $i < j$, are false. In the following, we write $A(e_j) \text{+>} \neg A(e_i)$ if $A(e_j)$ implicates $\neg A(e_i)$.

According to the standard theory, numerals also define a scale $\langle \dots, \textit{five}, \textit{four}, \textit{three}, \textit{two}, \textit{one} \rangle$. This predicts that a use of ‘*three*’ will normally implicate that the same sentence with ‘*three*’ replaced by ‘*four*’ is false:

- (3) A: How many children does John have?
 B: John has three children.
 +> John does not have four children.

This assumes that a numeral as e.g. ‘*three*’ is semantically equivalent to ‘*three or more*’. Hence, the sentence ‘*John has three children*’ is true semantically even if John has four, five or more children. This assumption is supported by examples as the following:

- (4) (a) A: Does John have three children?
 B: Yes, he has even four!
 (b) All of the students solved three problems, some even five or six.

If ‘*three*’ would mean ‘*exactly three*’, then B could not answer with ‘*yes*’ in (4a), and (4b) would be a contradiction.

One of the major issues of the standard theory is the question: what exactly constitutes a scale? Not just any set of phrases $\langle e_1, \dots, e_n \rangle$ ordered by logical strength can form a valid scale. It is often assumed that the e_i s have to be of equal complexity. But even this condition is no guarantee for the e_i s forming a scale. Modified numerals are a case in point:

- (5) (a) John has *more than two* children.
 (b) John has *more than three* children.
 (c) John has *more than four* children.

As ‘*more than $n + 1$* ’ implies ‘*more than n* ’, then, assuming that $\langle \textit{more than four}, \textit{more than three}, \textit{more than two} \rangle$ forms a scale, an utterance of (5b) would implicate that John does not have more than four children, and, hence, that he has exactly four. Intuitively, the use of the modifier ‘*more than*’ indicates that the speaker does not know the exact number. This was taken as evidence that comparatively modified numerals do not form scales and, hence, do not give rise to implicatures (Krifka 1999; Fox and Hackl 2006).

3 Inferences from Comparatively Modified Numerals

Although comparatively modified numerals do not give rise to the neo–Gricean scalar implicature, they allow for some conversational inferences. The following examples are taken from either Cummins (2011), Cummins et al. (2012), or Cummins (2013). Consider:

(6) London has more than 1000 inhabitants.

As an answer to the question ‘*How many inhabitants has London?*’, (6) is obviously underinformative. It conveys the impression that London is a small city, and a naïve addressee would clearly exclude the possibility that London has a million, let alone 10 million inhabitants. A similar effect is observable in the following example:

- (7) (a) More than 80 people got married today.
 (b) More than 90 people got married today.
 (c) More than 100 people got married today.

The sentences give rise to different expectations about the size of the interval that must contain the true number of people married that day. It should be larger for (7c) than for (7b). Whereas (7c) seems to be compatible with the real number being 120, (7b) suggests that less than 100 got married. Also (7a) seems to suggest a smaller interval than (7c). In (8), the expected interval becomes even smaller. One would expect that the true number is either 97, 98, or 99.

(8) More than 96 people got married today.

In addition to the inferences about the interval that must contain the true value, comparatively modified numerals also give rise to certain *ignorance* implicatures. For example, (8) conveys that the speaker does not know whether more than 97 people got married that day. It has been claimed that, in general, ‘*more than n*’ gives rise to the clausal implicatures ‘*possibly n + 1*’ and ‘*possibly more than n + 1*’ (Cummins 2013).

Central for the following discussion are the experiments by Cummins, Sauerland and Solt (2012), who provide support for the claim that the use of ‘*more than n*’ invites additional inferences about the possible range of values. In their experiments, test subjects had to answer a questionnaire on the Amazon Mechanical Turk platform. The items consisted of a short statement containing a modified numeral followed by a question which asked subjects to estimate the true number. An example is shown in Fig. 1.

Information A newspaper reported the following.

“[Numerical expression] people attended the public meeting about the new highway construction project.”

Question Based on reading this, how many people do you think attended the meeting?

Between _____ and _____ people attended [range condition]
 _____ people attended [single number condition]

Fig. 1 Example item from Cummins et al. (2012, p. 146)

	Low	High		Low	High
$n = 100$	100 (40)	150 (24)	$n = 93$	93 (32)	100 (34)
	101 (28)	125 (12)		94 (29)	95 (14)
		120 (8)		90 (14)	99 (7)
		200 (7)			150 (6)
		1000 (5)			125 (5)
$n = 110$	110 (46)	120 (28)			
	111 (31)	150 (24)			
	100 (9)	200 (7)			
		119 (6)			
		130 (5)			
		115 (5)			

Fig. 2 Experimental results from Cummins et al. (2012)

Cummins et al. considered the modifiers ‘*more than*’ and ‘*at least*’. We are only concerned with the cases in which the numerical expression of the test item was ‘*more than n* ’. Cummins et al. further distinguished two types of estimates: in the *single number* condition, subjects were required to provide a natural number; in the *range* condition, they had to provide a lower and an upper bound (*between m and n*) for the true number. In the following, we only consider the results for the range condition, which are shown in Fig. 2.

The ‘*Low*’ column shows the different lower bounds for ‘*more than n* ’, followed in brackets by the number of subjects providing the lower bound. For example, 40 subjects gave a lower bound of 100 for the ‘*more than 100*’ condition, and 28 subjects the lower bound 101. The ‘*High*’ column shows the different upper bounds.³ The experimental data overall support the introspective data described before. However, they also show that lower and upper bounds tend to be *round* numbers. They are always round if n is round. Even for $n = 93$, which is not round, the bounds are round numbers, or clearly chosen due to their proximity to a round number (99); the only exceptions are 93 and 94.

To sum up, there is evidence that comparatively modified numerals invite inferences which bear some similarity to standard scalar implicature. In particular, subjects feel entitled to make guesses about the intervals which contain the true values.

4 An Optimality Theoretic Model

In (Cummins et al. 2012), the implicature from modified numerals were explained as a kind of weaker implicature without being too explicit about the exact status of these weaker implicature. In a theoretical article, Chris Cummins’s proposed an optimality theoretic model of the experimental results (Cummins 2013).

³The table provides no indication of how lower and upper bounds were paired. So, the estimated intervals cannot be reconstructed.

Optimality Theory (OT) was introduced to phonology in the early 1990s (Prince and Smolensky 1993). It describes grammar as a symbolic system in which surface forms are considered solutions to a constraint optimisation problem. The constraints are violable and hierarchically ranked according to their strength such that the stronger constraints strictly dominate the weaker constraints. In a production model, the linguistic surface forms are the optimal output forms for an underlying input form. Cummins (2013) proposes an OT production model in which the constraints are special instances of Gricean maxims. The output forms are the modified numerals ‘more than *n*’, for example, ‘more than 50’, ‘more than 60’, ‘more than 70’, etc. The input from which these surface forms are generated are the speaker’s information states.

In the experiments of (Cummins et al. 2012), subjects first read a text containing the modified numeral ‘more than *n*’, and had then to estimate lower and upper bounds for the true number. For the experimental item shown in Fig. 1, the text may have said, for example, that ‘More than 100 people attended the public meeting about the new highway construction project.’ Cummins’s OT model is a model of the author producing the text. It tells us for which speaker information states the text is an optimal output. The assumption is that, as the OT model is common knowledge, the subjects can use this knowledge for inferring the underlying information state, and thereby for inferring the lower and upper bounds.

The optimal output for an information state is calculated with the help of a tableau as that shown in (9). The left upper corner of (9) shows the information state of knowing that between 101 and 109 people came to the meeting. In the column under the information state all possible outputs are listed. Cummins assumes that only modified numerals enter the competition. The constraints are listed in the top row.⁴ First, there is the QUALITY constraint, which asks the speaker to say only what he believes to be true. This constraint is assumed never to be violated. It is therefore always to be ranked highest. Next come the QUANTITY constraint, which asks the speaker to choose the most informative output, and the SALIENCE constraint, which asks the speaker to choose round numbers. It is assumed that different speakers can come with different rankings of QUANTITY and SALIENCE. Hence, from the hearer perspective, these constraints can be arbitrarily ordered. In (9), this is indicated by the dashed vertical line between the last two columns.

(9)

	101–109	QUALITY	QUANTITY	SALIENCE
	more than 110	*!		**
	more than 100			
	more than 90		*!	**
	more than 80		*!*	

⁴The full model also contains a constraint preferring the re-use of previously mentioned (*primed*) numerals. For the special problem of implicatures from ‘more than *n*’ for *round* numerals *n*, this constraint can be left out of consideration. It becomes relevant for non-round numerals, as in ‘more than 93’. We also changed constraint naming: Cummins’s tableaux do not include (Quality), and we call the quantity maxim by its conventional name of QUANTITY instead of INFORMATIVITY, which Cummins chose.

As the speaker believes that the true number of people coming to the meeting does not exceed 109, it would violate the QUALITY constraint to say that *'More than 110 came'*. This is indicated by the * in the column below QUALITY. All other numerals satisfy the constraint. The numerals 90 and 80 are less informative than 100. This is indicated by the * in the column below QUANTITY. The violation is stronger for 80 than for 90, hence, 80 receives two stars and 90 only one star. The SALIENCE constraint measures the *roundness* of the chosen numeral. In the given example, 100 is rounder than 80, and 80 is rounder than 90 and 110. The violation of SALIENCE, shown in the last column, grows stronger with decreasing roundness.

An output *F* wins over another output *G*, if the first constraint for which they differ is violated more strongly by *G* than by *F*. *'More than 100'* is the only output which satisfies all constraints. It therefore wins the competition. The '!' in other rows marks the first violation which makes the corresponding output lose the competition with the winning form.

Cummins's measure of roundedness is based on the corpus studies of Jansen and Pollmann (2001). As an indicator of the degree of roundness Jansen and Pollmann considered the frequencies of numerals in approximation contexts created by, for example, Dutch *'ongeveer'*, German *'etwa'*, French *'environ'*, and English *'about'*. The more frequent a numeral is in such contexts, the *rounder* it is assumed to be.

Cummins's OT model makes predictions about which numerals *n* are optimal in *'more than n'* contexts. He provides an explanation for why round numerals allow for a larger range of upper bounds. He also claims that his model explains the ignorance implicature from *'more than n'* to *the speaker believes it possible that n + 1* ($\diamond n + 1$).

The wider range of upper bounds for round numerals is explained as a consequence of the free ranking of the QUANTITY and SALIENCE constraints. For example, if the speaker knows that the true number of people attending some event lies in the open interval (150, 200), then, assuming that 100 is rounder than 150, it follows that a speaker ranking SALIENCE over QUANTITY will prefer the less precise *'more than 100'* over *'more than 150'*. If, in contrast, the speaker knows that the true number lies in some interval above 100, for example in the interval (110, 120), then there is no constraint ranking which could make the choice of *'more than 90'* optimal. This restricts the allowed range for the true number to (90, 100) for *'more than 90'*, and allows a wide range for *'more than 100'*. However, if the belief interval is (150, 200) and SALIENCE is ranked higher than QUANTITY, then the model not only predicts that *'more than 100'* wins over *'more than 150'*, but also that *'more than 10'* wins over *'more than 100'*, as 10 is rounder than 100.⁵ Hence, the speaker should choose *'more than 10'*. The model also predicts, as there is no higher number rounder than 100, that *'more than 100'* wins over all alternatives *'more than n'* for numbers *n* greater than 100. This is a consequence of the *'the winner takes it all'* property of OT models. It entails that the most salient number *N*, whichever it is, has no upper bound, and all other numbers greater than *N* can only be used if QUANTITY is ranked

⁵In fact, this depends on which number is considered to be roundest. For example, the frequency graph in (Jansen and Pollmann 2001, p. 193) shows an even higher peak for 20 than for 10. However, pure frequencies may not be the best criteria of roundness.

over SALIENCE. The model entails, therefore, that the ignorance implicature only exist for numbers which are preceded by a more salient number. For the maximally salient numbers, the implicature from ‘*more than N*’ to $\diamond N + 1$ does not hold true. If someone who believes the true number to lie in (150, 200) can utter ‘*more than 100*’, then the addressee cannot infer from this utterance that the speaker believes 101 to be possible.⁶

As the model imposes no restrictions on the possible speaker information states, no inferences about upper bounds can be drawn if QUANTITY is ranked over SALIENCE. For example, if all intervals $(110, L_u)$ are admissible information states of the speaker, then from an utterance of ‘*more than 110*’ no restriction on the upper limit L_u can be inferred. It could be arbitrarily large. Cummins obviously assumes some restrictions on possible information states, but they are not made explicit in his model.⁷

We sympathise with Cummins’s approach of explaining the observed implicatures from a production model. However, the model has shortcomings which warrant considering alternatives. In particular, it seems necessary to have a model that explicitly describes the set of admissible information states. Without it no inferences can be made about upper bounds. We also will drop the optimality theoretic framework.⁸ We prefer a representation which describes the communication process and the experimental situation more directly. As we are mainly interested in the general architecture of such a representation, we consider only a limited subset of the experimental results. For example, we leave out of consideration non-round numerals, as, for example, 93.⁹ The main question in the next section is how to account for the influence of roundedness on the estimated difference between lower and upper bounds even if it is assumed that the speaker tries to be as informative as possible, i.e. even if, in OT terms, the speaker ranks QUANTITY over SALIENCE.

5 A Causal Bayesian Network Model

Our aim is to explain the experimental data shown in Fig. 2 on ‘*more than 100*’. The main effort will be to set up a model of the author producing the text appearing in the questionnaire shown in Fig. 1, for example, the text ‘*More than 100 people attended the public meeting about the new highway construction project.*’

In Fig. 3, a model of the experimental situation which starts with a reported (imaginary) event and concludes with the test subject’s answer is shown. Each node in this

⁶This reveals a (minor) inconsistency in Cummins’s argumentation (2013), as he claims that the ignorance implicature always holds true.

⁷Excluding examples with anaphoric uses, Cummins always considers information states (n, m) for which n and m are *adjacent* round numbers. Making this aspect explicit would presumably reveal the same application of Jansen and Pollmann’s sequence rule as in our own model.

⁸It may also be debated that an OT model with free constraint ranking is still in the spirit of optimality theory.

⁹For these numerals, Cummins has also to apply a different model with an additional constraint.

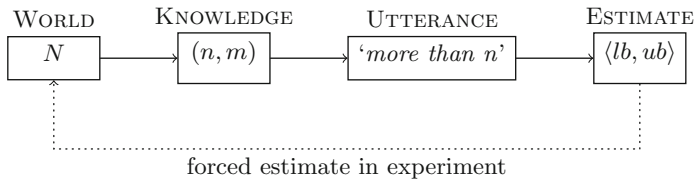


Fig. 3 A model of the Cummins, Sauerland, and Solt experiment (2012)

sequence corresponds to a random variable in a causal Bayesian network. The values of the first variable named **WORLD** are natural numbers N , those of the third variable named **UTTERANCE** are sentences containing ‘more than n ’ for some natural number n , and the values of the last variable named **ESTIMATE** are pairs $\langle lb, ub \rangle$ of lower and upper bounds for the true number N . In order to complete this model, it has to be known which values the variable named **KNOWLEDGE** can take, and it has to be known what the conditional probabilities of the various values are.

We will identify the knowledge states with open intervals (n, m) using Jansen and Pollmann’s unmodified *sequence rule* (Jansen and Pollmann 2001). It is obviously not possible to determine the conditional probabilities. Fortunately, this is not necessary as we will see once we know how to define implicature. We address this issue next.

5.1 Implicature in Bayesian Networks

In a Bayesian network, the different random variables carry information about each other. For our purposes, we can assume that communication is describable by a linear sequence of random variables $(\mathcal{X}_i, P_i)_{i=1}^n$ with conditional probabilities $P_i(x_i|x_{i-1})$ defined for $x_i \in \mathcal{X}_i$ and $x_{i-1} \in \mathcal{X}_{i-1}$. With the product probability $P(x_1, \dots, x_n) = P_1(x_1) \times \dots \times P_n(x_n|x_{n-1})$, we can define the probability that the i th variable has a value in some set X given that the j th variable has a value in Y as follows¹⁰:

$$P_{i|j}(X|Y) := \frac{P(\pi_i^{-1}(X) \cap \pi_j^{-1}(Y))}{P(\pi_j^{-1}(Y))} \quad \text{with } \pi_k(x_1, \dots, x_n) = x_k. \quad (1)$$

In a linear Bayesian network describing communication, one variable will describe the utterances the speaker produces. Let us assume it is the j th variable, and that the utterance is F . Then we can say that the utterance of F *implicates* that some other variable \mathcal{X}_i has a value in X iff $P_{i|j}(X|F) = 1$.

It follows that, in order to determine implicature, it is only necessary to know which values x_i of a variable \mathcal{X}_i are *consistent* with an utterance of F . Hence, it suffices to know for which x_i $P_{i|j}(x_i|F) > 0$.

¹⁰Whenever a probability appears in the denominator, we implicitly assume that it is greater than zero. If X or Y is a singleton set, for example $Y = \{y\}$, the brackets around y are omitted in argument positions of probability distributions.

This definition is justified by Grice’s contention (1989, p. 86) that ‘*what is implicated is what it [sic!] is required that one assume a speaker to think in order to preserve the assumption that he is observing the Cooperative Principle (and perhaps some conversational maxims as well), ...*’ The most obvious interpretation is that what is implicated by an utterance of F is the information that is carried by F about values of variables representing the speaker’s beliefs.¹¹

5.2 Knowledge States and the Sequence Rule

All estimated lower and upper bounds shown in Fig. 2 are round numbers, or differ only by a value of one from a round number. As we have seen, the speaker’s preference for round numbers is also a key element of Cummins’s OT model. In the Bayesian network in Fig. 3, the (imaginary) speaker’s first task is to estimate the true number of people taking part in some event. We assume with Cummins’s that this estimate is some interval. More specifically, we assume that it is an open interval (n, m) with at least two elements defined by two round numbers n and m . To further restrict the admissible pairs of numbers n, m , we make use of the corpus study by Jansen and Pollmann (2001) who found that pairs of numbers n, m in the approximative construction ‘*about n or m*’ follow the following rule:

- (10) *Sequence rule:*¹² For context ‘*about n or m*’, the numbers n, m follow each other directly in a sequence:

$$L_{k,l} = \{i (k \times 10^l) \in \mathbb{N} \mid i \in \mathbb{N}\} \text{ with } k = 1, 2, 1/2, 1/4, \text{ and } l \geq 0. \quad (2)$$

The following list in (11) shows all sequences which are relevant to us.

- (11)

$L_{1,0} = 1, 2, 3, 4, 5, \dots$	$L_{2,1} = 20, 40, 60, 80, \dots$
$L_{2,0} = 2, 4, 6, 8, 10, \dots$	$L_{1/4,2} = 25, 50, 75, \dots$
$L_{1/2,1} = 5, 10, 15, 20, \dots$	$L_{1/2,2} = 50, 100, 150, 200, \dots$
$L_{1,1} = 10, 20, 30, 40, 50,$	$L_{1,2} = 100, 200, 300, 400, 500,$
\dots	\dots

¹¹This definition deviates from common usage in so far as it includes literal meaning in implicated meaning, whereas what is implicated is generally understood to be the meaning that is communicated but not literally expressed. For simplicity, we use this inclusive definition, which can be easily modified to bring it more in line with common usage. We may mention that Grice’s original characterisation of conversational implicatures also does not exclude literal meaning (Grice 1989, p. 30f).

¹²This is the *unmodified* sequence rule. Jannsen and Pollmann (2001) see no justification for the inclusion of $k = 1/4$. However, we can ignore this modification for our purposes. For $k = 1/4$, l must be greater than 1 to make sure that all elements of $L_{1/4,l}$ are natural.

Let us write $(n, m) \sqsubset L_{k,l}$ if n and m are consecutive numbers in sequence $L_{k,l}$. We assume that (n, m) is a possible information state if n and m are consecutive numbers in some sequence $L_{k,l}$. In addition, we assume that the speaker's estimate is correct, i.e. that the true number N is an element of (n, m) . Translated into conditional probabilities $P((n, m)|N)$, this means:

$$P((n, m)|N) > 0 \text{ iff } \exists k \exists l > 0 : (n, m) \sqsubset L_{k,l} \wedge N \in (n, m). \quad (3)$$

As mentioned before, it suffices for implicature calculation to know when probabilities are greater zero. In order to rule out intervals with one or zero elements, we assume that $l > 0$.

5.3 Speaker's Strategy and Experimental Data

The last parameter to be fixed for the model in Fig. 3 is the probability with which the speaker chooses 'more than n ' given that his knowledge state is (n', m') . In game theoretic terminology, this conditional probability can be called the speaker's *strategy*. We make the assumption that the speaker can only produce $F_n =$ 'more than n ' if $n' = n$:

$$S(F_n|(n', m')) > 0 \text{ iff } n = n'. \quad (4)$$

The situations represented by condition (4) correspond to those situations in which QUANTITY is ranked higher than SALIENCE in Cummins's optimality theoretic model. As we have seen, for these cases, the optimality theoretic model cannot make predictions about upper bounds. In Sect. 5.4, we will shortly discuss cases in which the speaker could have been more informative; for example, cases in which he could have been more informative because he knew that exactly 93 people came to an event, but nevertheless chose the less informative lower bound 'more than 90'.

This completes the set-up of our model. The probabilities with which the test subjects choose their estimates and the nature of the estimates are provided by the experimental data.

The table in (12) lists the belief intervals which have probability greater zero given that the speaker uttered F_n for the numbers $n = 80, 90, 100$. To each interval, the sequence to which the interval belongs is added. More precisely, if \mathcal{F} is the name of the random variable representing utterances, and Θ the random variable for the speaker's belief, then the intervals (n, m) listed in table (12) are those for which $P_{\Theta|\mathcal{F}}((n', m')|F_n) > 0$.

(12) F_n	belief intervals	
'more than 80'	$L_{1/2,1}$ (80, 85)	$L_{1,1}$ (80, 90)
	$L_{2,1}$ (80, 100)	
'more than 90'	$L_{1/2,1}$ (90, 95)	$L_{1,1}$ (90, 100)
'more than 100'	$L_{1/2,1}$ (100, 105)	$L_{1,1}$ (100, 110)
	$L_{2,1}$ (100, 120)	$L_{1/4,2}$ (100, 125)
	$L_{1/2,2}$ (100, 150)	$L_{1,2}$ (100, 200)

Given our definition of implicature, this table says that an utterance of 'more than 80' implicates that the speaker believes that the true number is an element of (80, 85), (80, 90), or (80, 100). Likewise, an utterance of 'more than 100' implicates that the speaker estimates the true number to lie in one of the intervals with lower bound 100 and upper bounds 105, 110, 120, 125, 150, and 200. Accordingly, following from the assumption that the speaker's beliefs are true, an utterance of 'more than 100' implicates that the true number lies in the interval (100, 200), and an utterance of 'more than 80' that it lies in the interval (80, 100). The model accounts for the different ranges of estimated intervals without having to rank the preference for round numbers over the preference for information. Equation (4) entails that the speaker chooses the most informative version of 'more than n '. The model also accounts for the ignorance implicature 'more than n ' implicating possibly $n + 1$.

It is difficult to say how test subjects construed the task in the questionnaire. They were not explicitly asked for *implicatures*, only for estimated lower and upper limits of the true number N . It is plausible to assume that they provided probable and safe limits within which the speaker's belief intervals lie. The experimental results for 'more than 100' are repeated in (13). We find a good match between the predicted set of possible upper bounds and elicited upper bounds. The only predicted upper bounds which cannot be found in the experimental data are 105 and 110, and the only upper bound not predicted is 1000.

(13)	Low	High
$n = 100$	100 (40)	150 (24)
	101 (28)	125 (12)
		120 (8)
		200 (7)
		1000 (5)

A plausible reason for the missing 105 and 110 is that the resulting intervals would be too small. As test subjects had no information about speakers, the belief intervals (100, 105) and (100, 110) would have been too unlikely to be the only speaker types. On the other side, the estimated upper bound of 1000 may be the result of a special safety strategy of test subjects.¹³

¹³In the cognitive literature it is also argued that the distances on the mental number line are rescaled according to some logarithmic rule (Dehaene 2011). This may have an effect on choosing upper bounds which is not accounted for in our model but may explain why some people estimate the true number to lie between '100 and 1000.'

5.4 Extensions of the Basic Model

In the basic model we only considered the case in which the assumed speaker of the test sentence believes that the true number N can lie anywhere in an open interval (n, m) . When he produces an utterance with ‘*more than*’, then he always chooses n as the lower bound. This was encoded in Eq. (4). In this section, we consider other strategies which may lead to utterances containing ‘*more than* n ’. The reasoning is admittedly even more speculative than in the previous sections. The aim is to show that the constraints over possible estimates for the interval containing the true number N that may be produced by the *test subjects* remain valid also when considering other speaker strategies which might have produced the test sentences. A second aim is to show how the sequence rule *might* be used to capture a variety of other reasons leading to utterances with numerals modified by ‘*more than*’.

We first consider the following possibility: If the speaker says that ‘*More than 90 people attended the public meeting about the new highway construction project*’, it is conceivable that he knows that exactly 93 people attended but chose the round number 90 as lower bound for ‘*more than* n ’. In Cummins’s model one has to assume here that SALIENCE outranks QUANTITY. If the speaker knows that exactly N people attended, the dominance of SALIENCE over QUANTITY means that the speaker must choose the roundest number n smaller or equal to N . However, as we have seen, the restriction to the *roundest* number smaller than N seems to be too strong a requirement. We will now show how a prediction about possible lower bounds n can be derived from the sequence rule. It can be assumed that the different sequences $L_{k,l}$ resulting from different choices of numbers $k = 1, 2, 1/2, 1/4$ and $l \in \mathbb{N}$ represent different levels of precision with which the speaker wants to describe the true situation. In the following model, it is assumed that the speaker, who knows that exactly N people attended, first chooses a level of precision $L_{k,l}$ with which he wants to describe the situation, and then the largest number n in $L_{k,l}$ which is smaller than N as a lower bound for ‘*more than* n ’. The following constraint, which modifies the conditions stated in Eqs. (3) and (4), captures this idea. It says that $F_n =$ ‘*more than* n ’ is a possible utterance if there is an interval (n, m) of adjacent numbers in some sequence L_{kl} which contains N :

$$S(F_n|N) > 0 \text{ iff } \exists k \exists l > 0 \exists m > n : (n, m) \sqsubset L_{k,l} \wedge N \in (n, m). \quad (5)$$

For $N = 93$, this rule predicts that the speaker can produce the following sentences:

- (14) (a) More than 90 people attended. $((n, m) = (90, 95), (90, 100))$
 (b) More than 80 people attended. $((n, m) = (80, 100))$
 (c) More than 75 people attended. $((n, m) = (75, 100))$
 (d) More than 50 people attended. $((n, m) = (50, 100))$

These are all possibilities. Numbers like $n = 20$ or $n = 10$ are ruled out because there is no sequence $L_{k,l}$ for which there exist intervals $(20, m)$ or $(10, m)$ that would contain 93 as an element. What is important with respect to the experimental data of

Cummins et al. (2012) is that the strategy defined in (5) does not introduce expected intervals (n, m) different from the intervals introduced by the basic rule in (4): If the addressee hears that e.g. ‘more than 80 people attended’, then she can infer that the true number is an element of $(80, 100)$. So, even when the addressee does not know whether the speaker’s information state is an interval or an exact number, she should arrive at the same estimates.

Let us consider one further scenario. It is conceivable that the speaker counted the people attending the meeting and stopped counting at, say, 91. He may be certain that at least two more attended, and, maybe, up to a hundred. When asked, how many were there, he may well answer ‘more than 90’. Also for such a scenario we can derive a plausible model from Jansen and Pollmann’s sequence rule. In our example, the speaker knows that at least 93 attended. So let N be any number, not necessarily round, for which the speaker verified by counting that at least N attended. Then the following strategy produces plausible ‘more than n ’ utterances with round lower bounds n :

$$S(F_n | (N, m)) > 0 \text{ iff } \exists k \exists l > 0 : (n, m) \sqsubset L_{k,l} \wedge N \in (n, m). \tag{6}$$

The predicted intervals for $N = 93$ are the same as in (14). This also means that the hearer, when confronted with a ‘more than n ’ message, has no reason to change her estimates for the interval containing the true number.

As mentioned before, the modifications presented in this section are increasingly speculative. However, they demonstrate how the model can be extended to cover more scenarios in which the speaker can produce a comparatively modified numeral. In terms of linguistic depth, the achievements may seem modest. Even with the extensions presented, the model does not account for vague interpretations of unmodified numerals, as, for example, in ‘A hundred people came to the talk’, which can mean that about 100 people came. It also does not account for an utterance of ‘more than 93’. For these examples, the model has to take into account additional parameters, e.g. previously mentioned numerals, and contextually required precision. This leads to nontrivial extensions, which we leave to future research.

Before we end this section, we provide an overview of how roundness can influence the experimental results in the setting of Cummins et al. (2012). In our basic model, roundness entered the definition of conditional probabilities of belief intervals. However, roundness can also influence the choice of utterances and the interval estimation by the test subject. These possibilities are shown in Fig. 4.

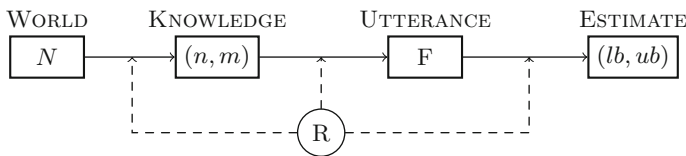


Fig. 4 Influence of roundness R on probabilities

If roundness determines the formation of belief intervals, then the (imaginary) speaker of the ‘*more than n*’ sentence is restricted to belief states (n, m) with round neighbours in a Jansen and Pollmann sequence, as assumed in our basic model. If roundness enters the choice of utterance, then the speaker may produce ‘*more than 100*’ although he knows that there are more than 110. This is the case we have considered in this section. There is also the possibility that roundness enters the choice of the test subject’s estimate of the interval containing the true number. This case we have not considered so far. For example, the test subject may estimate that the true number is between 103 and 150, but when asked for lower and upper bounds, she may respond between 100 and 150, simply because (100, 150) is a particularly salient interval. However, to disentangle how roundness exactly influences which module in Fig. 4 needs further experimental investigations.

6 Implicature in Which Sense?

In this section we consider the question what type of implicature are inferences from ‘*more than n*’ to upper bounds for the true n . Krifka (1999) maintained that ‘*more than n*’ does not generate implicatures. Cummins et al. (2012) considered ‘*more than n*’ implicature to be *weak* implicatures compared to standard scalar implicatures. Cummins (2013) explained them as inferences about the speaker’s beliefs which are inferred from an optimality theoretic production model. In this model, there seems to be no principled difference between implicature from ‘*more than n*’ and those from unmodified numerals, for example, from ‘*John has three children*’ to John has exactly three children.

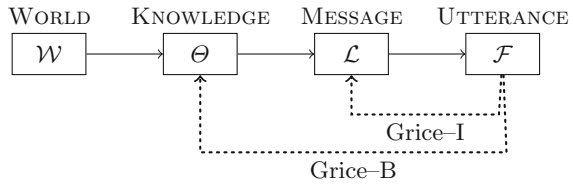
We quoted before a passage from Grice which says that an implicature is what is required that one assume a speaker to think in order to preserve the assumption that he is following the conversational maxims. This suggests that an implicature is an inference about some belief of the speaker. However, in (Grice 1989, Chap. 14), Grice defined communicated meaning as meaning that the speaker *intended* to communicate. This makes an implicature an inference about intentions, and not just about some belief he happens to entertain. Accordingly, we can distinguish between two types of implicature:

- (15) (a) (Grice–B): A belief of the speaker. (*Logic of Conversation*, Grice 1989, Chap. 2)
 (b) (Grice–I): The part of the intended meaning which is not literally expressed. (*Meaning*, Grice 1989, Chap. 14)

These implicature can be distinguished in a causal model if distinct nodes for beliefs and intentions are introduced, as shown in Fig. 5.

The standard examples of scalar implicature in (16) are clearly part of what has been communicated. Hence, they constitute Grice–I implicature.

Fig. 5 Grice-B and Grice-I implicature



- (16) (a) Some of the students failed the exam. \rightarrow not all failed.
- (b) John has three children. \rightarrow He has not more than three.

However, the implicatures about upper bounds in (7), repeated here as (17), are normally not part of what the speaker intended to communicate. In this sense they are *weaker* implicature as Cummins et al. (2012) maintained.

- (17) (a) More than 80 people got married today. \rightarrow between 80 and 100
- (b) More than 90 people got married today. \rightarrow between 90 and 100
- (c) More than 100 people got married today. \rightarrow between 100 and 200

This becomes also apparent from the causal model we considered in the previous section. Implicature from ‘more than n ’ were there considered inferences about the speaker’s knowledge.

This also reconciles Krifka’s (1999) view that ‘more than n ’ does not generate implicature with the experimental results of (Cummins et al. 2012), if one assumes that Krifka’s claim relates to Grice-I implicature, and not to Grice-B implicature.

7 Conclusion

The main aim of this paper was to show the usefulness of causal networks for analysing the experimental setting in experiments of the type performed by Cummins et al. (2012). Causal networks provide a high level description of communication by showing the causal dependencies between different modules in language production and interpretation. They are a very natural framework to use. We have seen how to explicate the notion of *implicature* in these networks as inferences from one special variable representing utterances to properties of other variables. This provided us with a general definition, and we could show that, in spite of its abstractness, it helps to make fine distinctions within the group of quantity implicature.

We have seen that the model provides a good match for upper bounds of ‘more than 100’. It explains why differences between lower and upper bounds are larger for round numbers. It also shows that this pattern is consistent with the assumption that the speaker tries to be as informative as possible. Hence, the data are consistent with the ignorance implicature from ‘more than n ’ to ‘it is possible that $n + 1$ ’. The model also predicts that ‘more than n ’ can only be used with round numbers.

We have argued that pragmatic inferences from, for example, ‘*more than 100*’ to ‘*less than 200*’ are not Grice-I implicatures, hence, that they are not part of the speaker’s intended message. This makes them different from implicatures triggered by unmodified numerals, as, for example, by ‘*John has three children*’ implicating that ‘*John has exactly three children*’. In this sense, we agree with (Krifka 1999; Fox and Hackl 2006) that comparatively modified numerals do not give rise to standard scalar implicatures. The experimental results of (Cummins et al. 2012) seem to be the consequence of general inferences about the speaker’s probable knowledge state.

References

- Benz, A. (2012). Errors in pragmatics. *Journal of Logic, Language, and Information*, 21, 97–116.
- Cummins, C. (2011). The interpretation and use of numerically-quantified expressions. Ph.D thesis, University of Cambridge
- Cummins, C., Sauerland, U., & Solt, S. (2012). Granularity and scalar implicature in numerical expressions. *Linguistics and Philosophy*, 35, 135–169.
- Cummins, C. (2013). Modelling implicatures from modified numerals. *Lingua*, 132, 103–114.
- Dehaene, S. (2011). *The number sense: How the mind creates mathematics. Revised and Updated Edition*. Oxford: Oxford University Press.
- Fox, D., & Hackl, M. (2006). The universal density of measurement. *Linguistics and Philosophy*, 29, 537–586.
- Grice, H. P. (1989). *Studies in the way of words*. Cambridge: Harvard University Press.
- Jansen, C. J., & Pollmann, M. M. (2001). On round numbers: pragmatic aspects of numerical expressions. *Journal of Quantitative Linguistics*, 8(3), 187–201.
- Krifka, M. (1999). At least some determiners aren’t determiners. In K. Turner (Ed.), *The semantics/pragmatics interface from different points of view. Current research in the semantics/pragmatics interface* (Vol. 1, pp. 257–292). Oxford: Elsevier.
- Levinson, S. C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.
- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.
- Prince, A. & Smolensky, P. (1993). Optimality theory: constraint interaction in generative grammar. Technical Report 2, Rutgers University, Centre for Cognitive Science

Measurement-Theoretic Foundations of Logic for Better Questions and Answers

Satoru Suzuki

Abstract Groenendijk and Stokhof (Varieties of formal semantics, 143–170, 1984) propose a version of partition-based semantics to analyse questions and answers. In order to compare the goodness of answers, they use qualitative orderings among answers. Groenendijk and Stokhof’s qualitative method is too coarse-grained to achieve the goal. In order to overcome the limitations, van Rooij (Formal theories of information: from Shannon to semantic information theory and general concepts of information, LNCS, 161–192, 2009) proposes an information-theoretic approach based on conditional entropy. The aims of this paper are to construct a qualitative approach that is as fine-grained as the information-theoretic one, in terms of measurement theory on the basis of Domotor (Information and inference, 148–194, 1970), and to propose a new version of complete logic for better questions and answers—Better-Question-and-Answer Logic (BQAL)—the model of the language of which is based on the qualitative and measurement-theoretic approach.

Keywords (Conditional) entropy · Information theory · Logic for better questions and answers · Measurement theory · Qualitative conditional entropy relation · Representation theorem

1 Introduction

Groenendijk and Stokhof (1984) consider the *goodness* of an answer in terms of the *partition* induced by the *question*.¹ A_Q is defined as the set of all cells of the partition Q that are *compatible* with an answer A :

$$A_Q := \{q \in Q : q \cap A \neq \emptyset\}.$$

¹ We owe this section to van Rooij (2009).

S. Suzuki (✉)
Faculty of Arts and Sciences, Komazawa University, 1-23-1 Komazawa, Setagaya-ku,
Tokyo 154-8525, Japan
e-mail: bxs05253@nifty.com

© Springer International Publishing Switzerland 2015
H. Zeevat and H.-C. Schmitz (eds.), *Bayesian Natural Language Semantics and Pragmatics*, Language, Cognition, and Mind 2,
DOI 10.1007/978-3-319-17064-0_3

In order to compare the goodness of answers, they use a *qualitative ordering* $>_Q$ among answers with relative to Q . Groenendijk and Stokhof propose that when an answer A is *incompatible* with *more cells* of Q than an answer A' , that is, $A_Q \subsetneq A'_Q$, A should be counted as a better answer to the question than A' . Van Rooij (2009) gives the following example:

Example 1 The answer A :

(At least) John is sick.

to the question Q :

Who of John and Mary are sick?

is incompatible with 2 of the 4 cells of the partition. The answer A' :

If Mary is not sick, then neither is John is.

is incompatible with 1 cell of the partition. According to Groenendijk and Stokhof, A should be a better answer to the question than A' .

What if $A_Q = A'_Q$? If A does not give extra *irrelevant information*, that is, $A' \subsetneq A$, A is considered to be a better answer than A' . Combining both constraints, Groenendijk and Stokhof define A 's being a better answer to a question Q than A' , in symbols, $A >_Q A'$ as follows:

$$A >_Q A' \text{ iff (i) } A_Q \subsetneq A'_Q, \text{ or} \\ \text{(ii) } A_Q = A'_Q \text{ and } A' \subsetneq A.$$

Van Rooij gives the following example that *Groenendijk and Stokhof's qualitative method cannot deal with*, for neither $A_Q \subsetneq A'_Q$, $A'_Q \subsetneq A_Q$, nor $A_Q = A'_Q$ holds:

Example 2 The answer A :

(At least) John and Mary are sick.

to the question Q :

Who of John, Mary and Sue are sick?

is felt to be better than the answer A' :

(At least) Sue is sick.²

Groenendijk and Stokhof's qualitative method is *too coarse-grained* to deal with this example. In order to overcome the limitations, van Rooij proposes an *information-theoretic* approach based on *conditional entropy*. He proposes that an answer A is better than an answer A' to a question Q iff the *information value* $IV_Q(A)$ of A with respect to Q is higher than that of A' , or in case both of them are the same, the

²In Sect. 7, we apply Better-Question-and-Answer Logic (BQAL) to Example 2.

informativity $I(A)$ of A should be less than that of A' , intuitively speaking, A should be *less surprising* than A' :

$$(1) \quad A >_Q A' \quad \text{iff} \quad \begin{array}{l} \text{(i) } IV_Q(A) > IV_Q(A'), \text{ or} \\ \text{(ii) } IV_Q(A) = IV_Q(A') \text{ and } I(A) < I(A'). \end{array}$$

Then van Rooij provides the following example of a better question:

Example 3 If I want to find out who of John, Mary and Sue are sick, the question:

Who of John and Mary are sick?

is felt to be better than the question:

Is Sue sick?³

Generalising the idea, he proposes that a question Q is better than a question Q' with respect to a *mutually exclusive and exhaustive set* T of hypotheses iff the *expected information value* of Q is higher than that of Q' , or if both of them are the same, a partition (question) Q' is *more fine-grained than* (\subset^*) a partition (question) Q :

$$(2) \quad Q >_T Q' \quad \text{iff} \quad \begin{array}{l} \text{(i) } EIV_T(Q) > EIV_T(Q'), \text{ or} \\ \text{(ii) } EIV_T(Q) = EIV_T(Q') \text{ and } Q' \subset^* Q. \end{array}$$

Both IV_Q and EIV_H are defined in terms of *conditional entropy*. $IV_Q(A)$ is defined as the *reduction of entropy of* Q when A is learnt:

$$IV_Q(A) := H(Q) - H_A(Q),$$

where $H(Q)$ is the entropy of Q , and $H_A(Q)$ is the entropy of Q conditional on an answer A . So we can rewrite (1) as follows:

$$(3) \quad A >_Q A' \quad \text{iff} \quad \begin{array}{l} \text{(i) } H_A(Q) < H_{A'}(Q), \text{ or} \\ \text{(ii) } H_A(Q) = H_{A'}(Q) \text{ and } I(A) < I(A'). \end{array}$$

$EIV_T(Q)$ is defined as the *average reduction of entropy of* T when an answer to Q is learnt:

$$\begin{aligned} EIV_T(Q) &:= \sum_{C \in Q} P(C) \cdot IV_T(C) = \sum_{C \in Q} P(C) \cdot (H(T) - H_C(Q)) \\ &= H(T) - \left(\sum_{C \in Q} P(C) \cdot H_C(T) \right) = H(T) - H_Q(T), \end{aligned}$$

³In Sect. 7, we apply BQAL to Example 3.

where $H_Q(T)$ is the entropy of T conditional on a question Q . So we can rewrite (2) as follows:

$$(4) \quad Q >_T Q' \text{ iff } \begin{array}{l} \text{(i)} H_Q(T) < H_{Q'}(T), \text{ or} \\ \text{(ii)} H_Q(T) = H_{Q'}(T) \text{ and } Q' \subset^* Q. \end{array}$$

The aims of this paper are

1. to construct a *qualitative* approach that is as *fine-grained* as the information-theoretic one, in terms of measurement theory on the basis of Domotor (1970), and
2. to propose a new version of *complete logic for better questions and answers*—Better-Question-and-Answer Logic (BQAL)—the model of the language of which is based on the *qualitative* and *measurement-theoretic* approach.

As a result, the qualitative and measurement-theoretic approach enables us to give an answer to the question how fine-grainedly we can express the betterness of answers to a question and the betterness of questions with respect to a mutually exclusive and exhaustive set of hypotheses in such a *weak* language as the language of *sentential logic* enriched with *sentential operators*. Of course, we can construct a logic for better questions and answers by introducing probability function symbols directly into its language. But the former logic is more *interesting* than the latter logic in the sense of the *expressibility with limited resources*.

The structure of this paper is as follows. In Sect. 2, we give descriptions of measurement-theoretic settings, define a qualitative information structure, formulate the representation theorem for a qualitative information relation, define a finite qualitative entropy difference structure, formulate the representation theorem for a qualitative entropy difference relation, and define a qualitative conditional entropy relation. In Sect. 3, on the basis of the measurement-theoretic settings, we define the language $\mathcal{L}_{\text{BQAL}}$ of BQAL. In Sect. 4, we define a model \mathfrak{M} of $\mathcal{L}_{\text{BQAL}}$, and provide BQAL with a truth definition and a validity definition. In Sect. 5, we provide BQAL with its proof system, and give some theorems of BQAL. In Sect. 6, we touch upon the soundness theorem of BQAL, and outline the proof of the completeness theorem of BQAL. In Sect. 7, we apply BQAL to two examples.

2 Measurement-Theoretic Settings

2.1 Representation Theorem and Uniqueness Theorem

There are two main problems with measurement theory⁴:

⁴Roberts (1979) gives a comprehensive survey of measurement theory. The mathematical foundation of measurement had not been studied before Hölder (1901) developed his axiomatisation for the

1. the *representation problem*—justifying the assignment of numbers to objects, and
2. the *uniqueness problem*—specifying the transformation up to which this assignment is unique.

A solution to the former can be furnished by a *representation theorem*, which establishes that the specified conditions on a qualitative relational system are (necessary and) sufficient for the assignment of numbers to objects that represents (preserves) all the relations in the system. A solution to the latter can be furnished by a *uniqueness theorem*, which specifies the transformation up to which this assignment is unique. To construct a model of the language of BQAL, we use two sorts of qualitative orderings:

1. orderings of information,
2. orderings of (conditional) entropy.

2.2 Qualitative Information Structure

We define a qualitative information structure for orderings of information as follows⁵:

Definition 1 (*Qualitative Information Structure*) Let \mathcal{W} be a non-empty set of possible worlds, \mathcal{F} a Boolean algebra of subsets of \mathcal{W} , \perp a probabilistic independence relation on \mathcal{F} , and \lesssim a relation on \mathcal{F} . Then $(\mathcal{W}, \mathcal{F}, \lesssim, \perp)$ is said to be a qualitative information structure iff the following conditions are satisfied for all variables running over \mathcal{F} .

- (I₁) $\emptyset \perp A$,
- (I₂) If $A \perp B$, then $B \perp A$,
- (I₃) If $A \perp B$, then $B^C \perp A$,
- (I₄) When $B \cap C = \emptyset$, if $A \perp B$ and $A \perp C$, then $A \perp B \cup C$,
- (I₅) $\mathcal{W} < \emptyset$, where $A < B := B \not\lesssim A$,
- (I₆) $A \lesssim \emptyset$,
- (I₇) $A \lesssim B$ or $B \lesssim A$,
- (I₈) If $A \lesssim B$ and $B \lesssim C$, then $A \lesssim C$,
- (I₉) If $A \perp B$ and $A \cap B = \emptyset$, then $A \sim \emptyset$ or $B \sim \emptyset$,
where $A \sim B := A \lesssim B$ and $B \lesssim A$,
- (I₁₀) When $C \cap A = \emptyset$ and $C \cap B = \emptyset$, $A \lesssim B$ iff $A \cup C \lesssim B \cup C$,

(Footnote 4 continued)

measurement of mass. Krantz et al. (1971), Suppes et al. (1989) and Luce et al. (1990) are seen as milestones in the history of measurement theory.

⁵We owe this subsection to Domotor (1970).

(I₁₁) When $C \perp A, B$ and $C \prec \emptyset$, $A \lesssim B$ iff $A \cap C \lesssim B \cap C$,

(I₁₂) When $B \cap D = \emptyset$, if $A \lesssim B$ and $C \lesssim D$, then $A \cup C \lesssim B \cup D$,

(I₁₃) When $A \perp C$ and $B \perp D$, if $A \lesssim B$ and $C \lesssim D$, then $A \cap C \lesssim B \cap D$,

(I₁₄) If $A_i \perp A_j$ for any i, j ($i \neq j$ and $i, j \leq n$),

then for any B , there is A_{n+1} such that for any i ($i \leq n$), $A_i \perp A_{n+1}$ and $B \sim A_{n+1}$.

Remark 1 (Classification) The above conditions can be classified into three groups:

1. The conditions (I₁)–(I₄) concern the properties of \perp .
2. The conditions (I₅)–(I₈) concern the properties of \lesssim .
3. The conditions (I₉)–(I₁₄) concern the properties of the interaction between \perp and \lesssim .

Remark 2 (Archimedeanity) (I₁₄) states *Archimedeanity*.

Domotor proves the representation theorem for \lesssim .

Theorem 1 (Representation Theorem for \lesssim , Theorem 4 of Domotor (1970)) *Let $(\mathcal{W}, \mathcal{F}, \lesssim, \perp)$ be a qualitative information structure. Then there exists a finitely additive probability measure P on \mathcal{F} such that $(\mathcal{W}, \mathcal{F}, P)$ is a probability space, and*

- (i) $I(A) = -\log_2 P(A)$,
- (ii) $A \lesssim B$ iff $I(A) \leq I(B)$,
- (iii) $A \perp B$ iff $I(A \cap B) = I(A) + I(B)$.

Remark 3 (Additivity) The condition (ii) states the *additivity of amount of information*.

2.3 Finite Qualitative Entropy Difference Structure

We define some notions about partitions as follows⁶:

Definition 2 (\subset^* , \cdot , $+^*$, \perp^* , \mathcal{B} , \mathcal{E} , \equiv^* , \mathcal{C} , \mathcal{A} , \mathcal{P}^C , and $\hat{\mathcal{P}}$)

- Let \mathbf{P} denote the set of all partitions over \mathcal{W} .
- The fine-grained-than relation (\subset^*) between $\mathcal{P}_1 \in \mathbf{P}$ and $\mathcal{P}_2 \in \mathbf{P}$ is defined as follows:

$$\mathcal{P}_1 \subset^* \mathcal{P}_2 \quad \text{iff} \quad \text{for any } A \in \mathcal{P}_1, \text{ there is } B \in \mathcal{P}_2 \text{ such that } A \subset B.$$

- The product of partitions is defined as follows:

$$\mathcal{P}_1 \cdot \mathcal{P}_2 := \{A \cap B : A \in \mathcal{P}_1 \text{ and } B \in \mathcal{P}_2\}.$$

⁶We owe this subsection to Domotor (1970).

- The sum of partitions is defined as follows:

$$\mathcal{P}_1 +^* \mathcal{P}_2 := \prod_{\mathcal{P}_1 \subset^* \mathcal{P}, \mathcal{P}_2 \subset^* \mathcal{P}} \{\mathcal{P}\},$$

where \prod denotes the generalisation of the operation.

- $\mathcal{P}_1 \perp^* \mathcal{P}_2$ iff, for any $A, B \in \mathcal{F}$, if $A \in \mathcal{P}_1$ and $B \in \mathcal{P}_2$, then $A \perp B$.
- If $A \in \mathcal{F}$, the partition $\{A, A^C\}$ is called a Bernoulli partition. The variables $\mathcal{B}, \mathcal{B}_1, \mathcal{B}_2, \dots$ will run over Bernoulli partitions.
- A partition \mathcal{P} is called equiprobable iff, for any $A, B \in \mathcal{F}$, if $A, B \in \mathcal{P}$, then $A \sim B$. The variable for equiprobable partitions will be $\mathcal{E}, \mathcal{E}_1, \mathcal{E}_2, \dots$
- We call two partitions equivalent modulo \sim , in symbol,

$$\mathcal{P}_1 \equiv^* \mathcal{P}_2, \text{ iff } \mathcal{P}_2 = [B_1|A_1]([B_2|A_2](\dots([B_n|A_n]\mathcal{P}_1)\dots)),$$

where $[B_i|A_i]\mathcal{P}$ is the partition that is the result of replacing B_i in \mathcal{P} by A_i , and, for any $i \leq n$, ($A_i \sim B_i$ and $B_i \in \mathcal{P}_1$).

- $\mathcal{O} := \{\emptyset, \mathcal{W}\}$.
- $\mathcal{A} := \{\{w\} : w \in \mathcal{W}\}$.
- $\mathcal{P}^C := \prod_{\mathcal{P}, \mathcal{Q}=\mathcal{A}, \mathcal{P} \perp^* \mathcal{Q}} \{\mathcal{Q}\}$.
- Let $(\mathcal{P} = \mathcal{P}_1 \odot \mathcal{P}_2) := (\mathcal{P} = \mathcal{P}_1 \cdot \mathcal{P}_2 \text{ and } \mathcal{P}_1 \perp^* \mathcal{P}_2 \text{ and } \mathcal{P}_1, \mathcal{P}_2 \neq^* \mathcal{O})$. Let $\mathbf{B} := \{\mathcal{P} \in \mathbf{P} : \neg \exists \mathcal{P}_1, \mathcal{P}_2 (\mathcal{P} = \mathcal{P}_1 \odot \mathcal{P}_2)\}$. Moreover, let \mathcal{D} enumerate \mathbf{B} , so that $\mathbf{B} = \{\mathcal{D}_i\}_{i \leq k}$. Then we define

$$\hat{\mathcal{P}} := (d_1, d_2, \dots, d_k),$$

where, if $\mathcal{P} = \mathcal{D}_i$ ($1 \leq i \leq k$), then $d_i = 1$, and, for any j ($j \neq i$ and $1 \leq j \leq k$), $d_j = 0$; otherwise $\hat{\mathcal{P}} = \hat{\mathcal{Q}}_1 + \hat{\mathcal{Q}}_2$ for some $\mathcal{Q}_1, \mathcal{Q}_2 \in \mathbf{P}$. $\hat{\mathcal{O}} := (0, 0, \dots, 0)$.

Remark 4 (Vector Space) $\hat{\mathcal{P}} \in \mathcal{V}(\mathbf{B})$, where \mathbf{B} is the basis of the k dimensional vector space $\mathcal{V}(\mathbf{B})$.

By means of Definition 2, we define a finite qualitative entropy difference structure for ordering of conditional entropy as follows:

Definition 3 (Finite Qualitative Entropy Difference Structure) Let \mathcal{W} be a non-empty finite set of possible worlds, \mathbf{P} an algebra of finite partitions over \mathcal{W} , \perp^* a probabilistic independence relation on \mathbf{P} , and \preceq^* a relation on $\mathbf{P} \times \mathbf{P}$. Then the quadruple $(\mathcal{W}, \mathbf{P}, \preceq^*, \perp^*)$ is said to be a finite qualitative entropy difference structure iff the following conditions are satisfied for all variables running over \mathbf{P} :

- (D₁) If $\mathcal{P}_1 \subset^* \mathcal{P}_2$, then $(\mathcal{P}_2, \mathcal{P}) \lesssim^* (\mathcal{P}_1, \mathcal{P})$,
 (D₂) If $(\mathcal{B}, \mathcal{O}) \sim^* (\mathcal{E}, \mathcal{O})$, then $(\mathcal{O}, \mathcal{P}) \prec^* (\mathcal{B}, \mathcal{P})$,
 where $\mathcal{P}_1 \sim^* \mathcal{P}_2 := (\mathcal{P}_1 \lesssim^* \mathcal{P}_2 \text{ and } \mathcal{P}_2 \lesssim^* \mathcal{P}_1)$, and $\mathcal{P}_1 \prec^* \mathcal{P}_2 := \mathcal{P}_2 \not\lesssim^* \mathcal{P}_1$,
 (D₃) $(\mathcal{P}_1, \mathcal{P}_2) \lesssim^* (\mathcal{Q}_1, \mathcal{Q}_2)$ or $(\mathcal{Q}_1, \mathcal{Q}_2) \lesssim^* (\mathcal{P}_1, \mathcal{P}_2)$,
 (D₄) If $(\mathcal{P}_1, \mathcal{P}_2) \lesssim^* (\mathcal{Q}_1, \mathcal{Q}_2)$, then $(\mathcal{Q}_2, \mathcal{Q}_1) \lesssim^* (\mathcal{P}_2, \mathcal{P}_1)$,
 (D₅) If, for i ($i < n$), $(\mathcal{P}_i, \mathcal{Q}_i) \lesssim^* (\mathcal{R}_i, \mathcal{T}_i)$, then $(\mathcal{R}_n, \mathcal{T}_n) \lesssim^* (\mathcal{P}_n, \mathcal{Q}_n)$,
 when $\sum_{i \leq n} \hat{\mathcal{P}}_i = \sum_{i \leq n} \hat{\mathcal{R}}_i$ and $\sum_{i \leq n} \hat{\mathcal{Q}}_i = \sum_{i \leq n} \hat{\mathcal{T}}_i$,
 (D₆) If $|\mathcal{P}| = |\mathcal{E}|$, then $(\mathcal{P}, \mathcal{O}) \lesssim^* (\mathcal{E}, \mathcal{O})$, where $|\cdot|$ denotes the cardinality of a partition,
 (D₇) $(\mathcal{P} \cup \{\emptyset\}, \mathcal{O}) \prec^* (\mathcal{P}, \mathcal{O})$.

Remark 5 (Dependence) The qualitative entropy difference relation \lesssim^* depends on two factors: the underlying algebra of partitions \mathbf{P} and the probabilistic independence relation \perp^* defined on this algebra. \perp^* is used in the definition of $\hat{\mathcal{P}}_i$ in (D₅).

Domotor proves the representation theorem for \lesssim^* .

Theorem 2 (Representation Theorem for \lesssim^* , Theorem 8 of Domotor (1970))
($\mathscr{H}, \mathbf{P}, \lesssim^, \perp^*$) is a finite qualitative entropy difference structure iff there exists a function $H : \mathbf{P} \rightarrow \mathbb{R}$ satisfying the following conditions for all variables running over \mathbf{P} :*

- (i) $H(\mathcal{P}) = - \sum_{A \in \mathcal{P}} P(A) \cdot \log_2 P(A)$,
 (ii) $H(\mathcal{P}_1 | \mathcal{P}_2) = - \sum_{A \in \mathcal{P}_1} \sum_{B \in \mathcal{P}_2} P(A \cap B) \cdot \log_2 P(A|B)$,
 (iii) $(\mathcal{P}_1, \mathcal{P}_2) \lesssim^* (\mathcal{Q}_1, \mathcal{Q}_2)$ iff $H(\mathcal{P}_1) - H(\mathcal{P}_2) \leq H(\mathcal{Q}_1) - H(\mathcal{Q}_2)$,
 (iv) If $\mathcal{P}_1 \perp^* \mathcal{P}_2$, then $H(\mathcal{P}_1 \cdot \mathcal{P}_2) = H(\mathcal{P}_1) + H(\mathcal{P}_2)$,
 (v) If $\mathcal{P}_1 \subset^* \mathcal{P}_2$, then $H(\mathcal{P}_2) \leq H(\mathcal{P}_1)$,
 (vi) $H(\mathcal{O}) = 0$,
 (vii) If $(\mathcal{B}, \mathcal{O}) \sim^* (\mathcal{E}, \mathcal{O})$, then $H(\mathcal{B}) = 1$,
 (viii) $H(\mathcal{P}_1 | \mathcal{P}_2) = H(\mathcal{P}_1 \cdot \mathcal{P}_2) - H(\mathcal{P}_2)$.

We define a qualitative conditional entropy relation $\mathcal{P}_1 | \mathcal{P}_2 \lesssim^* \mathcal{Q}_1 | \mathcal{Q}_2$ by a qualitative entropy difference relation as follows:

Definition 4 (Qualitative Conditional Entropy Relation)

$$\mathcal{P}_1 | \mathcal{P}_2 \lesssim^* \mathcal{Q}_1 | \mathcal{Q}_2 := (\mathcal{P}_1 \cdot \mathcal{P}_2, \mathcal{P}_2) \lesssim^* (\mathcal{Q}_1 \cdot \mathcal{Q}_2, \mathcal{Q}_2).$$

3 Language of BQAL

On the basis of the measurement-theoretic settings of Sect. 2, we can provide Better-Question-and-Answer Logic (BQAL) with its language. We define the betterness of answers to a question and the betterness of questions with respect to a mutually exclusive and exhaustive set of hypotheses in terms of qualitative information and qualitative entropy difference. Moreover, the language of BQAL includes a necessity operator \Box . The aim of introducing \Box is as follows. Let \mathscr{W} be a non-empty set of possible worlds. Then, in each $w \in \mathscr{W}$, there may be different universes, \mathscr{W}_w , relative to w . For not every $w \in \mathscr{W}$ might be able to be taken into consideration. To make this point explicit, we introduce an accessibility relation R on \mathscr{W} . By means of R , we can define \mathscr{W}_w as $\{w' \in \mathscr{W} : R(w, w')\}$. $R(w, w')$ is interpreted to mean that, in w , $w' \in \mathscr{W}$ can be imagined as an alternative to w . We need \Box to *verbalise* the behaviour of R in BQAL. So we define the language $\mathcal{L}_{\text{BQAL}}$ of better-question-and-answer logic BQAL as follows:

Definition 5 (Language) Let \mathcal{S} denote a set of sentential variables, \Box and Δ unary sentential operators, \leq and \Downarrow binary sentential operators, and \leq^* a quaternary sentential operator. The language $\mathcal{L}_{\text{BQAL}}$ of BQAL is given by the following BNF grammar:

$$\begin{aligned}
 \varphi &::= s \mid \top \mid \neg\varphi \mid \varphi \wedge \varphi \\
 \alpha &::= \varphi \mid \varphi \Downarrow \varphi \mid \varphi \leq \varphi \mid \neg\alpha \mid \alpha \wedge \alpha \mid \Box\alpha \\
 \Phi &::= \neg \left(\left(\neg \bigwedge_{i=1}^n (\neg\varphi_i) \right) \wedge \neg\top \right) \wedge \neg \left(\top \wedge \bigwedge_{i=1}^n (\neg\varphi_i) \right) \\
 &\quad \wedge \bigwedge_{1 \leq i \neq j \leq n} (\neg((\varphi_i \wedge \varphi_j) \wedge \top) \wedge \neg(\neg\top \wedge \neg(\varphi_i \wedge \varphi_j))) \\
 \mathbb{A} &::= \varphi \mid \alpha \mid (\Phi, \Phi) \leq^* (\Phi, \Phi) \mid \Delta_{1 \leq i \leq n} \Phi_i \mid \neg\mathbb{A} \mid \mathbb{A} \wedge \mathbb{A}
 \end{aligned}$$

such that $s \in \mathcal{S}$.

- \perp , \vee , \rightarrow , \leftrightarrow , and \diamond are introduced by the standard definitions.
- $\varphi \Downarrow \psi$ is interpreted to mean that φ is probabilistically independent from ψ .
- $\Box\alpha$ is interpreted to mean that necessarily α .
- $\varphi \leq \psi$ is intuitively interpreted to mean that the **answer** that φ is **less surprising** than or **as surprising as** the **answer** that ψ .
- $(\Phi, \Psi) \leq^* (X, T)$ is intuitively interpreted to mean that the **difference in uncertainty** between the **question** as to whether Φ and the **question** as to whether Ψ is felt to be **less** than or **as little as** the **difference in uncertainty** between the **question** as to whether X and the **question** as to whether T .
- $\Delta_{1 \leq i \leq n} \Phi_i$ denotes the **symmetric difference** of Φ_1, \dots, Φ_n .
- \leq , \approx , \leq^* and \approx^* are defined as follows:

$$\begin{aligned}
 - \varphi \leq \psi &::= \neg(\psi \leq \varphi). \\
 - \varphi \approx \psi &::= (\varphi \leq \psi) \wedge (\psi \leq \varphi).
 \end{aligned}$$

- $(\Phi, \Psi) \triangleleft^* (X, T) := \neg((X, T) \leq^* (\Phi, \Psi))$.
- $(\Phi, \Psi) \approx^* (X, T) := ((\Phi, \Psi) \leq^* (X, T)) \wedge ((X, T) \leq^* (\Phi, \Psi))$.

- \triangleright_X and \triangleright_X^* are defined as follows:

$$\begin{aligned} - \varphi \triangleright_X \psi &:= \\ &((\Phi \wedge X, T) \triangleleft^* (\Psi \wedge X, T)) \\ &\vee (((\Phi \wedge X, T) \approx^* (\Psi \wedge X, T)) \wedge (\varphi \triangleleft \psi)), \text{ where} \end{aligned}$$

$$\begin{cases} X := \left(\bigvee_{i=1}^m \chi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq m} ((\chi_i \wedge \chi_j) \leftrightarrow \perp), \\ T := ((\tau_1 \leftrightarrow \perp) \wedge (\tau_2 \leftrightarrow \top)) \vee ((\tau_1 \leftrightarrow \top) \wedge (\tau_2 \leftrightarrow \perp)). \end{cases}$$

$$\begin{aligned} - \Phi \triangleright_X^* \Psi &:= \\ &((X \wedge \Phi, \Phi) \triangleleft^* (X \wedge \Psi, \Psi)) \\ &\vee \left(((X \wedge \Phi, \Phi) \approx^* (X \wedge \Psi, \Psi)) \wedge \bigwedge_{i=1}^m \bigvee_{j=1}^l (\psi_i \rightarrow \varphi_j) \right), \text{ where} \end{aligned}$$

$$\begin{cases} \Phi := \left(\bigvee_{i=1}^l \varphi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq l} ((\varphi_i \wedge \varphi_j) \leftrightarrow \perp), \\ \Psi := \left(\bigvee_{i=1}^m \psi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq m} ((\psi_i \wedge \psi_j) \leftrightarrow \perp), \\ X := \left(\bigvee_{i=1}^n \chi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq n} ((\chi_i \wedge \chi_j) \leftrightarrow \perp). \end{cases}$$

- $\varphi \triangleright_X \psi$ is intuitively interpreted to mean that the **answer** that φ is **better** than the answer that ψ to the **question** as to whether/which/who $\dots X$.
- $\Phi \triangleright_X^* \Psi$ is intuitively interpreted to mean that the **question** as to whether/which/who $\dots \Phi$ is **better** than the **question** as to whether/which/who $\dots \Psi$ on a mutually exclusive and exhaustive set of **hypotheses** as to whether/which $\dots X$.

The set of all well-formed formulae of $\mathcal{L}_{\text{BQAL}}$ will be denoted by $\Phi_{\mathcal{L}_{\text{BQAL}}}$.

4 Semantics of BQAL

4.1 Model of $\mathcal{L}_{\text{BQAL}}$

We define a structured model \mathfrak{M} of $\mathcal{L}_{\text{BQAL}}$ as follows:

Definition 6 (*Model*) \mathfrak{M} is a quintuple $(\mathcal{W}, R, V, \rho, \rho^*)$ in which

- \mathcal{W} is a non-empty (**not necessarily finite**) set of possible worlds,
- R is a binary accessibility relation on \mathcal{W} ,
- V is a truth assignment to each $s \in \mathcal{S}$ for each $w \in \mathcal{W}$, and
- ρ is a **qualitative information structure assignment** that assigns to each $w \in \mathcal{W}$ a qualitative information structure $(\mathcal{W}_w, \mathcal{F}_w, \lesssim_w, \perp_w)$ in which
 - $\mathcal{W}_w := \{w' \in \mathcal{W} : R(w, w')\}$ that is not necessarily finite,
 - \mathcal{F}_w is a Boolean algebra of subsets of \mathcal{W}_w with \emptyset as zero element and W_w as unit element, and
 - \lesssim_w is a relation on \mathcal{F}_w that satisfies all of the conditions (I_1) – (I_{14}) of Definition 1.
 - \perp_w is probabilistic independence relation on \mathcal{F}_w .
- ρ^* is a **qualitative entropy difference structure assignment** that assigns to each $w \in \mathcal{W}$ a qualitative entropy difference structure $(\mathcal{W}_w, \mathbf{P}_w, \lesssim_w^*, \perp_w^*)$ in which
 - \mathbf{P} an algebra of partitions over \mathcal{W}_w , and
 - \lesssim_w^* is a relation on $\mathbf{P}_w \times \mathbf{P}_w$ that satisfies all of the conditions (D_1) – (D_7) of Definition 3.
 - \perp_w^* is probabilistic independence relation on \mathbf{P}_w .

4.2 Truth and Validity

We provide BQAL with the following truth definition relative to \mathfrak{M} and then define validity as follows:

Definition 7 (*Truth and Validity*) The notion of $\varphi \in \Phi_{\mathcal{L}_{\text{BQAL}}}$ being true at $w \in \mathcal{W}$ in \mathfrak{M} , in symbols $(\mathfrak{M}, w) \models_{\mathcal{L}_{\text{BQAL}}} \varphi$ is inductively defined as follows:

- $(\mathfrak{M}, w) \models_{\mathcal{L}_{\text{BQAL}}} s$ iff $V(w)(s) = \mathbf{true}$,
- $(\mathfrak{M}, w) \models_{\mathcal{L}_{\text{BQAL}}} \top$,
- $(\mathfrak{M}, w) \models_{\mathcal{L}_{\text{BQAL}}} \varphi \wedge \psi$ iff $(\mathfrak{M}, w) \models_{\mathcal{L}_{\text{BQAL}}} \varphi$ and $(\mathfrak{M}, w) \models_{\mathcal{L}_{\text{BQAL}}} \psi$,
- $(\mathfrak{M}, w) \models_{\mathcal{L}_{\text{BQAL}}} \neg\varphi$ iff $(\mathfrak{M}, w) \not\models_{\mathcal{L}_{\text{BQAL}}} \varphi$,
- $(\mathfrak{M}, w) \models_{\mathcal{L}_{\text{BQAL}}} \varphi \downarrow \psi$ iff $\llbracket \varphi \rrbracket_w^{\mathfrak{M}} \perp_w \llbracket \psi \rrbracket_w^{\mathfrak{M}}$,
- $(\mathfrak{M}, w) \models_{\mathcal{L}_{\text{BQAL}}} \varphi \leq \psi$ iff $\llbracket \varphi \rrbracket_w^{\mathfrak{M}} \lesssim_w \llbracket \psi \rrbracket_w^{\mathfrak{M}}$, in which $\llbracket \varphi \rrbracket_w^{\mathfrak{M}} := \{w' \in \mathcal{W} : R(w, w') \text{ and } (\mathfrak{M}, w') \models_{\mathcal{L}_{\text{BQAL}}} \varphi\}$,
- $(\mathfrak{M}, w) \models_{\text{BQAL}} \Box\alpha$ iff, for any w' such that $R(w, w')$, $(\mathfrak{M}, w') \models_{\text{BQAL}} \alpha$.
- $(\mathfrak{M}, w) \models_{\mathcal{L}_{\text{BQAL}}} (\Phi, \Psi) \leq^* (X, T)$ iff $(\{\llbracket \varphi_i \rrbracket_w^{\mathfrak{M}}\}_{1 \leq i \leq k}, \{\llbracket \psi_i \rrbracket_w^{\mathfrak{M}}\}_{1 \leq i \leq l}) \lesssim_w^* (\{\llbracket \chi_i \rrbracket_w^{\mathfrak{M}}\}_{1 \leq i \leq m}, \{\llbracket \tau_i \rrbracket_w^{\mathfrak{M}}\}_{1 \leq i \leq n})$, where

$$\left\{ \begin{array}{l} \Phi := \left(\bigvee_{i=1}^k \varphi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq k} ((\varphi_i \wedge \varphi_j) \leftrightarrow \perp), \\ \Psi := \left(\bigvee_{i=1}^l \psi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq l} ((\psi_i \wedge \psi_j) \leftrightarrow \perp), \\ X := \left(\bigvee_{i=1}^m \chi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq m} ((\chi_i \wedge \chi_j) \leftrightarrow \perp), \\ T := \left(\bigvee_{i=1}^n \tau_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq n} ((\tau_i \wedge \tau_j) \leftrightarrow \perp). \end{array} \right.$$

- $(\mathfrak{M}, w) \models_{\text{BQAL}} \Delta_{1 \leq i \leq n} \Phi_i$ iff $\Theta_{1 \leq i \leq n} \{ \llbracket \varphi_i^j \rrbracket_w^{\mathfrak{M}_1} \}_{1 \leq j \leq m}$, where

$$\Phi_i := \left(\bigvee_{j=1}^m \varphi_i^j \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq j \neq k \leq m} ((\varphi_i^j \wedge \varphi_i^k) \leftrightarrow \perp),$$

and, Θ is the **symmetric difference of partitions** (for $\mathcal{P}, \mathcal{Q} \in \mathbf{P}_w$, $\Theta(\mathcal{P}, \mathcal{Q}) = (\mathcal{P} \cdot \mathcal{Q}^c) +^* (\mathcal{P}^c \cdot \mathcal{Q})$).

If $(\mathfrak{M}, w) \models_{\mathcal{L}_{\text{BQAL}}} \varphi$ for all $w \in \mathcal{W}$, we write $\mathfrak{M} \models_{\mathcal{L}_{\text{BQAL}}} \varphi$ and say that φ is valid in \mathfrak{M} . If φ is valid in all models for $\mathcal{L}_{\text{BQAL}}$, we write $\models_{\mathcal{L}_{\text{BQAL}}} \varphi$ and say that φ is valid.

Remark 6 The truth clauses of $(\Phi, \Psi) \leq^* (X, T)$ and $\Delta_{1 \leq i \leq n} \Phi_i$ reflect the idea of *contextual definition* which provide an analysis of each sentence where Φ, Ψ, X, T or Φ_i occurs without giving a explicit definition or giving necessary and sufficient conditions for the application of Φ, Ψ, X, T or Φ_i in *isolation*. Both $(\Phi, \Psi) \leq^* (X, T)$ and $\Delta_{1 \leq i \leq n} \Phi_i$ are *minimal contexts* in which Φ, Ψ, X, T or Φ_i can occur. These minimal contexts can be embedded into some larger contexts in a fully *compositional* way. So the truth clauses of $(\Phi, \Psi) \leq^* (X, T)$ and $\Delta_{1 \leq i \leq n} \Phi_i$ can satisfy the principle of *compositionality*.

By means of Definition 7, we can provide $\varphi \triangleright_X \psi$ and $\Phi \triangleright_X^* \Psi$ with their truth conditions:

Proposition 1 (Truth Conditions of $\varphi \triangleright_X \psi$ and $\Phi \triangleright_X^* \Psi$)

- (5) $(\mathfrak{M}, w) \models_{\mathcal{L}_{\text{BQAL}}} \varphi \triangleright_X \psi$, where

$$X := ((\chi_1 \leftrightarrow \perp) \wedge (\chi_2 \leftrightarrow \top)) \vee ((\chi_1 \leftrightarrow \top) \wedge (\chi_2 \leftrightarrow \perp)).$$

iff

- (i) $(\{ \llbracket \varphi \wedge \chi_i \rrbracket_w^{\mathfrak{M}_1} \}_{1 \leq i \leq n}, \mathcal{O}) <_w^* (\{ \llbracket \psi \wedge \chi_i \rrbracket_w^{\mathfrak{M}_1} \}_{1 \leq i \leq n}, \mathcal{O})$, or
- (ii) $(\{ \llbracket \varphi \wedge \chi_i \rrbracket_w^{\mathfrak{M}_1} \}_{1 \leq i \leq n}, \mathcal{O}) \sim_w^* (\{ \llbracket \psi \wedge \chi_i \rrbracket_w^{\mathfrak{M}_1} \}_{1 \leq i \leq n}, \mathcal{O})$ and $\llbracket \varphi \rrbracket_w^{\mathfrak{M}_1} <_w \llbracket \psi \rrbracket_w^{\mathfrak{M}_1}$,

- (6) $(\mathfrak{M}, w) \models_{\mathcal{L}_{\text{BQAL}}} \Phi \triangleright_X^* \Psi$, where

$$\left\{ \begin{array}{l} \Phi := \left(\bigvee_{i=1}^l \varphi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq l} ((\varphi_i \wedge \varphi_j) \leftrightarrow \perp), \\ \Psi := \left(\bigvee_{i=1}^m \psi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq m} ((\psi_i \wedge \psi_j) \leftrightarrow \perp), \\ X := \left(\bigvee_{i=1}^n \chi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq n} ((\chi_i \wedge \chi_j) \leftrightarrow \perp). \end{array} \right.$$

iff

- (i) $\{\llbracket \chi_i \rrbracket_w^{\mathfrak{M}}\}_{1 \leq i \leq n} \mid \{\llbracket \varphi_i \rrbracket_w^{\mathfrak{M}}\}_{1 \leq i \leq l} \prec_w^* \{\llbracket \chi_i \rrbracket_w^{\mathfrak{M}}\}_{1 \leq i \leq n} \mid \{\llbracket \psi_i \rrbracket_w^{\mathfrak{M}}\}_{1 \leq i \leq m}$, or
- (ii) $\{\llbracket \chi_i \rrbracket_w^{\mathfrak{M}}\}_{1 \leq i \leq n} \mid \{\llbracket \varphi_i \rrbracket_w^{\mathfrak{M}}\}_{1 \leq i \leq l} \sim_w^* \{\llbracket \chi_i \rrbracket_w^{\mathfrak{M}}\}_{1 \leq i \leq n} \mid \{\llbracket \psi_i \rrbracket_w^{\mathfrak{M}}\}_{1 \leq i \leq m}$ and $\{\llbracket \psi_i \rrbracket_w^{\mathfrak{M}}\}_{1 \leq i \leq m} \subset^* \{\llbracket \varphi_i \rrbracket_w^{\mathfrak{M}}\}_{1 \leq i \leq l}$.

Remark 7 (Semantic Expressibility in BQAL)

- (3) in Introduction is expressed semantically by (5) in BQAL.
- (4) in Introduction is expressed semantically by (6) in BQAL.

5 Syntax (Axiomatisation) of BQAL

5.1 Proof System of BQAL

On the basis of Domotor (1969), we can translate the part of $\sum_{1 \leq i \leq n} \hat{\mathcal{P}}_i = \sum_{1 \leq i \leq n} \hat{\mathcal{R}}_i$

and $\sum_{1 \leq i \leq n} \hat{\mathcal{Q}}_i = \sum_{1 \leq i \leq n} \hat{\mathcal{T}}_i$ of (D5) of Definition 3 into a system of identities among partitions by means of the following proposition:

Proposition 2 (Symmetric Difference of Partitions)

$$\sum_{1 \leq i \leq n} \hat{\mathcal{P}}_i = \sum_{1 \leq i \leq n} \hat{\mathcal{R}}_i \quad \text{iff} \quad \begin{array}{l} \Theta_{1 \leq i \leq n} \mathcal{P}_i = \Theta_{1 \leq i \leq n} \mathcal{R}_i \\ \text{and } \Theta_{1 \leq i < j \leq n} (\mathcal{P}_i \cdot \mathcal{P}_j) = \Theta_{1 \leq i < j \leq n} (\mathcal{R}_i \cdot \mathcal{R}_j) \\ \text{and } \dots \text{ and } \Theta_{1 \leq i_1 < \dots < i_n \leq n} (\mathcal{P}_{i_1} \cdots \mathcal{P}_{i_n}) = \Theta_{1 \leq i_1 < \dots < i_n \leq n} (\mathcal{R}_{i_1} \cdots \mathcal{R}_{i_n}) \end{array}$$

By using a syntactic counterpart Δ of Θ in Proposition 2, we provide BQAL with the following proof system.

Definition 8 (Proof System of BQAL)

1. all tautologies of sentential logic,
2. $((\varphi_1 \leftrightarrow \varphi_2) \wedge (\psi_1 \leftrightarrow \psi_2)) \rightarrow ((\varphi_1 \Downarrow \psi_1) \leftrightarrow (\varphi_2 \Downarrow \psi_2))$
(Replacement of Equivalents on \Downarrow),

3. $\perp \Downarrow \varphi$
(Syntactic Counterpart of (I_1)),
4. $(\varphi \Downarrow \psi) \rightarrow (\psi \Downarrow \varphi)$
(Syntactic Counterpart of (I_2)),
5. $(\varphi \Downarrow \psi) \rightarrow (\neg\psi \Downarrow \varphi)$
(Syntactic Counterpart of (I_3)),
6. $((\varphi \wedge \chi) \leftrightarrow \perp) \rightarrow (((\varphi \Downarrow \psi) \wedge (\varphi \Downarrow \chi)) \rightarrow (\varphi \Downarrow (\psi \vee \chi)))$
(Syntactic Counterpart of (I_4)),
7. $((\varphi_1 \leftrightarrow \varphi_2) \wedge (\psi_1 \leftrightarrow \psi_2)) \rightarrow ((\varphi_1 \leq \psi_1) \leftrightarrow (\varphi_2 \leq \psi_2))$
(Replacement of Equivalents on \leq),
8. $\top \leq \perp$
(Syntactic Counterpart of (I_5)),
9. $\varphi \leq \perp$
(Syntactic Counterpart of (I_6)),
10. $(\varphi \leq \psi) \vee (\psi \leq \varphi)$
(Syntactic Counterpart of (I_7)),
11. $((\varphi \leq \psi) \wedge (\psi \leq \chi)) \rightarrow (\varphi \leq \chi)$
(Syntactic Counterpart of (I_8)),
12. $((\varphi \Downarrow \psi) \wedge ((\varphi \wedge \psi) \leftrightarrow \perp)) \rightarrow ((\varphi \leftrightarrow \perp) \vee (\psi \leftrightarrow \perp))$
(Syntactic Counterpart of (I_9)),
13. $((\chi \wedge \varphi) \leftrightarrow \perp) \wedge ((\chi \wedge \psi) \leftrightarrow \perp) \rightarrow ((\varphi \leq \psi) \leftrightarrow ((\varphi \vee \chi) \leq (\psi \vee \chi)))$
(Syntactic Counterpart of (I_{10})),
14. $((\chi \Downarrow \varphi) \wedge (\chi \Downarrow \psi) \wedge (\chi \leq \perp)) \rightarrow ((\varphi \leq \psi) \leftrightarrow ((\varphi \wedge \chi) \leq (\psi \wedge \chi)))$
(Syntactic Counterpart of (I_{11})),
15. $((\psi \wedge \tau) \leftrightarrow \perp) \rightarrow (((\varphi \leq \psi) \wedge (\chi \leq \tau)) \rightarrow ((\varphi \vee \chi) \leq (\psi \vee \tau)))$
(Syntactic Counterpart of (I_{12})),
16. $((\varphi \Downarrow \chi) \wedge (\psi \Downarrow \tau)) \rightarrow (((\varphi \wedge \chi) \leq (\psi \wedge \tau)))$
(Syntactic Counterpart of (I_{13})),
17. $\Box(\alpha \rightarrow \beta) \rightarrow (\Box\alpha \rightarrow \Box\beta)$ (K),
18. $\bigwedge_{i=1}^l \bigvee_{j=1}^m (\varphi_i \rightarrow \psi_j) \rightarrow ((\Psi, X) \leq^* (\Phi, X))$, where

$$\left\{ \begin{array}{l} \Phi := \left(\bigvee_{i=1}^l \varphi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq l} ((\varphi_i \wedge \varphi_j) \leftrightarrow \perp), \\ \Psi := \left(\bigvee_{i=1}^m \psi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq m} ((\psi_i \wedge \psi_j) \leftrightarrow \perp), \\ X := \left(\bigvee_{i=1}^n \chi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq n} ((\chi_i \wedge \chi_j) \leftrightarrow \perp) \end{array} \right.$$

(Syntactic Counterpart of (D_1)),

19. $((\Phi, \Psi) \approx^* (X, \Psi)) \rightarrow ((\Psi, T) \leq^* (\Phi, T))$, where

$$\left\{ \begin{array}{l} \Phi := ((\varphi_1 \vee \varphi_2) \leftrightarrow \top) \wedge (\varphi_1 \wedge \varphi_2) \leftrightarrow \perp, \\ \Psi := ((\psi_1 \leftrightarrow \perp) \wedge (\psi_2 \leftrightarrow \top)) \vee ((\psi_1 \leftrightarrow \top) \wedge (\psi_2 \leftrightarrow \perp)), \\ X := \left(\bigvee_{i=1}^m \chi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq l} ((\varphi_i \wedge \varphi_j) \leftrightarrow \perp) \wedge \bigwedge_{1 \leq i \neq j \leq m} (\chi_i \approx \chi_j), \\ T := \left(\bigvee_{i=1}^n \tau_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq n} ((\tau_i \wedge \tau_j) \leftrightarrow \perp) \end{array} \right.$$

(Syntactic Counterpart of (D_2)),

20. $((\Phi, \Psi) \leq^* (X, T)) \vee ((X, T) \leq^* (\Phi, \Psi))$, where

$$\left\{ \begin{array}{l} \Phi := \left(\bigvee_{i=1}^k \varphi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq k} ((\varphi_i \wedge \varphi_j) \leftrightarrow \perp), \\ \Psi := \left(\bigvee_{i=1}^l \psi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq l} ((\psi_i \wedge \psi_j) \leftrightarrow \perp), \\ X := \left(\bigvee_{i=1}^m \chi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq m} ((\chi_i \wedge \chi_j) \leftrightarrow \perp), \\ T := \left(\bigvee_{i=1}^n \tau_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq n} ((\tau_i \wedge \tau_j) \leftrightarrow \perp) \end{array} \right.$$

(Syntactic Counterpart of (D_3)),

21. $((\Phi, \Psi) \leq^* (X, T)) \rightarrow ((T, X) \leq^* (\Psi, \Phi))$, where

$$\left\{ \begin{array}{l} \Phi := \left(\bigvee_{i=1}^k \varphi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq k} ((\varphi_i \wedge \varphi_j) \leftrightarrow \perp), \\ \Psi := \left(\bigvee_{i=1}^l \psi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq l} ((\psi_i \wedge \psi_j) \leftrightarrow \perp), \\ X := \left(\bigvee_{i=1}^m \chi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq m} ((\chi_i \wedge \chi_j) \leftrightarrow \perp), \\ T := \left(\bigvee_{i=1}^n \tau_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq n} ((\tau_i \wedge \tau_j) \leftrightarrow \perp) \end{array} \right.$$

(Syntactic Counterpart of (D_4)),

22.

$$\begin{aligned}
& \left((\Delta_{1 \leq i \leq n} \Phi_i \leftrightarrow \Delta_{1 \leq i \leq n} \Psi_i) \wedge (\Delta_{1 \leq i < j \leq n} (\Phi_i \wedge \Phi_j) \leftrightarrow \Delta_{1 \leq i < j \leq n} (\Psi_i \wedge \Psi_j)) \wedge \cdots \wedge \right. \\
& (\Delta_{1 \leq i_1 < \cdots < i_n \leq n} (\Phi_{i_1} \wedge \cdots \wedge \Phi_{i_n}) \leftrightarrow \Delta_{1 \leq i_1 < \cdots < i_n \leq n} (\Psi_{i_1} \wedge \cdots \wedge \Psi_{i_n})) \wedge \\
& (\Delta_{1 \leq i \leq n} X_i \leftrightarrow \Delta_{1 \leq i \leq n} T_i) \wedge (\Delta_{1 \leq i < j \leq n} (X_i \wedge X_j) \leftrightarrow \Delta_{1 \leq i < j \leq n} (T_i \wedge T_j)) \wedge \cdots \wedge \\
& \left. (\Delta_{1 \leq i_1 < \cdots < i_n \leq n} (X_{i_1} \wedge \cdots \wedge X_{i_n}) \leftrightarrow \Delta_{1 \leq i_1 < \cdots < i_n \leq n} (T_{i_1} \wedge \cdots \wedge T_{i_n})) \right) \\
& \rightarrow \left(\bigwedge_{i=1}^{n-1} ((\Phi_i, \Psi_i) \leq^* (X_i, T_i)) \rightarrow ((X_n, T_n) \leq^* (\Phi_n, \Psi_n)) \right),
\end{aligned}$$

where

$$\left\{ \begin{array}{l}
\Phi_i := \left(\bigvee_{j=1}^{m_1} \varphi_i^j \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq j \neq k \leq m_1} ((\varphi_i^j \wedge \varphi_i^k) \leftrightarrow \perp), \\
\Psi_i := \left(\bigvee_{j=1}^{m_2} \psi_i^j \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq j \neq k \leq m_2} ((\psi_i^j \wedge \psi_i^k) \leftrightarrow \perp), \\
X_i := \left(\bigvee_{j=1}^{m_3} \chi_i^j \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq j \neq k \leq m_3} ((\chi_i^j \wedge \chi_i^k) \leftrightarrow \perp), \\
T_i := \left(\bigvee_{j=1}^{m_4} \tau_i^j \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq j \neq k \leq m_4} ((\tau_i^j \wedge \tau_i^k) \leftrightarrow \perp)
\end{array} \right.$$

(Syntactic Counterpart of (D_5)),23. $(\Phi, \Psi) \leq^* (X, \Psi)$, where

$$\left\{ \begin{array}{l}
\Phi := \left(\bigvee_{i=1}^n \varphi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq n} ((\varphi_i \wedge \varphi_j) \leftrightarrow \perp), \\
\Psi := ((\psi_1 \leftrightarrow \perp) \wedge (\psi_2 \leftrightarrow \top)) \vee ((\psi_1 \leftrightarrow \top) \wedge (\psi_2 \leftrightarrow \perp)), \\
X := \left(\bigvee_{i=1}^n \chi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq n} ((\chi_i \wedge \chi_j) \leftrightarrow \perp) \wedge \bigwedge_{1 \leq i \neq j \leq n} (\chi_i \approx \chi_j)
\end{array} \right.$$

(Syntactic Counterpart of (D_6)),24. $(\Phi \vee \perp, \Psi) \leq^* (\Phi, \Psi)$, where

$$\left\{ \begin{array}{l}
\Phi := \left(\bigvee_{i=1}^n \varphi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq n} ((\varphi_i \wedge \varphi_j) \leftrightarrow \perp), \\
\Psi := ((\psi_1 \leftrightarrow \perp) \wedge (\psi_2 \leftrightarrow \top)) \vee ((\psi_1 \leftrightarrow \top) \wedge (\psi_2 \leftrightarrow \perp))
\end{array} \right.$$

(Syntactic Counterpart of (D_7)),

25. modus ponens.

A proof of $\varphi \in \Phi_{\mathcal{L}_{\text{BQAL}}}$ is a finite sequence of $\mathcal{L}_{\text{BQAL}}$ -formulae having φ as the last formula such that either each formula is an instance of an axiom or it can be obtained from formulae that appear earlier in the sequence by applying a modus ponens. If there is a proof of φ , we write $\vdash_{\text{BQAL}} \varphi$.

Remark 8 (Behaviour and Structure) The proof system of **BQAL** has no syntactic counterpart of (I_{14}) that is satisfied in \mathfrak{M} of $\mathcal{L}_{\text{BQAL}}$ of Definition 6. For $\mathcal{L}_{\text{BQAL}}$ is not so fine-grained as to express it. However, it is not a defect of **BQAL**. For **BQAL** is designed to capture *behaviour* about \leq and \leq^* , whereas (I_{14}) states the *Archimedeanity* that is a mere *structural* property required for the existence of a finitely additive probability measure.

5.2 Theorems and Their Interpretations

BQAL has nontrivial and good properties. Among them, first we show the following *basic* facts:

Proposition 3 (Strict Weak Order (\triangleright_T)) *Suppose that*

$$T := \left(\bigvee_{i=1}^n \tau_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq n} ((\tau_i \wedge \tau_j) \leftrightarrow \perp).$$

The following then hold:

- (7) $\vdash_{\text{BQAL}} (\varphi \triangleright_T \psi) \rightarrow \neg(\psi \triangleright_T \varphi)$,
- (8) $\vdash_{\text{BQAL}} (\neg(\varphi \triangleright_T \psi) \wedge \neg(\psi \triangleright_T \chi)) \rightarrow \neg(\varphi \triangleright_T \chi)$.

Remark 9 (Asymmetry, Negative Transitivity and Strict Weak Order)

- (7) states that a being-a-better-answer-than relation satisfies *asymmetry*.
- (8) states that a being-a-better-answer-than relation satisfies *negative transitivity*.
- (7) and (8) imply that a being-a-better-answer-than relation is a *strict weak order*.

Proposition 4 (Strict Weak Order (\triangleright_T^*)) *Suppose that*

$$\left\{ \begin{array}{l} \Phi := \left(\bigvee_{i=1}^k \varphi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq k} ((\varphi_i \wedge \varphi_j) \leftrightarrow \perp), \\ \Psi := \left(\bigvee_{i=1}^l \psi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq l} ((\psi_i \wedge \psi_j) \leftrightarrow \perp), \\ X := \left(\bigvee_{i=1}^m \chi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq m} ((\chi_i \wedge \chi_j) \leftrightarrow \perp), \\ T := \left(\bigvee_{i=1}^n \tau_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq n} ((\tau_i \wedge \tau_j) \leftrightarrow \perp) \end{array} \right.$$

The following then hold:

$$\begin{aligned} (9) \quad & \vdash_{\text{BQAL}} (\Phi \triangleright_T^* \Psi) \rightarrow \neg(\Psi \triangleright_T^* \Phi), \\ (10) \quad & \vdash_{\text{BQAL}} (\neg(\Phi \triangleright_T^* \Psi) \wedge \neg(\Psi \triangleright_T^* X)) \rightarrow \neg(\Phi \triangleright_T^* X). \end{aligned}$$

Remark 10 (Strict Weak Order) (9) and (10) imply that a being-a-better-question-than relation is also a *strict weak order*.

Now we show one of essential facts in comparing answers to a given question:

Proposition 5 (Incompatible Cells and Betterness of Answers) *Suppose that*

$$X := \left(\bigvee_{i=1}^n \chi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq n} ((\chi_i \wedge \chi_j) \leftrightarrow \perp),$$

and that, for any p, q ($0 \leq q < p \leq n$), $\mathbf{DC}_{(p,q)}(\varphi, \psi, X)$ is defined as the *disjunction of all conjunctions*:

$$\bigwedge_{i=1}^n (\delta_i((\varphi \wedge \chi_i) \leftrightarrow \perp) \wedge \varepsilon_i((\psi \wedge \chi_i) \leftrightarrow \perp))$$

such that exactly p of the δ_i 's and q of ε_i 's, are the empty symbol strings, the rest of them being \neg 's, and that

$$\mathbf{DDC}(\varphi, \psi, X) := \mathbf{DC}_{(1,0)}(\varphi, \psi, X) \vee \mathbf{DC}_{(2,0)}(\varphi, \psi, X) \vee \cdots \vee \mathbf{DC}_{(n,n-2)}(\varphi, \psi, X) \vee \mathbf{DC}_{(n,n-1)}(\varphi, \psi, X).$$

The following then holds:

$$(11) \quad \vdash_{\text{BQAL}} \left(\mathbf{DDC}(\varphi, \psi, X) \wedge \bigwedge_{1 \leq i \neq j \leq n} (\chi_i \approx \chi_j) \right) \rightarrow (\varphi \triangleright_X \psi).$$

Remark 11 (11) implies that a question that can be incompatible with more cells of a given question is a better answer to it if all cells of it are equiprobable.

Next we show one of essential facts in comparing questions on a mutually exclusive and exhaustive set of hypotheses:

Proposition 6 (Fine-Grainedness and Betterness of Questions) *Suppose that*

$$\left\{ \begin{array}{l} \Phi := \left(\bigvee_{i=1}^l \varphi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq l} ((\varphi_i \wedge \varphi_j) \leftrightarrow \perp), \\ \Psi := \left(\bigvee_{i=1}^m \psi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq m} ((\psi_i \wedge \psi_j) \leftrightarrow \perp), \\ X := \left(\bigvee_{i=1}^n \chi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq n} ((\chi_i \wedge \chi_j) \leftrightarrow \perp), \text{ where } m < l < n. \end{array} \right.$$

The following then holds:

$$(12) \quad \vdash_{\text{BQAL}} \left(\bigwedge_{1 \leq i \neq j \leq l} (\varphi_i \approx \varphi_j) \wedge \bigwedge_{1 \leq i \neq j \leq m} (\psi_i \approx \psi_j) \wedge \bigwedge_{1 \leq i \neq j \leq n} (\chi_i \approx \chi_j) \right) \rightarrow (\Phi \triangleright_X^* \Psi).$$

Remark 12 (12) implies that a more fine-grained question is a better question on a mutually exclusive and exhaustive set of hypotheses if the questions and the hypothesis set each have their equiprobable cells.

6 Metalogic of BQAL

It is a routine to prove the soundness of BQAL.

Theorem 3 (Soundness) *For any $\varphi \in \Phi_{\mathcal{L}_{\text{BQAL}}}$, if $\vdash_{\text{BQAL}} \varphi$, then $\models_{\text{BQAL}} \varphi$.*

We now turn to the task of proving the completeness of BQAL. We prove it by using the ideas of Segerberg (1971) and modifying *filtration* in such a way that completeness can be established by Theorem 1 (Representation Theorem for \lesssim) and Theorem 2 (Representation Theorem for \lesssim^*). We cannot go into detail because of limited space, but the outline of the proof is as follows: We begin by defining some new concepts.

Definition 9 (*Stuffedness and Value Formula*) *Suppose that Θ is a set of formulae such that Θ is closed under subformulae. Let*

$$\begin{aligned} \Delta &:= \{ \varphi : \text{for some } \psi, \varphi \leq \psi \in \Theta \text{ or } \psi \leq \varphi \in \Theta \}, \text{ and} \\ \Delta_1^* \times \Delta_2^* &:= \{ (\Phi, \Psi) : \text{for some } (X, T), (\Phi, \Psi) \leq^* (X, T) \in \Theta \text{ or } (X, T) \leq^* (\Phi, \Psi) \in \Theta \}, \end{aligned}$$

and let Δ' ($\Delta_1^{*'}$ and $\Delta_2^{*'}$ respectively) be the closure of Δ (Δ_1^* and Δ_2^* respectively) under Boolean compounds. If Θ also satisfies the condition that $\varphi \leq \psi \in \Theta$, for any $\varphi, \psi \in \Delta'$ and that $(\Phi, \Psi) \leq^* (X, T)$, for any $(\Phi, \Psi), (X, Y) \in \Delta_1^* \times \Delta_2^*$, we say that Θ is **stuffed**. The formulae in Δ' , $\Delta_1^{*'}$ or $\Delta_2^{*'}$ are called the **value formulae** of Θ .

Remark 13 There is no occurrence of \leq and \leq^* in value formulae.

Definition 10 (*Base and Logical Finiteness*) We say that $\Lambda_0 \subset \Phi_{\mathcal{L}_{\text{BQAL}}}$ is a **base** (with respect to BQAL) for $\Lambda \subset \Phi_{\mathcal{L}_{\text{BQAL}}}$ if for any $\varphi \in \Lambda$ there is some $\varphi_0 \in \Lambda_0$ such that $\vdash_{\text{BQAL}} \varphi \leftrightarrow \varphi_0$. We say that Λ is **logically finite** (with respect to BQAL) if there is a **finite base** for Λ .

Then we can prove the next lemma:

Lemma 1 (*Logical Finiteness*) *If $\Lambda \subset \Phi_{\mathcal{L}_{\text{BQAL}}}$ is a finite set closed under subformulae, and if Θ is the smallest stuffed superset of Λ , then Θ is logically finite.*

We define BQAL-maximal consistency as follows:

Definition 11 (*BQAL-Maximal Consistent Set*) A finite set $\{\varphi_1, \dots, \varphi_n\} \subset \Phi_{\mathcal{L}_{\text{BQAL}}}$ is BQAL-consistent iff $\not\vdash_{\text{BQAL}} \neg(\varphi_1 \wedge \dots \wedge \varphi_n)$. An infinite set of formulae is BQAL-consistent iff all of its finite subsets are BQAL-consistent. $\Gamma \subset \Phi_{\mathcal{L}_{\text{BQAL}}}$ is a **BQAL-maximal consistent set** iff it is BQAL-consistent and for any $\varphi \notin \Gamma$, $\Gamma \cup \{\varphi\}$ is BQAL-inconsistent.

A canonical model for the modal part of BQAL is defined as follows:

Definition 12 (*Canonical Model for Modal Part*) We define $\mathfrak{U}^C := (\mathcal{X}^C, R^C, V^C)$ as a **canonical model for the modal part of BQAL** in which

- $\mathcal{X}^C := \{\Gamma \subset \Phi_{\mathcal{L}_{\text{BQAL}}} : \Gamma \text{ is BQAL-maximal consistent}\}$,
- for any $\Gamma, \Delta \in \mathcal{X}^C$, $R(\Gamma, \Delta)$ iff for any $\alpha \in \Phi_{\mathcal{L}_{\text{BQAL}}}$, if $\Box\alpha \in \Gamma$, then $\alpha \in \Delta$, and
- for any $\Gamma \in \mathcal{X}^C$,

$$V^C(\Gamma)(s) := \begin{cases} \mathbf{true} & \text{if } s \in \Gamma, \\ \mathbf{false} & \text{otherwise.} \end{cases}$$

We define an equivalence relation modulo Θ on \mathcal{X}^C as follows:

Definition 13 (*Equivalence Relation*) Let Θ be a stuffed set of formulae that are logically finite with respect to BQAL. We define, for $\Gamma, \Delta \in \mathcal{X}^C$,

$$\Gamma \equiv_{\Theta} \Delta \text{ iff } \Gamma \cap \Theta = \Delta \cap \Theta.$$

Then, \equiv_{Θ} is an **equivalence relation modulo Θ on \mathcal{X}^C** . We write $[\Gamma]_{\Theta}$ for the **equivalence class of Γ under \equiv_{Θ}** .

A filtration of \mathfrak{U}^C through Θ is defined as follows:

Definition 14 (*Filtration*) We define $\mathfrak{U}^\Theta := (\mathcal{X}^\Theta, R^\Theta, V^\Theta)$ as a **filtration** of \mathfrak{U}^C through Θ in which

- $\mathcal{X}^\Theta := \{[\Gamma]_\Theta : \Gamma \in \mathcal{X}^C\}$,
- R^Θ is a relation on \mathcal{X}^Θ such that
 1. if $R^C(\Gamma, \Delta)$, then $R^\Theta([\Gamma]_\Theta, [\Delta]_\Theta)$,
 2. if $R^\Theta([\Gamma]_\Theta, [\Delta]_\Theta)$ and $\Box\alpha \in \Gamma$, then $\alpha \in \Delta$, and
- V^Θ is a function such that for any $s \in \Theta$,

$$V^\Theta([\Gamma]_\Theta)(s) = V^C(\Gamma)(s).$$

Thus, for any $\xi \in \mathcal{X}^\Theta$,

$$\llbracket \varphi \rrbracket_\xi^{\mathfrak{U}^\Theta} := \{\eta : R^\Theta(\xi, \eta) \text{ and } (\mathfrak{U}^\Theta, \eta) \models_{\text{BQAL}} \varphi\}$$

is well-defined for any φ that contains neither \leq nor \leq^* . We can prove the Lindenbaum lemma:

Lemma 2 (Lindenbaum) *Every BQAL-consistent set of formulae is a subset of a BQAL-maximal consistent set of formulae.*

We can prove the partial truth lemma:

Lemma 3 (Partial Truth) *If $\varphi \in \Theta$ and φ contains neither \leq nor \leq^* , then for any $\Gamma \in \mathcal{X}^C$,*

$$(\mathfrak{U}^\Theta, [\Gamma]_\Theta) \models_{\text{BQAL}} \varphi \text{ iff } \varphi \in \Gamma.$$

We wish to supplement \mathfrak{U}^Θ with a qualitative information structure assignment ρ^Θ and a qualitative entropy difference structure assignment $\rho^{*\Theta}$ so as to obtain a structured model $\mathfrak{U}_\#^\Theta$ of $\mathcal{L}_{\text{BQAL}}$ for which the truth lemma holds for all formulae in Θ . Doing this contributes to solving the completeness problem of BQAL. \mathcal{F}_ξ^Θ is defined as follows:

Definition 15 (\mathcal{F}_ξ^Θ) For any $\xi \in \mathcal{X}^\Theta$, we define \mathcal{F}_ξ^Θ as the set of all $A \subset \mathcal{X}_\xi^\Theta := \{\eta : R^\Theta(\xi, \eta)\}$ such that for some value formula $\varphi \in \Theta$, $A = \llbracket \varphi \rrbracket_\xi^{\mathfrak{U}^\Theta}$.

We can prove the next lemma:

Lemma 4 (Boolean Algebra) *For any $\xi \in \mathcal{X}^\Theta$, \mathcal{F}_ξ^Θ is a **Boolean algebra** with \emptyset as zero element and $X_\xi^\Theta = \{\eta : R^\Theta(\xi, \eta)\}$ as unit element.*

\lesssim_ξ is defined as follows:

Definition 16 (\lesssim_ξ) For any $\xi \in \mathcal{X}^\Theta$, we define $A \lesssim_\xi B$ to hold between elements $A, B \in \mathcal{F}_\xi^\Theta$ iff there are value formulae $\varphi, \psi \in \Theta$ such that $A = \llbracket \varphi \rrbracket_\xi^{\mathfrak{U}^\Theta}$, $B = \llbracket \psi \rrbracket_\xi^{\mathfrak{U}^\Theta}$ and $\varphi \leq \psi \in \Gamma$ for any $\Gamma \in \xi$.

We can prove the next lemma:

Lemma 5 (\lesssim_ξ and \leq) *For any value formula $\varphi, \psi \in \Theta$ and any $\xi \in \mathcal{X}^\Theta$, $\llbracket \varphi \rrbracket_\xi^{\mathcal{M}^\Theta} \lesssim_\xi \llbracket \psi \rrbracket_\xi^{\mathcal{M}^\Theta}$ iff, for any $\Gamma \in \xi$, $\varphi \leq \psi \in \Gamma$.*

Since we assumed that Θ is logically finite, \mathcal{X}^Θ is finite. Hence for any $\xi \in \mathcal{X}^\Theta$, \mathcal{F}_ξ^Θ is finite. So the next lemma follows from Lemma 5.

Lemma 6 (Satisfaction of (I_1) – (I_{14})) *For any $\xi \in \mathcal{X}^\Theta$, \lesssim_ξ on \mathcal{F}_ξ^Θ satisfies (I_1) – (I_{14}) .*

Remark 14 Since \mathcal{F}_ξ^Θ is finite, (I_{14}) is automatically satisfied.

The next corollary follows from Theorem 1 (Representation Theorem for \lesssim), Lemmas 4 and 6.

Corollary 1 (Representation on \mathcal{F}_ξ^Θ) *For any $\xi \in \mathcal{X}^\Theta$, $(\mathcal{X}_\xi^\Theta, \mathcal{F}_\xi^\Theta, \lesssim_\xi, \perp_\xi)$ is a qualitative information structure, and there exists a finitely additive probability measure $P_\xi : \mathcal{F}_\xi^\Theta \rightarrow \mathbb{R}$ such that $(\mathcal{X}_\xi^\Theta, \mathcal{F}_\xi^\Theta, P_\xi)$ is a finitely additive probability space, and for any $A, B \in \mathcal{F}_\xi^\Theta$ the conditions (i)–(iii) of Theorem 1 hold.*

We continue the proof about \lesssim^* as well as \lesssim .

Definition 17 (\mathbf{P}_ξ^Θ) For any $\xi \in \mathcal{X}^\Theta$, we define \mathbf{P}_ξ^Θ as the set of all partitions \mathcal{P} over $\mathcal{X}_\xi^\Theta := \{\eta : R^\Theta(\xi, \eta)\}$ such that for some value formulae $\varphi_1, \dots, \varphi_n \in \Theta$, $\mathcal{P} := \{\llbracket \varphi_i \rrbracket_\xi^{\mathcal{M}^\Theta}\}_{1 \leq i \leq n}$, where

$$\bigcup_{i=1}^n \llbracket \varphi_i \rrbracket_\xi^{\mathcal{M}^\Theta} = \mathcal{X}_\xi^\Theta \text{ and } \bigcap_{1 \leq i \neq j \leq n} (\llbracket \varphi_i \rrbracket_\xi^{\mathcal{M}^\Theta} \cap \llbracket \varphi_j \rrbracket_\xi^{\mathcal{M}^\Theta}) = \emptyset.$$

\lesssim_ξ is defined as follows:

Definition 18 (\lesssim_ξ^*) For any $\xi \in \mathcal{X}^\Theta$, we define $(\mathcal{P}, \mathcal{Q}) \lesssim_\xi^* (\mathcal{R}, \mathcal{T})$ to hold between elements $\mathcal{P}, \mathcal{Q}, \mathcal{R}, \mathcal{T} \in \mathbf{P}_\xi^\Theta$ iff there are value formulae $\varphi_1, \dots, \varphi_k, \psi_1, \dots, \psi_l, \chi_1, \dots, \chi_m, \tau_1, \dots, \tau_n \in \Theta$ such that $\mathcal{P} = \{\llbracket \varphi_i \rrbracket_\xi^{\mathcal{M}^\Theta}\}_{1 \leq i \leq k}$, $\mathcal{Q} = \{\llbracket \psi_i \rrbracket_\xi^{\mathcal{M}^\Theta}\}_{1 \leq i \leq l}$, $\mathcal{R} = \{\llbracket \chi_i \rrbracket_\xi^{\mathcal{M}^\Theta}\}_{1 \leq i \leq m}$, $\mathcal{T} = \{\llbracket \tau_i \rrbracket_\xi^{\mathcal{M}^\Theta}\}_{1 \leq i \leq n}$ and $(\Phi, \Psi) \leq^* (X, T) \in \Gamma$ for any $\Gamma \in \xi$, where

$$\left\{ \begin{array}{l} \Phi := \left(\bigvee_{i=1}^k \varphi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq k} ((\varphi_i \wedge \varphi_j) \leftrightarrow \perp), \\ \Psi := \left(\bigvee_{i=1}^l \psi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq l} ((\psi_i \wedge \psi_j) \leftrightarrow \perp), \\ X := \left(\bigvee_{i=1}^m \chi_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq m} ((\chi_i \wedge \chi_j) \leftrightarrow \perp), \\ T := \left(\bigvee_{i=1}^n \tau_i \leftrightarrow \top \right) \wedge \bigwedge_{1 \leq i \neq j \leq n} ((\tau_i \wedge \tau_j) \leftrightarrow \perp). \end{array} \right.$$

We can prove the next lemma:

Lemma 7 (\lesssim_ξ and \leq) *For any value formula $\varphi_1, \dots, \varphi_k, \psi_1, \dots, \psi_l, \chi_1, \dots, \chi_m, \tau_1, \dots, \tau_n \in \Theta$ and any $\xi \in \mathcal{X}^\Theta$, $(\{\llbracket \varphi_i \rrbracket_\xi^{\mathcal{U}_\xi^\Theta}\}_{1 \leq i \leq k}, \{\llbracket \psi_i \rrbracket_\xi^{\mathcal{U}_\xi^\Theta}\}_{1 \leq i \leq l}) \lesssim_\xi (\{\llbracket \chi_i \rrbracket_\xi^{\mathcal{U}_\xi^\Theta}\}_{1 \leq i \leq m}, \{\llbracket \tau_i \rrbracket_\xi^{\mathcal{U}_\xi^\Theta}\}_{1 \leq i \leq n})$ iff, for any $\Gamma \in \xi$, $(\Phi, \Psi) \leq^* (X, T) \in \Gamma$.*

The next lemma follows from Lemma 7.

Lemma 8 (Satisfaction of (D_1) – (D_7)) *For any $\xi \in \mathcal{X}^\Theta$, \lesssim_ξ^* on $\mathbf{P}_\xi^\Theta \times \mathbf{P}_\xi^\Theta$ satisfies (D_1) – (D_7) .*

The next corollary follows from Theorem 2 (Representation Theorem for \lesssim^*) and Lemma 8.

Corollary 2 (Representation on \mathbf{P}_ξ^Θ) *For any $\xi \in \mathcal{X}^\Theta$, $(\mathcal{X}^\Theta, \mathbf{P}_\xi^\Theta, \lesssim_\xi^*, \perp_\xi^*)$ is a finite qualitative entropy difference structure and there exists a function $H : \mathbf{P}_\xi^\Theta \rightarrow \mathbb{R}$ satisfying the conditions (D_1) – (D_7) of Theorem 2.*

$\mathcal{U}_\#^\Theta$ is defined as follows:

Definition 19 ($\mathcal{U}_\#^\Theta$) We define $\mathcal{U}_\#^\Theta$ as $(\mathcal{X}^\Theta, R^\Theta, V^\Theta, \rho^\Theta, \rho^{*\Theta})$, where ρ^Θ is a qualitative information structure assignment that assigns to each $\xi \in \mathcal{X}^\Theta$ a qualitative information structure $(\mathcal{X}_\xi^\Theta, \mathcal{F}_\xi^\Theta, \lesssim_\xi, \perp_\xi)$ in which

- $\mathcal{X}_\xi^\Theta := \{\xi' : R^\Theta(\xi, \xi')\}$,
- \mathcal{F}_ξ^Θ is a Boolean algebra of subsets of \mathcal{X}_ξ^Θ with \emptyset as zero element and \mathcal{X}_ξ^Θ as unit element,
- \lesssim_ξ is a relation on \mathcal{F}_ξ^Θ that satisfies all of the conditions (I_1) – (I_{14}) of Definition 1,
- \perp_ξ is probabilistic independence relation on \mathcal{F}_ξ^Θ ,

and where $\rho^{*\Theta}$ is a qualitative entropy difference structure assignment that assigns to each $\xi \in \mathcal{X}^\Theta$ a qualitative entropy difference structure $(\mathcal{X}_\xi^\Theta, \mathbf{P}_\xi^\Theta, \lesssim_\xi^*, \perp_\xi^*)$ in which

- \mathbf{P}_ξ^Θ is an algebra of partitions over \mathcal{X}_ξ^Θ ,

- \lesssim_{ξ}^* is a relation on $\mathbf{P}_{\xi}^{\Theta} \times \mathbf{P}_{\xi}^{\Theta}$ that satisfies all of the conditions (D_1) – (D_7) of Definition 3, and
- \perp_{ξ}^* is probabilistic independence relation on $\mathbf{P}_{\xi}^{\Theta}$.

We can prove the full truth lemma:

Lemma 9 (Full Truth) *For any $\varphi \in \Theta$ and any $\Gamma \in \mathcal{X}^{\mathcal{C}}$,*

$$(\mathfrak{U}_{\sharp}^{\Theta}, [\Gamma]_{\Theta}) \models_{\text{BQAL}} \varphi \text{ iff } \varphi \in \Gamma.$$

Remark 15 This lemma is the announced improvement of Lemma 3.

We can prove the completeness of BQAL as follows:

Theorem 4 (Completeness) *For any $\varphi \in \Phi_{\mathcal{L}_{\text{BQAL}}}$, if $\models_{\text{BQAL}} \varphi$, then $\vdash_{\text{BQAL}} \varphi$.*

Proof Suppose that $\not\models_{\text{BQAL}} \varphi_0$. Then $\{\neg\varphi_0\}$ is a BQAL-consistent set. By Lemma 2, $\{\neg\varphi_0\}$ is a subset of a BQAL-maximal consistent set Γ . Evidently, $\varphi_0 \notin \Gamma$. Let Ψ be the set of subformulae of BQAL which is finite and let Θ be the smallest stuffed extension of Ψ . By Lemma 1, Θ is logically finite with respect to BQAL. If $\mathfrak{U}_{\sharp}^{\Theta}$ is constructed as above, it follows from Lemma 9 that $(\mathfrak{U}_{\sharp}^{\Theta}, [\Gamma]_{\Theta}) \not\models_{\text{BQAL}} \varphi_0$. Therefore, $\not\models_{\text{BQAL}} \varphi_0$.

7 Application of BQAL to Examples 2 and 3

7.1 Example 2 Reconsidered

Suppose that

$$\begin{cases} J := \text{John is sick.} \\ M := \text{Mary is sick.} \\ S := \text{Sue is sick.} \end{cases}$$

and that

$$X := \frac{(((J \wedge M \wedge S) \vee \dots \vee (\neg J \wedge \neg M \wedge \neg S)) \leftrightarrow \top)}{\wedge(((J \wedge M \wedge S) \wedge (J \wedge M \wedge \neg S)) \leftrightarrow \perp) \wedge \dots \wedge (((\neg J \wedge M \wedge \neg S) \wedge (\neg J \wedge \neg M \wedge \neg S)) \leftrightarrow \perp)}.$$

The next then follows from Proposition 5:

$$\begin{aligned} & \vdash_{\text{BQAL}} (\neg(((J \wedge M \wedge S) \wedge (J \wedge M)) \leftrightarrow \perp) \wedge \neg(((J \wedge M \wedge S) \wedge S) \leftrightarrow \perp) \\ & \vee \dots \vee (((\neg J \wedge \neg M \wedge \neg S) \wedge (J \wedge M)) \leftrightarrow \perp) \wedge (((\neg J \wedge \neg M \wedge \neg S) \wedge S) \leftrightarrow \perp) \\ & \wedge ((J \wedge M \wedge S) \approx (J \wedge M \wedge \neg S)) \wedge \dots \wedge ((\neg J \wedge \neg M \wedge S) \approx (\neg J \wedge \neg M \wedge \neg S)) \\ & \rightarrow ((J \wedge M) \triangleright_X S). \end{aligned}$$

7.2 Example 3 Reconsidered

Suppose that

$$\left\{ \begin{array}{l} \Phi := (((J \wedge M) \vee \dots \vee (\neg J \wedge \neg M)) \leftrightarrow \top) \wedge (((J \wedge M) \wedge (J \wedge \neg M)) \leftrightarrow \perp) \\ \quad \wedge \dots \wedge (((\neg J \wedge M) \wedge (\neg J \wedge \neg M)) \leftrightarrow \perp), \\ \Psi := ((S \vee \neg S) \leftrightarrow \top) \wedge ((S \wedge \neg S) \leftrightarrow \perp), \\ \quad (((J \wedge M \wedge S) \vee \dots \vee (\neg J \wedge \neg M \wedge \neg S)) \leftrightarrow \top) \\ X := \wedge(((J \wedge M \wedge S) \wedge (J \wedge M \wedge \neg S)) \perp) \\ \quad \wedge \dots \wedge (((\neg J \wedge \neg M \wedge S) \wedge (\neg J \wedge \neg M \wedge \neg S)) \leftrightarrow \perp). \end{array} \right.$$

The next then follows from Proposition 6:

$$\begin{array}{l} \vdash_{\text{BQAL}} (((J \wedge M) \approx (J \wedge \neg M)) \wedge \dots \wedge ((\neg J \wedge M) \approx (\neg J \wedge \neg M)) \\ \wedge ((J \wedge M \wedge S) \approx (J \wedge M \wedge \neg S)) \wedge \dots \wedge ((\neg J \wedge \neg M \wedge S) \approx (\neg J \wedge \neg M \wedge \neg S))) \\ \rightarrow (\Phi \triangleright_X^* \Psi). \end{array}$$

8 Concluding Remarks

In this paper, we have constructed a qualitative approach that is as fine-grained as the information-theoretic one, in terms of measurement theory on the basis of Domotor (1970), and have proposed a new version of complete logic for better questions and answers—BQAL—the model of the language of which is based on the qualitative and measurement-theoretic approach.

This paper is only a part of a larger *measurement-theoretic* study. By means of measurement theory, we constructed or are trying to construct such logics as

1. (dynamic epistemic) preference logic (Suzuki 2008, 2009b, 2014),
2. dyadic deontic logic (Suzuki 2009a),
3. vague predicate logic (Suzuki 2011b, c),
4. threshold-utility-maximiser's preference logic (Suzuki 2010, 2011a),
5. interadjective-comparison logic (Suzuki 2012b),
6. gradable-predicate logic (Suzuki 2012a),
7. doxastic and epistemic logic (Suzuki 2013e),
8. multidimensional-predicate-comparison logic (Suzuki 2013b),
9. logic for preference aggregation represented by a Nash collective utility function (Suzuki 2013c),
10. preference aggregation logic for weighted utilitarianism (Suzuki 2013d), and
11. modal-qualitative-probability logic (Suzuki 2013a)

Acknowledgments The author would like to thank two anonymous reviewers of BNLSP 13 for their helpful comments.

References

- Domotor, Z. (1969). Probabilistic relational structures and their applications. Technical report No. 144, Institute for Mathematical Studies in the Social Sciences, Stanford University.
- Domotor, Z. (1970). Qualitative information and entropy structures. In J. Hintikka & P. Suppes (Eds.), *Information and inference* (pp. 148–194). Dordrecht: Reidel.
- Groenendijk, J., & Stokhof, M. (1984). On the semantics of questions and the pragmatics of answers. In F. Landman & F. Veltman (Eds.), *Varieties of formal semantics* (pp. 143–170). Dordrecht: Foris.
- Hölder, O. (1901). Die Axiome der Quantität und die Lehre vom Mass. Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig. *Mathematisch-Physikalische Klasse*, 53, 1–64.
- Krantz, D. H., et al. (1971). *Foundations of measurement* (Vol. 1). New York: Academic Press.
- Luce, R. D., et al. (1990). *Foundations of measurement* (Vol. 3). San Diego: Academic Press.
- Roberts, F. S. (1979). *Measurement theory*. Reading: Addison-Wesley.
- Segerberg, K. (1971). Qualitative probability in a modal setting. In J. E. Fenstad (Ed.), *Proceedings of the second Scandinavian logic symposium* (pp. 341–352). Amsterdam: North-Holland.
- Suppes, P., et al. (1989). *Foundations of measurement* (Vol. 2). San Diego: Academic Press.
- Suzuki, S. (2008). Preference logic and its measurement-theoretic semantics. In *Electronic Proceedings of the 8th Conference on Logic and the Foundations of Game and Decision Theory (LOFT 2008)*, Paper No. 42.
- Suzuki, S. (2009a). Measurement-theoretic foundation of preference-based dyadic deontic logic. In X. He, et al. (Eds.), *Proceedings of the Second International Workshop on Logic, Rationality, and Interaction (LORI-II)*. LNCS (Vol. 5834, pp. 278–291). Heidelberg: Springer.
- Suzuki, S. (2009b). Prolegomena to dynamic epistemic preference logic. In H. Hattori, et al. (Eds.), *New Frontiers in artificial intelligence*. LNCS (Vol. 5447, pp. 177–192). Heidelberg: Springer.
- Suzuki, S. (2010). Prolegomena to threshold utility maximiser's preference logic. In *Electronic Proceedings of the 9th Conference on Logic and the Foundations of Game and Decision Theory (LOFT 2010)*, Paper No. 44.
- Suzuki, S. (2011a). A measurement-theoretic foundation of threshold utility maximiser's preference logic. *Journal of Applied Ethics and Philosophy*, 3, 17–25.
- Suzuki, S. (2011b). Measurement-theoretic foundations of probabilistic model of JND-based vague predicate logic. In H. van Ditmarsch, et al. (Eds.), *Proceedings of the Third International Workshop on Logic, Rationality, and Interaction (LORI-III)*. LNCS (Vol. 6953, pp. 272–285). Heidelberg: Springer.
- Suzuki, S. (2011c). Prolegomena to salient-similarity-based vague predicate logic. In T. Onada, et al. (Eds.), *New Frontiers in artificial intelligence*. LNCS (Vol. 6797, pp. 75–89). Heidelberg: Springer.
- Suzuki, S. (2012a). Measurement-theoretic foundations of gradable-predicate logic. In M. Okumura, et al. (Eds.), *New Frontiers in artificial intelligence*. LNCS (Vol. 7258, pp. 82–95). Heidelberg: Springer.
- Suzuki, S. (2012b). Measurement-theoretic foundations of interadjective-comparison logic. In A. Aguilar-Guevara, et al. (Eds.), *Proceedings of Sinn und Bedeutung 16* (Vol. 2, pp. 571–584). Cambridge: MIT Working Papers in Linguistics.
- Suzuki, S. (2013a). Epistemic modals, qualitative probability, and nonstandard probability. In M. Aloni, et al. (Eds.), *Proceedings of the 19th Amsterdam Colloquium (AC 2013)* (pp. 211–218).
- Suzuki, S. (2013b). Measurement-theoretic bases of multidimensional-predicate logic. Manuscript.
- Suzuki, S. (2013c). Measurement-theoretic foundations of many-sorted preference aggregation logic for Nash collective utility function. Manuscript.
- Suzuki, S. (2013d). Measurement-theoretic foundations of preference aggregation logic for weighted utilitarianism. In T. Yamada (Ed.), *Electronic Post-Proceedings of the Third International Workshop on Philosophy and Ethics of Social Reality (SOCREAL 2013)* (pp. 68–79).

- Suzuki, S. (2013e). Remarks on decision-theoretic foundations of doxastic and epistemic logic (revised version). *Studies in Logic*, 6, 1–12.
- Suzuki, S. (2014). Measurement-theoretic foundations of dynamic epistemic preference logic. In E. McCready, et al. (Eds.), *Formal approaches to semantics and pragmatics. Studies in Linguistics and Philosophy* (Vol. 95, pp. 295–324). Heidelberg: Springer.
- van Rooij, R. (2009). Comparing questions and answers: A bit of logic, a bit of language, and some bits of information. In G. Sommaruga (Ed.), *Formal theories of information: From Shannon to semantic information theory and general concepts of information. LNCS* (Vol. 5363, pp. 161–192). Heidelberg: Springer.

Conditionals, Conditional Probabilities, and Conditionalization

Stefan Kaufmann

Abstract Philosophers investigating the interpretation and use of conditional sentences have long been intrigued by the intuitive correspondence between the probability of a conditional ‘*if A, then C*’ and the conditional probability of *C*, given *A*. Attempts to account for this intuition within a general probabilistic theory of belief, meaning and use have been plagued by a danger of trivialization, which has proven to be remarkably recalcitrant and absorbed much of the creative effort in the area. But there is a strategy for avoiding triviality that has been known for almost as long as the triviality results themselves. What is lacking is a straightforward integration of this approach in a larger framework of belief representation and dynamics. This paper discusses some of the issues involved and proposes an account of belief update by conditionalization.

Keywords Conditionals · Probability · Conditional probability · Stalnaker Bernoulli Models

1 Introduction

Most contemporary theories of the interpretation of conditionals are inspired by Ramsey’s (1929) paraphrase of the mental process involved in their interpretation, known as the **Ramsey Test (RT)**:

Thanks to Hans-Christian Schmitz and Henk Zeevat for organizing the ESSLLI 2014 workshop on Bayesian Natural-Language Semantics and Pragmatics, where I presented an early version of this paper. Other venues at which I presented related work include the research group “What if” at the University of Konstanz and the Logic Group at the University of Connecticut. I am grateful to the audiences at all these events for stimulating discussion and feedback. Thanks also to Hans-Christian Schmitz and Henk Zeevat for their patience during the preparation of this manuscript. All errors and misrepresentations are my own.

S. Kaufmann (✉)
University of Connecticut, Mansfield, USA
e-mail: stefan.kaufmann@uconn.edu

(RT) If two people are arguing ‘*If A will C?*’ and are both in doubt as to *A*, they are adding *A* hypothetically to their stock of knowledge and arguing on that basis about *C*... We can say they are fixing their degrees of belief in *C* given *A*.

For all its intuitive appeal, (RT) is too general and underspecified to be operationalized in a concrete theoretical approach. To turn it into a definition, one has to flesh out several abstract ideas it mentions. Most crucially, this concerns the notions of **stock of knowledge**, **adding A temporarily**, and **degrees of belief**.

In this paper I explore one particular way to read (RT), *viz.* a version of probabilistic semantics inspired by Ramsey and developed by Jeffrey (1964), Adams (1965, 1975), and many subsequent authors. The basic ingredients are familiar from propositional logic. Sentences denote **propositions**, modeled as sets of **possible worlds**. Epistemic states (Ramsey’s “stocks of knowledge”) are represented in terms of (subjective) **probabilities**. The addition of the antecedent proceeds by **conditionalization**.

Regarding the latter point, another question that needs to be addressed is whether one and the same update operation is appropriate for all conditionals. Conditionalization on a proposition is typically interpreted as modeling the process of **learning** (hence coming to believe) that the proposition is true. However, there are well-known examples of conditionals in which the relevant supposition is intuitively that the antecedent is **true** without its truth being **known** to the speaker. This can be illustrated with examples whose antecedents or consequents explicitly deny the relevant belief, such as ‘*if it’s raining and I don’t know it...*’. Examples of this kind are generally problematic for theories—whether probabilistic or not—which rely on a single operation to represent the hypothetical reasoning involved. I return to this issue towards the end of the paper, having introduced a formal framework in which the relevant distinction can be accounted for.

2 Probability Models

To get things started, I begin with a fairly simple model-theoretic interpretation. For now the goal is to stay close to the standard possible-worlds apparatus of propositional logic, adding only the structure required to model probability judgments. The setup I introduce in this subsection will ultimately not work, as we will see below. But it is a good place to start.

I assume that sentences of English are mapped to expressions of the standard language of propositional logic, and spell out the formal details in terms of the latter. Let \mathcal{A} be a set of atomic propositional letters, and \mathcal{L}_A^0 be the smallest set containing \mathcal{A} and closed under negation and conjunction. I write ‘ $\bar{\varphi}$ ’ and ‘ $\varphi\psi$ ’ for the negation

of φ and the conjunction of φ and ψ , respectively. A **probability model** is a standard possible-worlds model for \mathcal{L}_A^0 , augmented with a probability measure.

Definition 1 (*Probability model*) A **probability model** for language \mathcal{L}_A^0 is a tuple $\langle \Omega, \mathcal{F}, \text{Pr}, V \rangle$, where¹:

- Ω is a non-empty set (of possible worlds);
- \mathcal{F} is a σ -algebra on Ω ;
- Pr is a probability measure on \mathcal{F} ;
- V is a function mapping sentences to characteristic functions of elements of \mathcal{F} , subject to the following conditions, for all $\varphi, \psi \in \mathcal{L}_A^0$ and $\omega \in \Omega$:

$$V(\overline{\varphi})(\omega) = 1 - V(\varphi)(\omega)$$

$$V(\varphi\psi)(\omega) = V(\varphi)(\omega) \times V(\psi)(\omega)$$

To relate the terminology familiar from possible-worlds semantics to the statistical jargon of probability theory, we can equate possible worlds with **outcomes** and sets of possible worlds (i.e., propositions) with **events**. Definition 1 follows the practice, common in probability theory, of separating the representations of the space of outcomes and the algebra on which the measure is defined (here Ω and \mathcal{F} , respectively). There are good philosophical reasons for doing this, but for the purposes of this paper, no harm is done by assuming for simplicity that \mathcal{F} is the powerset of Ω . It then follows without further stipulation that all propositions denoted by sentences are in the domain of the probability function.

Note that V maps sentences not to sets of worlds, but to their characteristic functions. Statistically speaking, those sentence denotations are **indicator variables**—a special kind of **random variables** whose range is restricted to the set $\{0, 1\}$. In the present context this perspective was first proposed by Jeffrey (1991) and further developed by Stalnaker and Jeffrey (1994). The motivation for modeling the denotations of sentences in this way will become clear below. It is important in the formal setup, but I will occasionally collapse talk of sets of worlds and talk of their characteristic functions where concepts rather than implementation are at stake.

The function Pr assigns probabilities to propositions, not sentences, but based on it we can assign probabilities to sentences indirectly as the **expectations** of their values. Generally speaking (and not just in the case of variables with range $\{0, 1\}$), the expectation of a random variable is the weighted sum of its values, where the weights are the probabilities that it takes those values.

¹A σ -algebra on Ω is a non-empty set of subsets of Ω that is closed under complements and countable unions. A probability measure on \mathcal{F} is a countably additive function from \mathcal{F} to the real interval $[0, 1]$ such that $\text{Pr}(\Omega) = 1$.

Definition 2 (*Expectation*) Let $\theta : \Omega \mapsto \mathbb{R}$ be a random variable. The **expectation** of θ relative to a probability function Pr is defined as follows²:

$$E[\theta] = \sum_{x \in \text{range}(\theta)} x \times \text{Pr}(\theta = x).$$

For the purposes of interpreting the language \mathcal{L}_A^0 , the relevant variable for a given sentence φ is its interpretation $V(\varphi)$.

Definition 3 (*Probabilities of sentences*) Given a probability model for \mathcal{L}_A^0 (see Definition 1), a function P maps sentences to real numbers as follows:

$$P(\varphi) = E[V(\varphi)]$$

Since $V(\varphi)$ is an indicator variable for all φ in \mathcal{L}_A^0 , the expectation of $V(\varphi)$ is just the probability that φ is true:

$$(2) \quad \begin{aligned} P(\varphi) &= 0 \times \text{Pr}(V(\varphi) = 0) + 1 \times \text{Pr}(V(\varphi) = 1) \\ &= \text{Pr}(V(\varphi) = 1) \end{aligned}$$

Thus in this framework we can state the connection between the probabilities of sentences on the one hand and the probabilities of propositions on the other, in the disquotational slogan in (3), the probabilistic analog of the famous Tarskian truth definition.

$$(3) \quad P(\text{“snow is white”}) = \text{Pr}(\text{snow is white})$$

Despite the equality, however, it is important to keep in mind the formal distinction between Pr , a probability measure on an algebra of propositions, and P , an assignment of probabilities to sentences. Only the former is technically a **measure** on its domain, whereas the latter maps sentences to probabilities only indirectly. We will see below that this separation of a sentence’s denotation from its probability is useful in extending the interpretation to conditionals.

3 Conditionals and Conditional Probability

While Ramsey’s formulation in (RT) is open to a variety of interpretations, the reading that is at issue here—one which many philosophers have found eminently plausible—is that the “degree of C given A ” that speakers report and aim to adjust when using indicative conditionals, is just the conditional probability of C given A . This idea

²I write ‘ $\text{Pr}(\theta = x)$ ’ to refer to the probability of the event that θ has value x . This is an abbreviation of the more cumbersome ‘ $\text{Pr}(\{\omega \in \Omega \mid \theta(\omega) = x\})$.’ I also assume, here and throughout this paper, that the range of the random variable is finite. This is guaranteed for \mathcal{L}_A^0 under V in Definition 2, but becomes a non-trivial restriction in general. Nothing hinges on it, however: Giving it up would merely require that the summations in the definitions be replaced with integrals.

has wide appeal among philosophers, going back to Jeffrey (1964), Adams (1965, 1975), and Stalnaker (1970), among others.

3.1 Conditional Expectation

In the present framework, where the probabilities of sentences are defined as the expectations of their truth values, the corresponding idea is that the probabilities of conditionals are the **conditional expectations** of their consequents, given that their antecedents are true. The conditional expectation of a random variable is its expectation relative to the conditionalized probability distribution:

Definition 4 (*Conditional expectation*) Let θ, ξ be random variables and $y \in \text{range}(\xi)$. The **conditional expectation** of θ , given $\xi = y$, is defined as

$$E[\theta|\xi = y] = \sum_{x \in \text{range}(\theta)} x \times \Pr(\theta = x|\xi = y)$$

This suggests that we could supplement the function P , which maps sentences in \mathcal{L}_A^0 to expectations, with a two-place function mapping **pairs** of sentences to **conditional expectations**:

Definition 5 (*Conditional probabilities of sentences*) Given a model for \mathcal{L}_A^0 (see Definition 1), a two-place function $P(\cdot|\cdot)$ maps pairs of sentences to conditional expectations:

$$P(\varphi|\psi) = E[V(\varphi)|V(\psi) = 1]$$

It is then easy to show, along the lines of (2) above, that $P(\varphi|\psi)$ is defined just in case the conditional probability $\Pr(V(\varphi) = 1|V(\psi) = 1)$ is, and that, where defined, the two coincide. In this sense, we could say that $P(\cdot|\cdot)$ satisfies the probabilistic reading of (RT).

It is also quite clear, however, that this approach falls far short of giving us what we want. For what we ultimately want is an assignment of probabilities to conditional sentences, rather than the pairs of sentences that constitute them. I take it to be self-evident that we should favor an approach which treats conditionals on a par with the atomic sentences and Boolean compounds that make up the rest of the language. But if there is any need for a further argument for this claim, suffice it to point out that conditionals can be embedded and compounded with each other and with other sentences. The following are all well-formed and interpretable sentences of ordinary English:

- (4) a. If this match is wet, it won't light if you strike it.
 b. If this switch will fail if it is submerged in water, it will be discarded.
 c. If this vase will crack if it is dropped on wood, it will shatter if it is dropped on marble.

Compounds like those in (4) have sometimes been claimed to bear only a superficial resemblance to compounded conditionals. Semantically, the story goes, their constituents are subjected to some kind of re-interpretation as sentences in \mathcal{L}_A^0 , so that what looks like complex conditionals in are in fact simple ones (Adams 1975; Gibbard 1981, among others). But we should be wary of this kind of story, which opens up a suspiciously handy escape hatch in the face of technical problems posed by these sentences for one's favorite theory. There certainly is no **linguistic** evidence that they are anything other than compound conditionals. A theory which assigns probabilities to them in a systematic fashion may still be found wanting on empirical grounds; but at least it would lend itself to empirical verification in the first place.

What, then, would it take to extend P to conditionals? So far they are not even represented in the language. \mathcal{L}_A^0 , being closed under the usual Boolean operations only, comes with the means to express the material conditional (definable as $\overline{\psi\varphi}$); but this is not the intended interpretation of the natural-language conditional 'if ψ then φ ', nor is its probability equivalent to the conditional probability of φ given ψ .³

So let us first augment the language with a two-place propositional connective ' \rightarrow ' as a stand-in for the English 'if-then' construction. Let \mathcal{L}_A be the smallest set containing \mathcal{L}_A^0 and closed under \rightarrow . We can now state precisely what (RT) and Definition 5 imply in this framework: The probability assigned to $\psi \rightarrow \phi$ should be equal to the conditional probability assigned to its constituents, as shown in (5a). By the definition of P, this means that the **unconditional** expectation of the conditional should equal the **conditional** expectation of the consequent, given that antecedent is true (5b):

- (5) a. $P(\psi \rightarrow \varphi) = P(\varphi|\psi)$
 b. $E[V(\psi \rightarrow \varphi)] = E[V(\varphi)|V(\psi) = 1]$

To be quite clear, (5) is not a definition. Rather, it states a desideratum that one would like to achieve when extending the domain of V and P from \mathcal{L}_A^0 to \mathcal{L}_A . Interestingly, this extension is not at all straightforward. I turn to this issue next.

3.2 Triviality

The unification of (RT) on its probabilistic reading and a truth-conditional interpretation of conditionals has been a challenge ever since Lewis (1976) presented the first of his famous **triviality results**, which have since been deepened and extended in a

³This is not the place to rehearse the arguments for and against the material conditional as an adequate rendering of our intuitions about the meaning of the 'if-then' construction. The material analysis has its adherents in philosophy (Jackson 1979; Lewis 1986, among many others) and linguistics (see Abbott 2004, for recent arguments); but it is fair to say that, especially in the philosophical tradition, such proposals tend to be driven by frustration with technical obstacles (more on this in the next subsection), rather than pre-theoretical judgments. Empirically, the probabilistic interpretation of (RT) has strong and growing support (Evans and Over 2004; Oaksford and Chater 1994, 2003, 2007).

steady stream of subsequent work. Here I only summarize the first and simplest of the results. It should be kept in mind, however, that many attempts to circumvent them within the general truth-conditional framework have been tried and shown to fail.

Suppose, then, that the probabilistic (RT) holds and conditionals denote propositions in the usual sense—i.e., sets of worlds, or in the present framework, characteristic functions thereof. Then, since the denotations of conditionals are propositions, one can ask how their probabilities should be affected by conditionalization on other propositions: Technically, since $V(\psi \rightarrow \varphi)$ denotes a random variable, we should be able to derive its conditional expectation given that some arbitrary sentence is true.

Lewis (1976) was interested in the case that the conditioning sentence is φ , the consequent of the conditional, or its negation. For those cases, Lewis made a pair of assumptions which seem rather plausible: First, given φ , the probability of the conditional $\psi \rightarrow \varphi$ should be 1; and second, given $\bar{\varphi}$, it should be 0. Thus the equalities in (6) should hold.⁴

- (6) a. $P(\psi \rightarrow \varphi|\varphi) = 1$
 b. $P(\psi \rightarrow \varphi|\bar{\varphi}) = 0$

In the present framework, this amounts to the equalities in (7).

- (7) a. $E[V(\psi \rightarrow \varphi)|V(\varphi) = 1] = 1$
 b. $E[V(\psi \rightarrow \varphi)|V(\varphi) = 0] = 0$

Now, since we are assuming that all sentences denote variables with range $\{0, 1\}$, this should hold for conditionals as well. But then (7) implies that with probability 1, the conditional is equivalent to its consequent! This consequence is grossly counterintuitive. As Lewis already pointed out, it implies that conditional consequents are probabilistically independent of their antecedents, when in fact conditionals are typically used to convey that they are not.

The literature on triviality that ensued after the publication of Lewis's seminal argument is vast and varied, and this is not the place to do justice to its depth and breadth. See Hájek and Hall (1994), Hájek (1994, 2012), Edgington (1995), Bennett (2003) and references therein for overviews, and Kaufmann (2005) for more discussion in the present framework. In the next section, I turn away from this literature of mostly negative results to an approach which evades triviality.

4 Stalnaker Bernoulli Models

The challenge posed by the triviality results has proven to be formidable; but it is so only under certain assumptions which Lewis and many subsequent authors either took for granted or found too valuable to give up. Specifically, the assumptions are that (i) the denotations of conditionals are propositions in the usual sense—here, sets of

⁴Lewis argued for the plausibility of (6) by invoking the **Import-Export Principle**, which in its probabilistic version requires that $P(\psi \rightarrow \varphi|\chi)$ be equivalent to $P(\varphi|\psi\chi)$. I avoid this move here because this principle is not universally accepted (see, for instance, Adams 1975; Kaufmann 2009).

possible worlds (or their characteristic functions); and (ii) the values of conditionals at individual worlds are fixed and do not depend on the probability distribution. Giving up these assumptions opens up an elegant way around the triviality results. This was first pointed out by van Fraassen (1976). Jeffrey (1991) arrived at a similar approach (in spirit, if not in detail) from a different angle. The connection was made explicit by Stalnaker and Jeffrey (1994), which in turn inspired subsequent work including Kaufmann (2005, 2009).

4.1 Basic Idea

The main innovation is a move from possible worlds to **sequences** of possible worlds as the entities at which sentences receive their truth values and which constitute the points in the probabilistic sample space. The idea was given a compelling intuitive motivation by Stalnaker and Jeffrey (1994), with reference to Stalnaker's (1968) similarity-based interpretation of conditionals.

Stalnaker's proposal was that the truth value of '*if A, then C*' at a possible world w is the truth value of its consequent C at a world "at which A is true and which otherwise differs minimally from the actual world." What exactly such a theory predicts then depends on the notion of minimal difference. For instance, Stalnaker maintained that each world is maximally similar to itself; thus if A is true at w , then no alternative world enters the evaluation because w is the most similar A -world to itself. Another assumption favored by Stalnaker was that for any world w and proposition X (including the contradiction), there is a unique X -world that is most similar to w . This assumption allowed Stalnaker to ensure that the conditional had certain logical properties which he considered desirable, especially Conditional Excluded Middle, i.e., that for any pair A , C and world w , one of '*if A, then C*' and '*if A, then not-C*' is true at w .

Now, the assumption that there is a unique most similar world for each proposition has always been controversial.⁵ If we give it up, allowing for multiple maximally similar worlds, but still insist on evaluating the conditional consequent relative to a single antecedent-world, then the choice of world becomes essentially non-deterministic. Van Fraassen (1976) proposed a simple formal model of such a process: If we are at an antecedent-world, we are done choosing and evaluate the conditional by evaluating its consequent. Otherwise, we continue to choose worlds in a sequence of random trials (with replacement), where at each trial the probabilities that particular worlds will be chosen are determined by the original probability distribution. Thus the probabilities at each trial are independent and identically distributed. Hence van Fraassen's term "Stalnaker Bernoulli model," suggestive of the special utility of this kind of model in linking the intuitions behind Stalnaker's possible-worlds semantics for conditionals to a standard probability-theoretic framework.

⁵See Lewis (1973), Stalnaker (1981) for some relevant arguments.

4.2 Implementation

Following van Fraassen (1976) I start with a probability model and define a **product space** consisting of sets of denumerable sequences of possible worlds, along with a probability measure that is derived from the original one. As before, sentences of the language are mapped to (characteristic functions of) sets in the probability space; now these are sets of world sequences, rather than sets of worlds. The interpretation function is derived in the following way: Sentences in \mathcal{L}_A^0 are evaluated at world sequences in terms of the first world. Conditionals $\psi \rightarrow \varphi$ are evaluated at a sequence by eliminating the (possibly empty) initial sub-sequence of worlds at which the antecedent is false, then evaluating the consequent at the remaining “tail” of the sequence. At sequences at which the antecedent is false throughout, the value is undefined. The details are given in the following definition.⁶

Definition 6 (*Stalnaker Bernoulli model for \mathcal{L}_A*) Let $\langle \Omega, \mathcal{F}, \Pr, V \rangle$ be a probability model for \mathcal{L}_A^0 . The corresponding Stalnaker Bernoulli model for \mathcal{L}_A is the tuple $\langle \Omega^*, \mathcal{F}^*, \Pr^*, V^* \rangle$ such that:

- Ω^* is the set of denumerable sequences of worlds in Ω .
For ω^* in Ω^* , I use the following notation:
 - $\omega^*[n]$ is the n th world in ω^* (thus $\omega^*[n] \in \Omega$)
 - ω^{n*} is the tail of ω^* starting at $\omega^*[n]$ (thus $\omega^{n*} \in \Omega^*$)
- \mathcal{F}^* is the set of all products $X_1 \times \cdots \times X_n \times \Omega^*$, for $n \geq 1$ and $X_i \in \mathcal{F}$.
- $\Pr^*(\cdot)$ is a probability measure on \mathcal{F}^* defined as follows, for $X_i \in \mathcal{F}$:

$$\Pr^*(X_1 \times \cdots \times X_n \times \Omega^*) = \Pr(X_1) \times \cdots \times \Pr(X_n)$$

- V^* maps pairs of sentences in \mathcal{L}_A and sequences in Ω^* to values in $\{0, 1\}$ as follows, for $\varphi, \psi \in \mathcal{L}_A$:

$$\begin{aligned} \text{if } \varphi \in \mathcal{L}_A^0, \text{ then } V^*(\varphi)(\omega^*) &= V(\varphi)(\omega^*[1]) \\ V^*(\overline{\varphi})(\omega^*) &= 1 - V^*(\varphi)(\omega^*) \\ V^*(\varphi\psi)(\omega^*) &= V^*(\varphi)(\omega^*) \times V^*(\psi)(\omega^*) \\ V^*(\varphi \rightarrow \psi)(\omega^*) &= V^*(\psi)(\omega^{n*}) \\ &\text{for the least } n \text{ s.t. } V^*(\varphi)(\omega^{n*}) = 1 \end{aligned}$$

⁶In van Fraassen’s original version, a conditional is true, rather than undefined, at a sequence not containing any tails at which the antecedent is true. The difference is of no consequence for the cases I discuss here. In general, I find the undefinedness of the conditional probability in such cases intuitively plausible and preferable, as it squares well with widely shared intuition (in the linguistic literature, at least) that indicative conditionals with impossible antecedents give rise to presupposition failure. Moreover, it follows from the results below that the (un)definedness is fairly well-behaved, in the sense that the value of the conditional is defined with probability zero or one, according as the probability of the antecedent is zero or non-zero.

I use boldfaced letters like ‘ \mathbf{X} , \mathbf{Y} ’ to refer to elements in \mathcal{F}^* . Notice that even if \mathcal{F} is the powerset of Ω , \mathcal{F}^* is not the powerset of Ω^* , but a proper subset of the latter. For instance, consider two arbitrary worlds ω_1, ω_2 and two arbitrary sequences $\omega_a^* = \langle \omega_1, \omega_2, \dots \rangle$ and $\omega_b^* = \langle \omega_2, \omega_1, \dots \rangle$. The set $\{\omega_a^*, \omega_b^*\}$ is in the powerset of Ω^* but not in \mathcal{F}^* , although it is a **subset** of sets in \mathcal{F}^* . This is not a deficiency for the purposes that I am putting this model to, since the truth conditions do not allow for the case that, say, a sentence is true at all and only the sequences in $\{\omega_a^*, \omega_b^*\}$.

As before, the probabilities assigned to sentences are the expectations of their truth values. Recall that sentences containing conditionals differ from sentences in \mathcal{L}_A^0 in that their values are not guaranteed to be defined: If a given sequence does not contain any A -worlds, then the value of any conditional with antecedent A at that sequence is undefined, and so is the value of compounds containing the conditional in question. The assignment of probabilities to sentences therefore comes with a qualification:

Definition 7 (*Probabilities of sentences*) Given a Stalnaker Bernoulli model for \mathcal{L}_A (see Definition 6), a function P^* maps sentences to real numbers as follows:

$$P^*(\varphi) = E^*[V^*(\varphi) | V^*(\varphi) \in \{0, 1\}]$$

The conditioning event on the right-hand side in this definition is that the value of the sentence in question is defined. It turns out that this is not a very constraining condition, as the following outline of an argument shows (see also van Fraassen 1976; Kaufmann 2009).

First, from the definition of V^* it follows immediately that the values of sentences in \mathcal{L}_A^0 —i.e., sentences not containing any occurrences of the conditional connective—are defined at all sequences. Consider next a “first-order” conditional—that is, a sentence of the form $\varphi \rightarrow \psi$, where both φ and ψ are in \mathcal{L}_A^0 . There is only one reason why the value of this sentence at a sequence ω^* could be undefined, namely that its antecedent φ is false throughout ω^* (i.e., at $\omega^*[n]$ for all n). Proposition 1 establishes that whenever the antecedent has positive probability, the set of sequences with this property has probability 0. (Proofs of this and all subsequent results are given in the Appendix.)

Proposition 1 (van Fraassen 1976) For $X \in \mathcal{F}$, if $\Pr(X) > 0$, then $\Pr^*\left(\bigcup_{n \in \mathbb{N}} (\overline{X}^n \times X \times \Omega^*)\right) = 1$.

As an immediate consequence, the value of the conditional is **almost surely** defined when its antecedent has positive probability. By similar reasoning, the value of the conditional is **almost surely** undefined if the antecedent has zero probability.⁷

As it stands, Proposition 1 only applies to elements of \mathcal{F} , i.e., denotations of sentences in \mathcal{L}_A^0 . It can be generalized to other sets, in particular the denotations of

⁷In probability theory, an event happens “almost surely” if its probability is 1. This notion should not be confused with logical necessity.

conditionals, but I will not pursue this argument here because the points I aim to make in this paper can be made with reference to simple sentences.

Henceforth, to forestall any complexities of exposition arising from the possibility of undefinedness, I will implicitly limit the discussion to sentences, all of whose constituents have positive probability. With this assumption, their values are guaranteed to be defined almost surely, thus the conditioning event on the right-hand side of the equation in Definition 7 can be ignored.

Van Fraassen (1976) proved the following “Fraction Lemma” (credited to Ian Hacking), which turns out to be useful in calculating the probabilities of both simple and compounded conditionals.

Lemma 1 (van Fraassen 1976) *If $\Pr(X) > 0$, then $\sum_{n \in \mathbb{N}} \Pr(\bar{X})^n = 1/\Pr(X)$.*

As a straightforward consequence of the foregoing results, it is easy to ascertain that the probabilities of first-order conditionals are the corresponding conditional probabilities.

Theorem 1 (van Fraassen 1976) *For $A, C \in \mathcal{L}_A^0$, if $P(A) > 0$, then $P^*(A \rightarrow C) = P(C|A)$.*

In addition, it is easy to see that for all $\varphi \in \mathcal{L}_A^0$, $P^*(\varphi) = P(\varphi)$. Specifically, this is true for AC and A . Thus as an immediate corollary of Theorem 1, we have that

$$(8) \quad P^*(A \rightarrow C) = P^*(C|A)$$

This is why the interpretation in a Stalnaker Bernoulli model is not subject to the Lewis-style triviality results: The denotation of the conditional is **both** a proposition (specifically, the characteristic function of a set of sequences) **and** the corresponding conditional expectation.

Theorem 1 illustrates the simplest case of a correspondence between the probabilities assigned to conditional sentences in a Stalnaker Bernoulli model and the probabilities that their \mathcal{L}_A^0 -constituents receive in the original probability model. Similar calculations, albeit of increasing complexity, can be carried out for sentences of arbitrary complexity. This was hinted at by Stalnaker and Jeffrey (1994) and worked out for cases of conditional antecedents and consequents by Kaufmann (2009). The formulas in (9a) through (9d) serve as illustration; for more details and proofs, the reader is invited to consult Kaufmann (2009).

(9) For all A, B, C, D in \mathcal{L}_A^0 :

- a. If $P(A) > 0$ and $P(C) > 0$, then

$$\begin{aligned} & P^*((A \rightarrow B) \wedge (C \rightarrow D)) \\ &= \frac{P(ABCD) + P(D|C) + P(ABC\bar{C}) + P(B|A)P(CD\bar{A})}{P(A \vee C)} \end{aligned}$$

- b. If $P(A) > 0$, $P(C) > 0$, and $P(B|A) > 0$, then

$$\begin{aligned} & P^*((A \rightarrow B) \rightarrow (C \rightarrow D)) \\ &= \frac{P(ABCD) + P(D|C)P(ABC\bar{C}) + P(B|A)P(\bar{A}CD)}{P(A \vee C)P(B|A)} \end{aligned}$$

- c. If $P(B) > 0$ and $P(C) > 0$, then
 $P^*(B \rightarrow (C \rightarrow D)) = P(CD|B) + P(D|C)P(\bar{C}|B)$
- d. If $P(A) > 0$ and $P(B|A) > 0$, then
 $P^*((A \rightarrow B) \rightarrow D) = P(D|AB)P(A) + P(D\bar{A})$

More complex compounds also receive probabilities under this approach, but they are beyond the scope of this paper because it is not clear what the empirical situation is in those cases.

4.3 Back to Probability Models

The preceding subsection showed how to start from an ordinary probability model and obtain values and probabilities for sentences in the derived Stalnaker Bernoulli model, defined by V^* and Pr^* . Now Stalnaker and Jeffrey (1994) (and, following them, Kaufmann 2009) did not stop there, for their ultimate goal was to interpret sentences in \mathcal{L}_A **in the original probability model**—i.e., to extend the value assignment function V from \mathcal{L}_A^0 to \mathcal{L}_A .

Intuitively, the idea is to let $V(\varphi)(\omega)$ be the conditional expectation of $V^*(\varphi)$, given the set of sequences in Ω^* whose first world is ω^* —that is, the set $\{\omega^* \in \Omega^* | \omega^*[1] = \omega\}$. But this set may well have zero probability, since its probability is defined as $\text{Pr}(\{\omega\})$. Stalnaker and Jeffrey’s workaround is to define the expectation relative to the ω^* -subspace of Ω^* , obtained by effectively “ignoring” the first position of the sequences.

Definition 8 Given a probability model $\mathcal{M} = \langle \Omega, \mathcal{F}, \text{Pr}, V \rangle$ for \mathcal{L}_A^0 , the interpretation function is extended to \mathcal{L}_A as follows: For all $\varphi \in \mathcal{L}_A$ and $\omega \in \Omega$,

$$V(\varphi)(\omega) = \text{Pr}^* \left(\left\{ \omega^{2*} \mid \omega^*[1] = \omega \text{ and } V^*(\varphi)(\omega^*) = 1 \right\} \right)$$

where E^*, V^* are defined relative to the Stalnaker Bernoulli model based on \mathcal{M} .

It is then possible to calculate, for sentences in \mathcal{L}_A , the values they receive under V . The simple case of first-order conditionals is given in (10).

$$(10) \quad V(A \rightarrow C)(\omega) = \begin{cases} 1 & \text{if } V(A)(\omega) = V(C)(\omega) = 1 \\ 0 & \text{if } V(A)(\omega) = 1, V(C)(\omega) = 0 \\ P(C|A) & \text{if } V(A) = 0 \end{cases}$$

The first two cases on the right-hand-side of (10) are straightforward: The value of $V^*(A \rightarrow C)$ is true at either all or none of the sequences starting with an A -world, depending on the value of C at the first world. The third case is obtained by observing that the set of sequences starting with a \bar{A} -world at which the conditional is true is the union of all sets of sequences consisting of n \bar{A} -worlds followed by an AC -world, for

$n \in \mathbb{N}$. By an argument along the same lines as the proof of Lemma 1, the measure of this set under Pr^* is just the conditional probability of C , given A :

$$(11) \sum_{n \in \mathbb{N}} (\text{P}(\bar{A})^n \times \text{P}(AC)) = \text{P}(AC)/\text{P}(A) \text{ by Lemma 1}$$

Finally, we can show that the expectation of the values in (10) is the conditional probability of the consequent, given the antecedent, as desired:

$$(12) \begin{aligned} \text{P}(A \rightarrow C) &= E[V(A \rightarrow C)] \\ &= 1 \times \text{P}(AC) + 0 \times \text{Pr}(A\bar{C}) + \text{P}(C|A) \times \text{Pr}(\bar{A}) \\ &= \text{P}(C|A)[\text{P}(A) + \text{P}(\bar{A})] = \text{P}(C|A) \end{aligned}$$

More complex compounds involving conditionals have more complicated definitions of their value assignment. Kaufmann (2009) gives the details for the cases discussed in (9a) through (9d) above. The reader is referred to the paper for details.

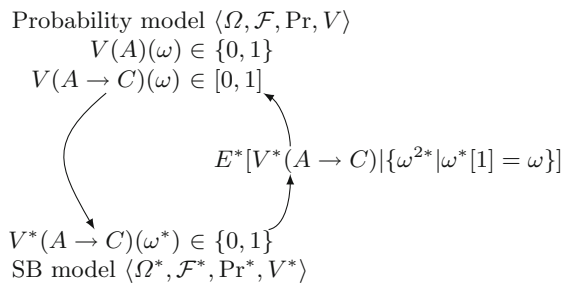
4.4 Interim Summary

The foregoing outlines a strategy for calculating values and probabilities for sentences in \mathcal{L}_A relative to a probability model \mathcal{M} , pictured schematically in Fig. 1. The extension from \mathcal{L}_A^0 to \mathcal{L}_A crucially requires the construction of the Stalnaker Bernoulli model \mathcal{M}^* , which is then used in the interpretation of sentences involving conditionals.

Now, it is clear that this clever strategy is also somewhat roundabout. All sentences receive values and probabilities in both models; moreover, due to the way one is defined in terms of the other, the probabilities actually coincide: $\text{P}(\varphi) \equiv \text{P}^*(\varphi)$ for all φ . This naturally raises the question why both models are required.

Little reflection is needed to see that once we raise this issue, it comes down to the question whether we still need the simpler probability model. For it is clear that the Stalnaker Bernoulli model is indispensable: It is crucially involved in the derivation of values and probabilities for conditionals and more complex sentences containing

Fig. 1 Value assignment for conditionals via the derived SB model



conditional constituents.⁸ But then, what is the utility of having the traditional, world-based model?

I will not venture to offer a conclusive answer to this question at this point; surely one factor in favor of the conventional possible-worlds model is its familiarity and relatively straightforward and well-understood interface with general topics in epistemology and metaphysics. So the question is to what extent these connections could be recreated in a framework that was primarily based on the Stalnaker Bernoulli approach.

Aside from such big-picture considerations, it is worth stressing that a general shift to the Stalnaker Bernoulli framework would bring additional functionality beyond the extension of the interpretation to \mathcal{L}_A . To wit, adopting the approach of assigning intermediate values to conditionals at worlds at which their antecedents are false, we face the problem that it there is no general rule of conditionalization involving conditionals—either for conditioning their denotations on other events, or for conditioning other events on them. This problem does not arise in the Stalnaker Bernoulli approach, where the probability of a conditional is the probability of a proposition (in addition to being a conditional probability).

To get such a project off the ground, however, much work has to be done. For one thing, the Stalnaker Bernoulli model would need some intuitive interpretation. Moreover, we would need plausible representations of such elementary notions as belief change by conditionalization.

Regarding the first of these desiderata—an intuitive interpretation of Stalnaker Bernoulli models—we can derive some inspiration from the suggestions of Stalnaker and Jeffrey, cited above. A set \mathbf{X} of world sequences represents two kinds of information: “factual” beliefs are encoded in the set of “first worlds” $\{\omega^*[1] \mid \omega^* \in \mathbf{X}\}$, which is where all sentences in \mathcal{L}_A^0 receive their truth values. The other kind of information consists in “conditional” beliefs, encoded in the sequences $\{\omega^{2*} \mid \omega^* \in \mathbf{X}\}$. Each sequence represents a possible outcome of a countable sequence of random choices of a world (with replacement), modeling a non-deterministic variant of the interpretation of conditionals in terms of a Stalnaker selection function.⁹

With this in mind, I will set the first issue aside and spend the rest of this paper focusing on the second issue, the definition of belief update by conditionalization in a Stalnaker Bernoulli model.

⁸As a historical side note, it is worth pointing out that some of the functionality delivered here by the Stalnaker Bernoulli model can also be achieved in a simpler model. This was shown by Jeffrey (1991), who developed the random-variable approach with intermediate truth values without relying on van Fraassen’s construction. But that approach has its limits, for instance when it comes to conditionals with conditional antecedents, and can be seen as superseded by the Stalnaker Bernoulli approach.

⁹In a related sense, one may also think of a given set of sequences as representing all paths following an introspective (i.e., transitive and euclidean) doxastic accessibility relation. I leave the further exploration of this connection for future work.

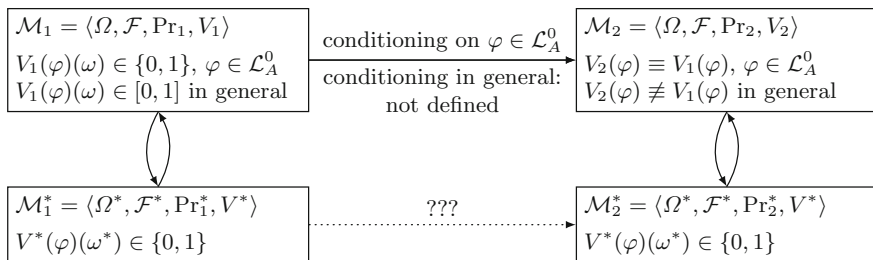


Fig. 2 Belief update in probability models with SB interpretations for conditionals

5 Conditionalization

What is the problem with conditionalization in Stalnaker Bernoulli models? Figure 2 gives a general overview. Suppose we start out, as we did in this paper, with \mathcal{M}_1 , a conventional probabilistic possible-worlds model. In it, the non-conditional sentences in \mathcal{L}_A^0 receive truth values at worlds, but in order to extend the value assignment to conditionals in accordance with the probabilistic interpretation of (RT), we take a detour via the derived Stalnaker Bernoulli model \mathcal{M}_1^* (as shown in Fig. 1). This allows us to extend V_1 to conditionals.

Now, suppose we want to update the belief model by conditioning on a new piece of information, say, that some sentence φ is true. If $\varphi \in \mathcal{L}_A^0$, this is not a problem: As usual in Bayesian update, we conditionalize by shifting the probability mass to the set of worlds at which φ is true, then renormalize the measure so as to ensure that we have once again a probability distribution, thus obtaining Pr_2 .

Now, a number of things are noteworthy about this update operation. First, along with the shift from Pr_1 to Pr_2 , the value assignment for conditionals must also change in order to ensure that the expectation of the new measure is the conditional probability. Thus we obtain a valuation function V_2 which agrees with V_1 on all sentences in \mathcal{L}_A^0 , but may differ in the values it assigns to conditionals and sentences containing them. The second noteworthy point about the update is that we have no systematic way of deriving this new assignment function V_2 from V_1 . After the update, the new values for conditionals have to be filled in by once again taking the detour via the derived Stalnaker Bernoulli model, now \mathcal{M}_2^* . Thirdly, while the update with a sentence in \mathcal{L}_A^0 just discussed merely presents a minor inconvenience (in calling for a recalculation of the values of sentences not in \mathcal{L}_A^0 via the derived Stalnaker Bernoulli model), matters are worse if the incoming information is conditional, i.e., the information that some sentence not in \mathcal{L}_A^0 is true. Whenever such a sentence takes values strictly less than 1, we do not have a way to conditionalize on it.

Notice, though, that these problems would not arise if we were using Stalnaker Bernoulli models throughout. In \mathcal{M}_1^* and \mathcal{M}_2^* , all sentences receive values in $\{0, 1\}$ almost surely (with the above caveats about the possibility of undefinedness and its intuitive justification). Moreover, notice that the two Stalnaker Bernoulli models share the same valuation function. For recall that $V^*(\varphi)(\omega^*)$ only depends on the

structure of ω^* , not on the probability distribution. And since conditionals denote propositions in these models (i.e., elements of \mathcal{F}^*), conditionalization involving those denotations can proceed in the familiar fashion. I take all of these facts to be compelling arguments in favor of the exclusive use of Stalnaker Bernoulli models.

5.1 Shallow Conditioning

How exactly should belief update be carried out in Stalnaker Bernoulli models? To see why this is an issue, consider what would seem to be the most straightforward way to define conditionalization on a sentence φ in \mathcal{M}_1^* . Presumably, similarly to the analogous procedure on \mathcal{M}_1 , this would involve shifting the probability mass onto sequences at which $V^*(\varphi)$ evaluates to 1, then renormalizing the measure to ensure that the result is again a probability distribution.¹⁰

Now, the problem with this approach is that the resulting probability distribution is not Pr_2^* , hence the resulting model is not \mathcal{M}_2^* . It is easy to see why this is the case. For concreteness, let us assume that the new information is that some sentence A in \mathcal{L}_A^0 is true. By the definition of the Stalnaker Bernoulli model, $V^*(A)(\omega^*) = 1$ whenever $V(A)(\omega^*[1]) = 1$ —in words, if **the first world** in ω^* is an A -world. This includes all sequences which begin with an A -world, regardless of what happens later on in them.

But this is not the set of sequences over which the probability is distributed in \mathcal{M}_2^* . For recall that \mathcal{M}_2^* was obtained from \mathcal{M}_1 by conditionalization on the information that A is true. As a result of this operation, in \mathcal{M}_2 the entire probability mass is assigned to worlds at which A is true. Hence, in \mathcal{M}_2^* , the probability mass is concentrated on sequences which consist **entirely** of A -worlds (i.e., sequences ω^* such that $V^*(A)(\omega^{n*}) = 1$ for all n).

Thus in order to fill in the missing step from \mathcal{M}_1^* to \mathcal{M}_2^* in Fig. 2, we would have to conditionalize on the set of sequences in (13b), rather than (13a).

- (13) a. $\{\omega^* \in \Omega^* \mid V^*(\varphi)(\omega^*) = 1\}$
 b. $\{\omega^* \in \Omega^* \mid V^*(\varphi)(\omega^{n*}) = 1 \text{ for all } n \geq 1\}$

However, the set in (13b) has zero probability whenever the probability of φ is less than 1. Indeed, for any $X \in \mathcal{F}$, the set X^* of sequences consisting entirely of X -worlds has probability 1 or 0, according as the probability of X is 1 or less than 1:

$$(14) \text{Pr}^*(X^*) = \lim_{n \rightarrow \infty} \text{Pr}(X)^n = \begin{cases} 1 & \text{if } \text{Pr}(X) = 1 \\ 0 & \text{otherwise} \end{cases}$$

Clearly a different definition is needed to work around this problem.

¹⁰An alternative way of achieving the same result would be to model belief update in terms of “truncation” of world sequences along the lines of the interpretation of conditionals, chopping off initial sub-sequences until the remaining tail verifies φ . I will not go into the details of this operation here; it corresponds to the operation of shallow conditioning discussed in this subsection, for the same reason that the probabilities of conditionals equal the corresponding conditional probabilities.

5.2 Deep Conditioning

It is not uncommon in discussions of probability to see “Bayes’s Rule” as a definition of conditional probability:

$$(BR) \quad \Pr(X|Y) = \frac{\Pr(X \cap Y)}{\Pr(Y)} \text{ when } \Pr(Y) > 0$$

Taken as a definition, this suggests that the conditional probability is generally undefined when the probability of the conditioning event is 0. But this view has many problematic consequences, not the least of them being that it predicts that the conditional probability is undefined in cases in which we in fact have clear intuitions that it exists and what it should be.¹¹ This point has been discussed in the philosophical literature (see for instance Stalnaker and Jeffrey 1994; Jeffrey 2004; Hájek 2003, 2011, and references therein), but the misconception that (BR) is the definition of conditional probability is apparently hard to root out.¹²

My own take on conditional probability follows the lead of Stalnaker and Jeffrey (1994) and Jeffrey (2004): (BR) is not a definition, but a restatement of the “Product Rule” (PR):

$$(PR) \quad \Pr(X|Y) \times \Pr(Y) = \Pr(X \cap Y)$$

Most crucially, (PR) should not be mistaken for a definition either, but as an axiom regulating the relationship between unconditional and conditional probabilities **when both are defined**.

Following this general strategy, in Definition 9 I give one definition of a two-place function $\Pr^*(\cdot|\cdot)$, then I proceed to show that it is in fact properly called a conditional probability.

Definition 9 (*Stalnaker Bernoulli model with conditional probability*)
An Stalnaker Bernoulli model with conditional probability is a tuple $\langle \Omega^*, \mathcal{F}^*, \Pr^*(\cdot), \Pr^*(\cdot|\cdot), V^* \rangle$, where $\langle \Omega^*, \mathcal{F}^*, \Pr^*(\cdot), V^* \rangle$ is a Stalnaker Bernoulli model (see Definition 6) and $\Pr^*(\cdot|\cdot)$ is a partial function mapping pairs of propositions in \mathcal{F}^* to real numbers as follows, for all \mathbf{X}, \mathbf{Y} in \mathcal{F}^* :

$$\Pr^*(\mathbf{X}|\mathbf{Y}) = \lim_{n \rightarrow \infty} \frac{\Pr^*(X_1 \cap Y_1 \times \cdots \times X_n \cap Y_n \times \Omega^*)}{\Pr^*(Y_1 \times \cdots \times Y_n \times \Omega^*)}$$

¹¹As a simple example, consider the task of choosing a point (x, y) at random from a plane. Fix some point (x^*, y^*) and consider the conditional probability that $y > y^*$, given $x = x^*$ (intuitively, the conditional probability that the randomly chosen point will lie above (x^*, y^*) , given that it lies on the vertical line through (x^*, y^*)). We have clear intuitions as to what this conditional probability is and how it depends on the location of the cutoff point x^* ; but the probability that the randomly chosen point lies on the line is 0.

¹²Notice, incidentally, that the view on conditional probability just endorsed is not at odds with the remarks on the undefinedness of the values of conditionals at world sequences throughout which the antecedent is false (see footnote 6). For one thing, technically the undefinedness discussed there does not enter the picture because some conditional probability is undefined. But that aside, I emphasize that I do not mean to claim that conditional probabilities given zero-probability events are **always** defined, only that they can be.

This definition opens up the possibility that $\Pr^*(\mathbf{X}|\mathbf{Y})$ is defined while the ratio $\Pr^*(\mathbf{X} \cap \mathbf{Y})/\Pr^*(\mathbf{Y})$ is not. For example, let $\mathbf{X} = \mathbf{Y} = Z^*$ for some $Z \in \mathcal{F}$ with $0 < \Pr(Z) < 1$, and note that in this case $\Pr^*(Z^*) = \lim_{n \rightarrow \infty} \Pr(Z)^n = 0$. Thus for instance, the quotient $\Pr^*(Z^*)/\Pr^*(Z^*)$ is undefined, hence Bayes's Rule is silent on the conditional probability of Z^* given Z^* . However, under Definition 9 the value of $\Pr^*(Z^*|Z^*)$ is defined (proofs are given in the appendix):

Proposition 2 *If $\Pr(Z) > 0$, then $\Pr^*(Z^*|Z^*) = 1$.*

Now, while $\Pr^*(\cdot|\cdot)$ may be defined when the quotient in (BR) is not, the next two results show that it is “well-behaved” with respect to the Product rule (PR).

Proposition 3 *If $\Pr^*(\mathbf{X}|\mathbf{Y})$ is defined, then $\Pr^*(\mathbf{X}|\mathbf{Y}) \times \Pr^*(\mathbf{Y}) = \Pr^*(\mathbf{X} \cap \mathbf{Y})$.*

As mentioned above, Stalnaker and Jeffrey (1994) likewise prefer to impose the product rule as an axiom, rather than using it as a definition. They also impose an additional condition which in the present framework falls out from the definition of conditional probability whenever it is defined:

Proposition 4 *If $\Pr^*(\mathbf{X}|\mathbf{Y})$ is defined and $\mathbf{Y} \subseteq \mathbf{X}$, then $\Pr^*(\mathbf{X}|\mathbf{Y}) = 1$.*

I conclude that $\Pr^*(\cdot|\cdot)$ can properly be called a “conditional probability.”¹³ At the same time, there are cases in which (BR) cannot be applied, yet $\Pr^*(\cdot|\cdot)$ is defined. This possibility is important in modeling belief update in Stalnaker Bernoulli models.

The remaining results in this section show that conditionalization according to Definition 9 is well-behaved. The first generalizes Proposition 1 above to probabilities conditionalized on a sequence Z^* .

Proposition 5 *If $\Pr(X \cap Z) > 0$, then $\Pr^*\left(\bigcup_{n \in \mathbb{N}} \left(\bar{X}^n \times X \times \Omega^*\right) \middle| Z^*\right) = 1$.*

Notice that Proposition 1 is a corollary of Proposition 5, substituting Ω for Z . Lemma 1 (van Fraassen's “Fraction Lemma”) can likewise be generalized to conditional probabilities.

Lemma 2 (Conditional Fraction Lemma) *If $\Pr(X \cap Z) > 0$, then $\sum_{n \in \mathbb{N}} \Pr(\bar{X}|Z)^n = 1/\Pr(X|Z)$.*

Lemma 2 allows us to determine the conditional probability assigned to the denotation of a conditional given a sequence Z^* .

¹³However, $\Pr^*(\mathbf{X}|\mathbf{Y})$ is itself not always defined: It is undefined if either $\Pr(Y_i) = 0$ for any i , or if the function does not converge as n approaches infinity.

6 Some Consequences for Conditionals

To sum up the preceding subsections, the Stalnaker Bernoulli framework offers two natural ways to interpret conditionals: by the rule in Definition 6, and by deep conditioning. The two are distinguished by the characterization of the conditioning event, i.e., the way in which the incoming information is used in singling out the set of sequences on which to concentrate the probability mass. For a given sentence φ , the two options are repeated in (17). I argued above that deep conditioning is what is required to model belief update by conditionalization.

- (17) a. $\{\omega^* \in \Omega^* | V^*(\varphi)(\omega^*) = 1\}$ *[shallow]*
 b. $\{\omega^* \in \Omega^* | V^*(\varphi)(\omega^{n*}) = 1 \text{ for all } n\}$ *[deep]*

In addition, I believe it is likely that **both** update operations are required to model the interpretation of conditionals. To see this, consider first what the difference comes down to in terms of the intuitive interpretation of the two versions of update.

Recall first the intuitive interpretation of a Stalnaker Bernoulli model as a representation of epistemic states. The idea was that a world sequence represents factual information in the first world, and conditional information in the tail following the first world. As already mentioned in footnote 9, from the modal-logic perspective on the representation of knowledge and belief, one can alternatively think of the sequences as representing all possible paths along an introspective doxastic accessibility relation. This intuitive interpretation of world sequences corresponds well with the proposal to model belief update by conditionalization on the set of sequences throughout which the incoming sentence is true (as opposed to the set of sequences at which the sentence evaluates to 1).

With all this in mind, I believe that we can discern in the distinction between shallow and deep conditioning an intuitively real difference in ways to “hypothetically add,” in Ramsey’s words, a sentence φ to one’s probabilistic “stock of knowledge”: Shallow update consists in assuming that φ is **true**, whereas deep update consists in assuming that φ is **learned**.

The distinction between these two modes of update underlies certain problematic examples which have been discussed widely in the literature on conditionals. Among them are simple conditionals like (18a), attributed to Richmond Thomason by van Fraassen (1980), and (18b) from Lewis (1986).

- (18) a. If my wife deceives me, I won’t believe it.
 b. If Reagan works for the KGB, I’ll never believe it.

In both cases, the conditional antecedent is intuitively not correctly paraphrased by any variant of the locution ‘*If I know/learn that*’ . . . , for otherwise it would not make sense together with the consequent. Yet these examples are perfectly well-formed and interpretable. What the examples show is that it is possible for speakers to suppose that something is the case unbeknownst to them. The present framework opens up a way to model the difference in a probabilistic setting.

Another difference between the two modes of update concerns the interpretation of right-nested conditionals, i.e., sentences of the form $B \rightarrow (C \rightarrow D)$. Here we see a tangible consequence of the fact that the update with the antecedent B determines the context for the interpretation of the consequent $C \rightarrow D$. Theorem 2 states the general result for deep conditioning with sentences of this form.

Theorem 2 *If $P(BC) > 0$, then $P^*(C \rightarrow D|B^*) = P(D|BC)$.*

Theorem 1 is a special case of Theorem 2, again obtained by substituting the tautology for B .

What is notable about this result is that it resembles the **Conditional Import-Export Principle**, i.e., the rule that the conditional probability of $\varphi \rightarrow \psi$ given χ should equal the conditional probability of ψ given $\varphi\chi$. Theorem 2 shows that the Import-Export Principle holds for right-nested conditionals under deep conditioning on the antecedent. On the other hand, it does not hold under the standard Stalnaker Bernoulli interpretation. Recall from (9c) above that the expectation of the values assigned to such right-nested conditionals is quite different:

$$(9c) \text{ If } P(B) > 0 \text{ and } P(C) > 0, \text{ then} \\ P^*(B \rightarrow (C \rightarrow D)) = P(CD|B) + P(D|C)P(\bar{C}|B)$$

There is much room for further explorations into the ramifications of this distinction, for right-nested conditionals as well as for the others listed in (9b) and (9d) above. These investigations are left to future work.

7 Conclusions

I have argued that it makes good sense to investigate the utility of Stalnaker Bernoulli models as a formal tool for the representation and analysis of belief states and their dynamics. The main arguments in favor of such a move draw on the viability of an account long these lines that is immune to Lewisian triviality results. But I also showed that once we start to investigate the matter seriously, a number of additional subtleties and advantages come into view which merit further study. The difference between shallow and deep conditioning and its potential applications in the analysis of counterexamples to the Import-Export Principle was one such example. Additionally, the framework promises to offer a straightforward account of belief updates involving conditionals, whose theoretical implementation and empirical verification remain to be carried out. And all of this, in a sense, leads up to the big question: whether it is time to do away with the time-honored but simplistic possible-worlds models, if only to get the probabilities of conditionals right.

Appendix: Proofs

Proposition 1 For $X \in \mathcal{F}$, if $\Pr(X) > 0$, then $\Pr^* \left(\bigcup_{n \in \mathbb{N}} \left(\overline{X}^n \times X \times \Omega^* \right) \right) = 1$.

Proof Notice that $\bigcup_{n \in \mathbb{N}} \left(\overline{X}^n \times X \times \Omega^* \right)$ is the set of all sequences containing at least one X -world, thus its complement is \overline{X}^* . Now

$$\begin{aligned} \Pr^* \left(\bigcup_{n \in \mathbb{N}} \left(\overline{X}^n \times X \times \Omega^* \right) \right) &= 1 - \Pr^* \left(\overline{X}^* \right) \\ &= 1 - \lim_{n \rightarrow \infty} \Pr^* \left(\overline{X}^n \times \Omega^* \right) = 1 - \lim_{n \rightarrow \infty} \Pr \left(\overline{X} \right)^n = 1 \text{ since } \Pr(\overline{X}) < 1. \quad \square \end{aligned}$$

Lemma 1 If $\Pr(X) > 0$, then $\sum_{n \in \mathbb{N}} \Pr \left(\overline{X} \right)^n = 1/\Pr(X)$.

Proof $\sum_{n \in \mathbb{N}} \Pr \left(\overline{X} \right)^n \times \Pr(X) = \sum_{n \in \mathbb{N}} \left(\Pr \left(\overline{X} \right)^n \times \Pr(X) \right)$
 $= \sum_{n \in \mathbb{N}} \Pr^* \left(\overline{X}^n \times X \times \Omega^* \right) = \Pr^* \left(\bigcup_{n \in \mathbb{N}} \left(\overline{X}^n \times X \times \Omega^* \right) \right) = 1$ by Proposition 1. \square

Theorem 1 For $A, C \in \mathcal{L}_A^0$, if $\Pr(A) > 0$, then $\Pr^*(A \rightarrow C) = \Pr(C|A)$.

Proof By Definition 6, the set of sequences ω^* such that $V^*(A \rightarrow C)(\omega^*) = 1$ is the union $\bigcup_{n \in \mathbb{N}} \left(\{\omega \in \Omega | V(A)(\omega) = 0\}^n \times \{\omega \in \Omega | V(C)(\omega) = 1\} \times \Omega^* \right)$. Since the sets for different values of n are mutually disjoint, the probability of the union is the sum of the probabilities for all n . Now, for all $X, Y \in \mathcal{F}$,

$$\begin{aligned} \Pr^* \left(\bigcup_{n \in \mathbb{N}} \left(\overline{Y}^n \times (X \cap Y) \times \Omega^* \right) \right) &= \sum_{n \in \mathbb{N}} \Pr^* \left(\overline{Y}^n \times (X \cap Y) \times \Omega^* \right) \\ &= \sum_{n \in \mathbb{N}} \left(\Pr \left(\overline{Y} \right)^n \times \Pr(X \cap Y) \right) = \sum_{n \in \mathbb{N}} \Pr \left(\overline{Y} \right)^n \times \Pr(X \cap Y) \\ &= \Pr(X \cap Y)/\Pr(Y) \text{ by Lemma 1.} \end{aligned}$$

In particular, let X, Y be the set of worlds in Ω at which $V(C)$ and $V(A)$ are true, respectively. \square

Proposition 2 If $\Pr(Z) > 0$, then $\Pr^*(Z^*|Z^*) = 1$.

Proof $\Pr^*(Z^*|Z^*) = \lim_{n \rightarrow \infty} \Pr^*(Z^n \times \Omega^* | Z^n \times \Omega^*) = 1$. \square

Proposition 3 If $\Pr^*(\mathbf{X}|\mathbf{Y})$ is defined, then $\Pr^*(\mathbf{X}|\mathbf{Y}) \times \Pr^*(\mathbf{Y}) = \Pr^*(\mathbf{X} \cap \mathbf{Y})$.

Proof Since $\Pr^*(\mathbf{X}|\mathbf{Y})$ is defined, $\Pr(Y_n) > 0$ for all n . Thus $\Pr^*(\mathbf{X}|\mathbf{Y}) \times \Pr^*(\mathbf{Y})$
 $= \lim_{n \rightarrow \infty} \frac{\Pr^*(X_1 \cap Y_1 \times \cdots \times X_n \cap Y_n \times \Omega^*)}{\Pr^*(Y_1 \times \cdots \times Y_n \times \Omega^*)} \times \lim_{n \rightarrow \infty} \Pr^*(Y_1 \times \cdots \times Y_n \times \Omega^*)$
 $= \lim_{n \rightarrow \infty} \left(\frac{\Pr^*(X_1 \cap Y_1 \times \cdots \times X_n \cap Y_n \times \Omega^*)}{\Pr^*(Y_1 \times \cdots \times Y_n \times \Omega^*)} \times \Pr^*(Y_1 \times \cdots \times Y_n \times \Omega^*) \right)$
 $= \lim_{n \rightarrow \infty} \Pr^*(X_1 \cap Y_1 \times \cdots \times X_n \cap Y_n \times \Omega^*) = \Pr^*(\mathbf{X} \cap \mathbf{Y})$ \square

Proposition 4 If $\Pr^*(\mathbf{X}|\mathbf{Y})$ is defined and $\mathbf{Y} \subseteq \mathbf{X}$, then $\Pr^*(\mathbf{X}|\mathbf{Y}) = 1$.

Proof For all $i \geq 1$, $X_i \cap Y_i = Y_i$ since $\mathbf{Y} \subseteq \mathbf{X}$, and $\Pr(Y_i) > 0$ since $\Pr^*(\mathbf{X}|\mathbf{Y})$ is defined. Thus $\Pr^*(\mathbf{X}|\mathbf{Y}) = \lim_{n \rightarrow \infty} \frac{\Pr^*(Y_1 \times \cdots \times Y_n \times \Omega^*)}{\Pr^*(Y_1 \times \cdots \times Y_n \times \Omega^*)} = 1$. \square

The following auxiliary result will be useful in the subsequent proofs.

Proposition 9 *If $\Pr(Z) > 0$, then for $X_i \in \mathcal{F}$ and $n \in \mathbb{N}$,*

$$\Pr^*(X_1 \times \cdots \times X_n \times \Omega^* | Z^*) = \Pr^*(X_1 \times \cdots \times X_n \times \Omega^* | Z^n \times \Omega^*)$$

Proof Immediate because $Z \subseteq \Omega$. \square

The significance of Proposition 9 derives from the fact that the sets of sequences at which a given sentence in \mathcal{L}_A is true can be constructed (using set operations under which \mathcal{F}^* is closed) out of sequence sets ending in Ω^* . This is obvious for the non-conditional sentences in \mathcal{L}_A^0 . For conditionals $\varphi \rightarrow \psi$ the relevant set is the union of sets of sequences consisting of n $\bar{\varphi}$ -worlds followed by a $\varphi\psi$ -world. For each n , the corresponding set ends in Ω^* and therefore can be conditioned upon Z^* as shown in Proposition 9. Since these sets for different numbers n are mutually disjoint, the probability of their union is just the sum of their individual probabilities.

Proposition 5 *If $\Pr(X \cap Z) > 0$, then $\Pr^*\left(\bigcup_{n \in \mathbb{N}} (\bar{X}^n \times X \times \Omega^*) \middle| Z^*\right) = 1$.*

Proof Since $\Pr(X \cap Z) > 0$, $\Pr(\bar{X} \cap Z) < \Pr(Z)$. Thus

$$\begin{aligned} \Pr^*\left(\bigcup_{n \in \mathbb{N}} (\bar{X}^n \times X \times \Omega^*) \middle| Z^*\right) &= 1 - \Pr^*(\bar{X}^* | Z^*) \\ &= 1 - \lim_{n \rightarrow \infty} \frac{\Pr^*((\bar{X} \cap Z)^n \times \Omega^*)}{\Pr^*(Z^n \times \Omega^*)} = 1 - \lim_{n \rightarrow \infty} \frac{\Pr(\bar{X} \cap Z)^n}{\Pr(Z)^n} \\ &= 1 - \lim_{n \rightarrow \infty} \left(\frac{\Pr(\bar{X} \cap Z)}{\Pr(Z)}\right)^n = 1 \text{ since } \Pr^*(\bar{X} \cap Z) < \Pr^*(Z) \end{aligned} \quad \square$$

Lemma 2 *If $\Pr(X \cap Z) > 0$, then $\sum_{n \in \mathbb{N}} \Pr(\bar{X}|Z)^n = 1/\Pr(X|Z)$.*

Proof Since $\Pr(X \cap Z) = \Pr(Z) \times \Pr(X|Z)$, both $\Pr(Z) > 0$ and $\Pr(X|Z) > 0$.

$$\begin{aligned} \sum_{n \in \mathbb{N}} \Pr(\bar{X}|Z)^n \times \Pr(X|Z) &= \sum_{n \in \mathbb{N}} \left(\Pr(\bar{X}|Z)^n \times \Pr(X|Z)\right) \\ &= \sum_{n \in \mathbb{N}} \frac{\Pr(\bar{X} \cap Z)^n \times \Pr(X \cap Z)}{\Pr(Z)^{n+1}} = \sum_{n \in \mathbb{N}} \frac{\Pr^*((\bar{X} \cap Z)^n \times (X \cap Z) \times \Omega^*)}{\Pr^*(Z^{n+1} \times \Omega^*)} \\ &= \sum_{n \in \mathbb{N}} \Pr^*(\bar{X}^n \times X \times \Omega^* | Z^{n+1} \times \Omega^*) \\ &= \sum_{n \in \mathbb{N}} \Pr^*(\bar{X}^n \times X \times \Omega^* | Z^*) \text{ by Proposition 9} \\ &= \Pr^*\left(\bigcup_{n \in \mathbb{N}} (\bar{X}^n \times X \times \Omega^*) \middle| Z^*\right) = 1 \text{ by Proposition 5.} \end{aligned} \quad \square$$

Theorem 2 *If $\Pr(BC) > 0$, then $\Pr^*(C \rightarrow D|B^*) = \Pr(D|BC)$.*

$$\begin{aligned}
 \text{Proof } P^*(C \rightarrow D|B^*) &= \lim_{n \rightarrow \infty} \frac{P^*((C \rightarrow D) \wedge B)^n \times \Omega^*}{P^*(B^n \times \Omega^*)} \\
 &= \lim_{n \rightarrow \infty} \frac{\sum_{i=0}^{n-1} P(\overline{C}B)^i \times P(CDB)}{P(B)^n} \\
 &= \lim_{n \rightarrow \infty} \sum_{i=0}^{n-1} P(\overline{C}|B)^i \times P(CD|B) \\
 &= P(CD|B)/P(C|B) \text{ by Lemma 2} \\
 &= P(D|BC)
 \end{aligned}$$

□

References

- Abbott, B. (2004). Some remarks on indicative conditionals. In R. B. Young (Ed.), *Proceedings of SALT* (Vol. 14, pp. 1–19). Ithaca: Cornell University.
- Adams, E. (1965). The logic of conditionals. *Inquiry*, 8, 166–197.
- Adams, E. (1975). *The logic of conditionals*. Dordrecht: D. Reidel.
- Bennett, J. (2003). *A philosophical guide to conditionals*. Oxford: Oxford University Press.
- Edgington, D. (1995). On conditionals. *Mind*, 104(414), 235–329.
- Eells, E., & Skyrms, B. (Eds.). (1994). *Probabilities and conditionals: Belief revision and rational decision*. Cambridge: Cambridge University Press.
- Evans, J. S. B. T., & Over, D. E. (2004). *If*. Oxford: Oxford University Press.
- Fetzer, J. H. (Ed.). (1988). *Probability and Causality* (Vol. 192), Studies in Epistemology, Logic, Methodology, and Philosophy of Science Dordrecht: D. Reidel.
- van Fraassen, B. C. (1976). Probabilities of conditionals. In W. L. Harper, R. Stalnaker, & G. Pearce (Eds.), *Foundations of probability theory, statistical inference, and statistical theories of science. The University of Western Ontario Series in Philosophy of Science* (Vol. 1, pp. 261–308) Dordrecht: D. Reidel.
- van Fraassen, B. C. (1980). Review of Brian Ellis, rational belief systems. *Canadian Journal of Philosophy*, 10, 457–511.
- Gibbard, A. (1981). Two recent theories of conditionals. In Harper, W. L., Stalnaker, R., & Pearce, G. (Eds.), (pp. 211–247).
- Hájek, A. (1994). Triviality on the cheap?. In Eells, E. & Skyrms, B. (Eds.), (pp. 113–140).
- Hájek, A. (2003). What conditional probability could not be. *Synthese*, 137, 273–323.
- Hájek, A. (2011). Conditional probability. In P. S. Bandyopadhyay & M. R. Forster (Eds.), *Philosophy of statistics. Handbook of the philosophy of science* (Vol. 7). Elsevier B.V. (Series editors: D.M. Gabbay, P. Thagard & J. Woods
- Hájek, A. (2012). The fall of Adams' thesis? *Journal of Logic, Language and Information*, 21(2), 145–161.
- Hájek, A., & Hall, N. (1994). The hypothesis of the conditional construal of conditional probability. In E. Eells & B. Skyrms (Eds.), (pp. 75–110).
- Harper, W. L., Stalnaker, R., & Pearce, G. (Eds.). (1981). *Ifs: Conditionals, belief, decision, chance, and time*. Dordrecht: D. Reidel.
- Jackson, F. (1979). On assertion and indicative conditionals. *Philosophical Review*, 88, 565–589.
- Jeffrey, R. C. (1964). If. *Journal of Philosophy*, 61, 702–703.
- Jeffrey, R. C. (1991). Matter-of-fact conditionals. *The Symposia Read at the Joint Session of the Aristotelian Society and the Mind Association at the University of Durham* (pp. 161–183). The Aristotelian Society, July 1991. Supplementary Volume 65.
- Jeffrey, R. C. (2004). *Subjective probability: The real thing*. Cambridge: Cambridge University Press.
- Kaufmann, S. (2005). Conditional predictions: A probabilistic account. *Linguistics and Philosophy*, 28(2), 181–231.

- Kaufmann, S. (2009). Conditionals right and left: Probabilities for the whole family. *Journal of Philosophical Logic*, 38, 1–53.
- Lewis, D. (1973). *Counterfactuals*. Cambridge: Harvard University Press.
- Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *Philosophical Review*, 85, 297–315.
- Lewis, D. (1986). Postscript to Probabilities of conditionals and conditional probabilities. *Philosophical papers* (Vol. 2, pp. 152–156). Oxford: Oxford University Press.
- Mellor, D. H. (Eds.). (1990). *Frank Ramsey: Philosophical papers*. Cambridge: Cambridge University Press.
- Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101, 608–631.
- Oaksford, M., & Chater, N. (2003). Conditional probability and the cognitive science of conditional reasoning. *Mind and Language*, 18(4), 359–379.
- Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford: Oxford University Press.
- Ramsey, F. P. (1929). *General propositions and causality*. Reprinted in Mellor (1990), (pp. 145–163).
- Stalnaker, R. (1968). A theory of conditionals. *Studies in logical theory. American Philosophical Quarterly, Monograph* (Vol. 2, pp. 98–112). Blackwell.
- Stalnaker, R. (1970). Probability and conditionals. *Philosophy of Science*, 37, 64–80.
- Stalnaker, R. (1981). A defense of conditional excluded middle. In W. Harper, et al. (Eds.), (pp. 87–104).
- Stalnaker, R., & Jeffrey, R. (1994). Conditionals as random variables. In E. Eells & B. Skyrms (Eds.), (pp. 31–46).

On the Probabilistic Notion of Causality: Models and Metalanguages

Christian Wurm

Abstract We present some formal, interpreted languages, which are based on propositional logic, but have each a new connective with a semantics based on (conditional) probabilities, which is related to the notion of causality. The first one is well-known as meta-language of Bayesian networks; the other two are new and seem to match better our intuition of a causal connective. We provide definitions of truth and validity, as well as some elementary model theory, in particular focussing on the questions: which properties of probability spaces can be axiomatized by formulas of the new languages, and which not? In particular, we can show that desirable properties of expressive power come at the cost of counterintuitive features.

Keywords Causality · Probabilistic model theory · Bayesianism · Logic · Probability spaces · Conditional independence

1 Introduction

This is a study in the probabilistic approach to the notion of causality by means of probabilistic models and logical metalanguages. The philosophical motivation is the following: in the probabilistic study of causality by means of graphs (see the next section), though we have good results, there are some things unsatisfying: we often want to speak about causal dependencies in a probability space directly in a logical language without a detour over graphs or multivariate functions. We therefore look at the (well-known) meta-language \mathcal{BL} with a direct interpretation in probability spaces. The spaces we consider are defined over possible worlds for classical propositional logic, and allow to define the relations of satisfaction and validity.

Still, \mathcal{BL} is rather alien to our intuitive way to talk about causality; we therefore introduce two new languages \mathcal{L} , \mathcal{L}' which are closer to our intuitions, having just a binary predicate for a causal relation. Next, we investigate the expressive power of

C. Wurm (✉)

Universität Düsseldorf, Dusseldorf, Germany

e-mail: cwurm@phil.uni-duesseldorf.de; cwurm@phil.hhu.de

© Springer International Publishing Switzerland 2015

H. Zeevat and H.-C. Schmitz (eds.), *Bayesian Natural Language Semantics*

and *Pragmatics*, Language, Cognition, and Mind 2,

DOI 10.1007/978-3-319-17064-0_5

our languages in terms of axiomatizing probability spaces, showing that at least \mathcal{L}' is sufficiently expressive to encode the probabilistic notions of causality, but that this expressive power comes with several drawbacks. This is a philosophically significant result, which might give new inspirations to Bayesian approaches to causality. But also the techniques applied are quite new: though there is a long common history of probability, logic and model theory (just by way of example, see Wachter et al. 2007; Schurz and Leitgeb 2008; Ebbinghaus and Flum 1995, Chap. 3), questions of axiomatization of probability spaces by logics enriched with probabilistic connectives have not been addressed yet to our knowledge. So beyond its philosophical relevance, this work can be seen as establishing new ties between logic and probability theory beyond the already existing ones; in particular, the axiomatic approach to probability spaces.

2 Background, Goals and Challenges

2.1 The Probabilistic Notion of Causality

As there has been considerable work on the probabilistic notion of causality, we will not question or justify this approach, but rather take it as a given frame. The central issue within the probabilistic study of causality is the question whether the information given by probabilistic interactions (in terms of conditional probabilities and independence of events) is sufficient to establish the *directionality* of causal influence. This is because we intuitively conceive of causality as an asymmetric relation, whereas in a probabilistic setting, the interaction between two random variables (or events) in terms of conditional probabilities is always symmetric. In order to be able in principle to establish an asymmetric connection, we need at least a third variable, which interacts in a certain way, which is called *intervention*. We will not give a general introduction to this kind of interaction, and refer to the literature (Pearl 2009; Koller and Friedman 2009; Woodward 2003 for general reference, Dehgani et al. 2012 for recent work and overview).

The usual approach in this line of research consists in Bayesian networks (BN). But these are in a sense too specific to give a good notion of causality as we intuitively conceive of it: apart from the fact that BN specify precise numeric values, which are of no particular interest to a study of causality, they can be arbitrarily redundant, so we always need the additional premise that they be minimal. What therefore is more useful is a metalanguage, as it has been scrutinized in Geiger and Pearl (1990), the basic statements of which have the form $X_1 \parallel X_2 | X_3$, which means, for random variables (or alternatively, events) X_1, X_2, X_3 : the distribution of X_1 is independent of the one of X_2 , if the value of X_3 is known. If we allow the X_i to be arbitrary finite sets of variables (conjunctions of events), this one ternary connective is fully sufficient to express the dependence and independence relations of Bayesian networks, if we grant us in addition the usual Boolean negation, thereby getting $X_1 \not\parallel X_2 | X_3$ as

$\neg(X_1 \parallel X_2 | X_3)$). This metalanguage of Bayesian networks allows us to encode the (conditional) dependence and independence of variables, as it can be encoded in any (minimal) finite BN, and therefore is sufficiently expressive to let us infer the directionality of causation, even though there is no predicate for directed causation (the predicate of conditional dependence and independence is obviously symmetric in its first two arguments). Consequently, this is a sufficiently rich logical language for the probabilistic notion of causality, which consists of propositional variables, Boolean connectives, and a single ternary connective as above. Henceforth, we will use the term “sufficiently expressive” in exactly this sense.

2.2 Goals and Challenges of This Paper

There is a problem though with the metalanguage of Bayesian networks: we intuitively think of causation as a *binary* relation, not a ternary one. The main research question for us here is the following: *can we enrich a classical logical language with a binary connective C , which is interpreted as causal dependence, such that the resulting language is sufficiently expressive?* This is the main research question for this paper. A number of questions follow up immediately:

1. In which way are we to interpret statements of our language in probability spaces?
2. What is the interpretation of C ?
3. How is causality related to truth: does $C(\phi_1, \phi_2)$ entail the truth of ϕ_1, ϕ_2 ?
4. Given that the interpretation of C will involve conditional probabilities, how should we interpret embedded statements?

As for questions 1 and 2: we use probability spaces over possible worlds, where we interpret logical statements in the usual way, and probabilities over *definable* sets of worlds (i.e. propositions). $C(\phi_1, \phi_2)$ for us intuitively means that $P(\sigma(\phi_2)) < P(\sigma(\phi_2) | \sigma(\phi_1))$ (by $\sigma(-)$ we mean the denotation of a formula). This basic semantics leaves however several options regarding the definition of satisfaction; we will investigate two possibilities for two languages $\mathcal{L}, \mathcal{L}'$, where \mathcal{L}' yields a sufficiently expressive language, \mathcal{L} does not. So there is a positive answer to our main research question.

Question 3 has a philosophical aspect: one might sustain causality only relates *instances* of events, if we think of it as an intrinsic, physical property; in this sense, causality presupposes truth. On the other hand, one might also sustain that causality only relates *types* of events in terms of expected co-occurrence frequency (conditional probability); but as soon as two events really occur, it does not make sense to speak of causality relating the occurrences. Our two languages $\mathcal{L}, \mathcal{L}'$ reflect these two options: in $\mathcal{L}, C(\phi, \chi)$ implies ϕ, χ , whereas in \mathcal{L}' , if ϕ, χ hold, then $C'(\phi, \chi)$ becomes trivially false. The interesting thing is that this is less a philosophical choice than a technical necessity, stemming from choices we have to make to ensure that \mathcal{L}' is sufficiently expressive. So question 3 has a technical answer, if we want to stick with our goal of being sufficiently expressive, which in turn has philosophical implications.

Question 4 turns out to be related: \mathcal{L} allows for meaningful embedding of causal statements; \mathcal{L}' does not. Again, the latter is not a matter of philosophical choice, but rather of technical necessity, if we want to make \mathcal{L}' sufficiently expressive. So here again technical problems relate to philosophical questions (in a rather negative way).

To sum up our main results: we can enrich a simple propositional logical language with a binary connective to yield a sufficiently expressive language for probabilistic causality, but then we have to interpret causality as being disconnected from truth, and we cannot assign a non-trivial meaning to embedded causal statements. We have to emphasize, though, that our results are not definitive in that they exclude the possibility of different solutions to the same problems: the main guideline for our treatment was that for our meta-languages, syntax, semantics and the relation between the two should be as simple and natural as possible. Maybe a more complicated ontology might yield different results.

2.3 Some Remarks on Truth and Probability

To justify our notion of probabilistic models, we make some remarks on the relation of truth and probability, in particular in how far they can be independent. If we consider probabilities in a context where also truth and falsity play a role, we often tend to equate truth with a probability of 1, and falsity with probability 0; propositions of which we do not know whether they are true or not get assigned something in between, so that probabilities result to be some sort of truth degrees. This is however not the Bayesian approach to probability: if we consider probabilities as being epistemic in nature, then probabilities come to affirm not truth degrees, but rather degrees of *belief* of truth. So truth and probability come apart and can be neatly separated: there are many things being true of which we do not believe they are true; and we do believe of many things that they are true which are not true. From this point of view, it is natural to consider the truth of a proposition independently from its probability.

For us, the main reason to treat probability and truth independently is the following: a certain proposition is something which is mostly only true and wrt. to a state of affairs, not in general (distinguishing “contingent propositions” from mathematical truths, to which, in fact, it makes little sense to assign probabilities). If truth is *situated*, that is, made only wrt. a certain state which can easily change, this means that the situated truth is independent (at least in one direction) from the general truth of a proposition. Same holds for probabilities: we can talk about the probability as the belief that something is the case wrt. a certain state of affairs (for example the current), and its probability in general. It is not the case that a speaker can make claims on the truth *and* the probability of a proposition p independently. But if he sustains its *situated truth*, he does not sustain its general truth, and nothing follows for the *general belief* that p might be the case. That in turn means: a speaker sustaining p can nonetheless assign a quite low general probability to it. This is mirrored by the fact that we all think that improbable things can happen—if a certain fact holds, it is not automatically assigned probability 1, because in a sense, the probability is for

the general case, the truth for the particular one, depending on assignments etc. We think that whereas truth by default is situated, probability by default is generic; and this is the way in which we will (on an intuitive level) treat the two.

It is a valid remark that generic and situated propositions might be different kinds of entities which should be separated by our ontology. While this might be feasible with a richer ontology and intensional language, we only work with the language of propositional classical logic. This does not allow any distinction between intensional and extensional propositions, and there can be no doubt that we want to be able to assign both truth values *and* probabilities to the same propositions. Still, the difference is reflected in \mathcal{L} by the fact that truth depends on the “actual world” (as opposed to validity), whereas probabilities are independent from actuality; for \mathcal{L}' , things are more complicated.

If we assign truth values and probabilities to propositions independently, we can separate the logical and the probabilistic meaning of connectives quite neatly, which is very useful. What is particularly interesting is that probabilities can be related to truth values in quite different ways, and for \mathcal{L} , \mathcal{L}' , there are different underlying philosophical conceptions, as we will see. Our general road is the following: we want propositions to be either true or false (depending on the model); but truth might depend on probabilities in a way that satisfaction has to be defined both wrt. an “actual world” and a probability distribution. This separates our approach from the one taken for example in Kaufmann (2009), where some propositions (conditionals) actually “denote” a probability rather than a truth value.

2.4 Intuitions on the Causal Connective

As we said, we intend to enrich the language of propositional logic with a single binary connective C (or C' , respectively) for the causal relation. If we say that this maps our intuition for the relation of causality, this is obviously because we think of it as analogous to the English (or whatever language we prefer) connective *because*. However, we should make the following clear from the outset: though C is intended to be a “causal” connective, that is, one corresponding to our usage of *because*, we have no intention to thereby model or cover every usage of *because* in natural language, but only some rather canonical and idealized portion of it. To justify this procedure, just note that the same holds for the logical constants: the classical \wedge , \neg can claim to exactly match the English “and”, “not” at most for the domain of mathematics, and hardly beyond. There is no bad thing in modelling just some idealized, canonical usage; but the question we have to ask is then: in what consists this canonical usage? This question seems hard to answer, but there are some things we can definitely exclude. For example, we should model *because* as a relation between states of affairs; and not as a relation involving states of mind of persons, or as giving a motivation for some action of some person. So 1. should be out, 2. should be in;

- (1) a. Sheila was right, because she thought that from the beginning.
 b. Mary left because she was disappointed.
- (2) a. There is smoke because there is a fire.
 b. Because he paid little attention to traffic, he had that accident.

We believe there are deep underlying reasons for the fact that we are unable to capture usages as (1) in terms of probabilities: humans generally believe that things behave according to certain rules, and thus we have the correspondence $\text{frequency} \approx \text{probability} \approx \text{causality}$; but whenever we think that actions are performed by agents having possible choices, this breaks down: we can virtually do for anything for any reason, and so the general case does not (fully) determine the particular one. We cannot dwell on this topic, so we just point to Woodward (2006), who treats a strongly related issue, and having said this as a broad guideline, we have to comment on some other properties of *because*.

What is the truth-functional meaning of *because*? One line of reasoning goes as follows: (i) the “truth-functional portion of the meaning of *because* seems to be equal to the meaning of *and*, put differently: *because* presupposes the truth of its argument clauses, “*because A, B*” cannot be true without A, B being true. But one might also argue: (ii) given the considerations of the last section, we can sustain that “*because A, B*” contains two separate assertions: 1. in general, A being true causes B being true, and 2. in this particular situation, both A and B are true. Given this, the second assertion is uninteresting for our purposes, and we can say that a formal model of *because* only should model the first part of its meaning. In the language \mathcal{L} , we follow the reasoning (i), in \mathcal{L}' , we follow the reasoning (ii), and only the latter yields a sufficiently expressive language.

Regarding the probabilistic meaning of a formal counterpart of *because*, there are surely many things one can say; however, we will choose the most blunt and simple semantics we can think of, both in \mathcal{L} and \mathcal{L}' : we take “*because A, B*” to have the probabilistic meaning: the probability of B is (strictly) smaller than the probability of B given A. It seems easy to find examples in natural language where this semantics does not work (even with “canonical” *because*); yet, it seems hard to find a formal alternative which works better. Having said this, it seems clear we model *because* in its epistemic sense, that is, the sentences in (2) and the following are equivalent from a probabilistic point of view (because of Bayes theorem):

- (3) a. There is a fire, because there is smoke.
 b. He paid little attention to traffic, because he had that accident.

That is of course problematic, as in the sentences (2) and (3), we intuitively use *because* in a different sense: where in the former, we read *because* as giving a proper reason in a physical sense, in the latter it is only epistemic. But as we said, in itself the symmetry of the probabilistic semantics of *because* is unproblematic, as long as we are able to axiomatize in our languages \mathcal{L} , \mathcal{L}' probability distributions which are rich enough to allow to *infer* the direction of causality.

3 Probabilistic Models

Having laid out the philosophical justification for our approach, we postpone further difficulties and introduce the formalism. We directly start with the language of propositional logic PL , defined as follows: for var a countable set of variables (which we assume to be fixed for all that is to follow), we have

1. $p \in PL$, if $p \in var$;
2. if $\phi, \psi \in PL$, then $\phi \wedge \psi \in PL$;
3. if $\phi \in PL$, then $\neg\phi \in PL$;
4. nothing else is in PL .

This is the language PL ; it is well-known how to define \vee, \rightarrow in terms of \wedge, \neg , and so we will make use of them, too. A **possible world** over PL is a set $m \subseteq PL$, which is *maximally consistent*, that is, it is consistent and any proper superset of it is inconsistent.¹ Let M be the set of possible worlds over PL . We have the usual satisfaction relation \models , where for $m \in M, \phi \in PL$, we write $m \models \phi$ iff ϕ is true in m according to our interpretation of PL , that means in our case: $m \models \phi$ iff $\phi \in m$. In the sequel, we will enlarge and modify the language PL , but possible worlds will always be defined wrt. this language.

We can now define **probabilistic models** in a straightforward fashion.² Assume PL, M as given, and assume $\sigma : PL \rightarrow \wp(M)$ to be the interpretation function for PL in M , where $m \in \sigma(\phi)$ iff $m \models \phi$ (iff $\phi \in m$). We follow common usage by calling a set of the form $\sigma(\phi)$ a **proposition** (wrt. PL, M). A probabilistic model of PL, M is a probability space (\mathcal{M}, P, M) , where the usual probability axioms hold, that is:

1. $\mathcal{M} \subseteq \wp(M)$ is a Boolean algebra, $P : \mathcal{M} \rightarrow [0, 1]$,
2. $P(M) = 1$,
3. $P(\emptyset) = 0$, and
4. if M_1, \dots, M_n are pairwise disjoint, then $P(\bigcup_{i=1}^n M_i) = \sum_{i=1}^n P(M_i)$.

In addition, $\mathcal{M} \subseteq \wp(M)$ is such that for all $\phi \in PL, P(\sigma(\phi))$ is defined. Thus a probabilistic model must assign a probability to any (classical) proposition. In general, we call any $O \subseteq M$ an *event*, usually written E . Note that if PL is classical logic over a countable set of variables, the set M of possible worlds is *uncountable* (that can be shown by a standard argument over the satisfaction of subsets of var), so (\mathcal{M}, P, M) will either not be a discrete space, or most worlds will have probability 0. The same holds, by the way, *a fortiori* for any language with more connectives or less theorems. This already reflects our philosophical considerations: whereas truth of a proposition is defined wrt. to single world, our probability function does usually not even assign (non-zero) probabilities to single worlds, at least in almost all cases.

¹We do not define these notions here, and refer the reader to any logic textbook.

²These definitions roughly coincide with the canonical ones in Kaufmann (2009), though there are some minor differences.

4 Networks and Metalanguages

4.1 Bayesian Networks

A **multivariate function** is a function in many variables. A **discrete probability distribution** is a function $f : M \rightarrow [0, 1]$ such that $\sum_{m \in M} f(m) = 1$; this criterion is called consistency. In what follows, we will just write distribution, and we will always mean discrete probability distribution. A **conditional** (discrete probability) **distribution** is a function $f : M \times N \rightarrow [0, 1]$ such that for all $n \in N$, $\sum_{m \in M} f(m, n) = 1$. A **directed acyclic graph** (DAG) is a graph without cycles. Given a DAG (V, E) , $v \in V$, we define $\downarrow v := \{v' : (v', v) \in E\}$. Most readers will be familiar with this concept. DAGs are the graph structures which underlie Bayesian networks.

Definition 1 A **Bayesian network** BN is a structure (D, V, E) , such that (V, E) is a finite DAG, and D is a function from V to a set $D_v : v \in V$, where each D_v is a consistent, discrete (conditional) probability distribution $D_v = D(X_v | \{X_{v'} : v' \in \downarrow v\})$. We assume that for any BN (D, V, E) , $V \subset \mathbb{N}$ is an initial interval of \mathbb{N} .

We assume that the nodes form an initial interval of \mathbb{N} , because we want to identify the number of a node with the position of an argument variable of a multivariate function. This gives us a uniqueness which otherwise would be obtained only modulo re-ordering of the argument variables and domains. Moreover, we can now associate every argument and domain with a number, and for notational convenience dismiss with the order of variables all together. Bayesian networks themselves do not tell us much for themselves on probability distributions, we first need to interpret them.

Definition 2 We say a multivariate distribution MD over domains and variables $x_v : v \in V$ is a **model** of (D, V, E) , if

$$MD(x_1, \dots, x_{|V|}) = D(X_1 = x_1 | \{X_i = x_i : i \in \downarrow 1\}) \cdot \dots \cdot D(X_{|V|} = x_{|V|} | \{X_i = x_i : i \in \downarrow |V|\})$$

As positions are identified by graph nodes (uniqueness of variable order), we can easily obtain the following lemma:

Lemma 3 *Given a Bayesian network BN, there is exactly one MD which is a model of BN.*

We therefore denote the unique model of BN by $\|BN\|^{MD}$. A thing to keep in mind is: There are many BN, BN' such that $\|BN\|^{MD} = \|BN'\|^{MD}$. That is the most important reason why we have to distinguish between syntax and semantics of Bayesian networks: BNs contain a lot of information, which are irrelevant to their “meaning”, which is the probability distribution they induce. For example, we can model any relation also with a larger network: assume $(V, E), (V, E')$ are two graphs, where $E \subseteq E'$. Then any distribution which can be modelled by a BN over (V, E)

can be modelled with a BN over (V, E') . So what we are in particular interested in are *minimal* Bayesian network for a given distribution. This minimality requirement is usually only formulated with respect to the underlying graph, where it can be made precise through the subset relation on edges.

4.2 Independence and Subdistributions

Assume we have a $MD : A_1 \times \dots \times A_i \rightarrow [0, 1]$. Assume $\{\alpha, \beta\}$ is a partition of $\{1, \dots, i\}$. By \bar{A}_α we denote the Cartesian product of all $A_j : j \in \alpha$; by \bar{a}_α we denote its members. We say that the variables \bar{A}_α and \bar{A}_β are **independent**, if there exist consistent distributions $MD' : \bar{A}_\alpha \rightarrow [0, 1]$, $MD'' : \bar{A}_\beta \rightarrow [0, 1]$, such that for all $\bar{a}_\alpha \in \bar{A}_\alpha$, $\bar{a}_\beta \in \bar{A}_\beta$,

$$MD(\bar{a}_\alpha, \bar{a}_\beta) = MD'(\bar{a}_\alpha) \cdot MD''(\bar{a}_\beta) \quad (1)$$

That means, we can decompose the function into two independent functions by means of multiplication. That however only works for partitions of the entire set of variables. For a more general definition, we need the notion of a **subdistribution**. A MD defines a set \mathfrak{D} , which is a set of functions, its subdistributions. We obtain the subdistributions by firstly marginalizing, and secondly by division.

Firstly, we consider unconditional subdistributions. These are defined as follows. Given $MD : A_1 \times \dots \times A_i \rightarrow [0, 1]$, $\alpha \subseteq \{1, \dots, i\}$, $\beta = \{1, \dots, i\} - \alpha$, we define \mathfrak{D}_{MD}^α by

$$\mathfrak{D}_{MD}^\alpha(\bar{a}_\alpha) = \sum_{\bar{a}_\beta \in \bar{A}_\beta} MD(\bar{a}_\alpha, \bar{a}_\beta) \quad (2)$$

This is commonly called **marginalization** and one of the most well-known techniques in probability theory; therefore, we do not comment on it. We also get a set of conditional distributions, which are defined as follows: for $\beta \subseteq \alpha \subseteq \{1, \dots, i\}$, we define

$$\mathfrak{D}_{MD}^{\alpha|\beta}(\bar{a}_\alpha) = \frac{\mathfrak{D}_{MD}^{\alpha \cup \beta}(\bar{a}_{\alpha \cup \beta})}{\mathfrak{D}_{MD}^\beta(\bar{a}_\beta)} \quad (3)$$

In the sequel, we will stick to standard notations and write $MD(x_{l_1}, \dots, x_{l_i})$ for any subdistribution in variables l_1, \dots, l_i ; and write $MD(x_{l_1}, \dots, x_{l_i} | x_{k_1}, \dots, x_{k_j})$ for conditional distributions. Obviously, there is a close connection between these (conditional) subdistributions and the distributions specified by our BN; as a matter of fact, we can uniquely reconstruct the distributions of a BN given its model MD as conditional subdistributions. A proof of this fact is somewhat involved, but needs nothing but standard arithmetic. Though it might seem obvious, this actually shows

an important result: the semantics we have given is quite strong, in the sense that for a given graph, and a model, the distributions are completely determined:

Lemma 4 *For any multivariate distribution MD, directed acyclic graph G, there exists at most one BN over G of which MD is a model.*

4.3 Speaking About Networks and Distributions—The Metalanguage and Its Models

In this paper, we are at no point interested in precise numerical values, and we think this can be generally said for any work on the notion of causation. Both a BN and an MD come with these values. We rather want to specify probability spaces in terms of (conditional) dependence and independence of events, without any numeric values. We therefore recur on the well-known meta-language of Bayesian networks, and show how it can be directly interpreted in a probability space and probabilistic model. We denote the resulting relation of satisfaction by \models . For introduction of the metalanguage \mathcal{B} , its meaning and motivation, consider Pearl (2009), Geiger and Pearl (1990). Let (E, V, D) be a BN. The following notions refer to its underlying graph (E, V) , which we assume to be directed and acyclic. Forming the reflexive and transitive closure of E , we get a partial order which we denote \leq_E ; notions like ancestor and descendant refer to \leq_E . We say a set $P(v, w) \subseteq V$ is a *path* in (E, V) from v to w , if $P(v, w) := \{v, w, u_1, \dots, u_i\}$, and $(v, u_1), (u_1, u_2), \dots, (u_i, w) \in E \cup E^{-1}$.³ We say u is a *collider* in $P(v, w)$, if for $x, u, y \in P(v, w)$, $x \neq y$, we have $(x, u), (y, u) \in E$; u is a *passerby*, if $w \neq u \neq v$ and u is not a collider. We write

$$v \parallel w | U \text{ (wrt. a DAG } (E, V)),$$

where $v, w \in V, U \subseteq V, v \neq w$, and $v, w \notin U$, if and only if every path $P(v, w)$ in (E, V) contains either a collider $z \notin U$, or a passerby $z \in U$.

That means for the particular case where $U = \emptyset$, that $v \parallel w$ iff v, w are incomparable according to \leq_E and have no common ancestor. Moreover, we write $v \not\parallel w | U$, where $v \neq w$, and $v, w \notin U$, if we do not have $v \parallel w | U$. Call the language consisting of these statements \mathcal{B} . For every finite DAG $G = (E, V)$, there is a finite set of clauses which hold in G ; we call this the **Bayesian theory** of G , written $Th(G)$.

Our language of graphs, though it might seem, is not symmetric in the sense that it is not equally valid of a graph and its dual.⁴ It is symmetric in the first two arguments, but not in the third, and this is what makes the difference. The classical example is a graph $(\{1, 2, 3\}, \{(1, 2), (1, 3)\})$. Here we have $2 \parallel 3 | 1$. If we take its

³Be aware that we allow both arcs and duals; moreover, by acyclicity, at most one of (x, y) and (y, x) is in E ; this makes the following statements unique.

⁴By the dual of a DAG (E, V) we denote the graph (E^{-1}, V) , where $(v, v') \in E^{-1}$ iff $(v', v) \in E$. So the dual is the order-theoretic inverse.

dual $(\{1, 2, 3\}, \{(2, 1), (3, 1)\})$, we rather have $2 \parallel 3$ and $2 \not\parallel 3 \mid 1$. This is the simplest example for a so-called V-structures, and it is by these structures and so-called interventions that we can infer the direction of causality. In the presence of V-structure, the direction of edges can be reconstructed, and this is exactly where the probabilistic conception of causality becomes asymmetric. If we have $(x \parallel y \mid z_1, \dots, z_i)$, but we have $(x \not\parallel y \mid z_1, \dots, z_i, u)$ for some u , then from this we can infer that $x \leq_E u$, $y \leq_E u$. This is straightforward from the definition. The interesting thing is that this notion transfers this to dependence and independence in the model of a Bayesian network with the underlying graph, provided it is minimal.

We take the usual correspondence, by indices, between BN and multivariate distributions; also, we define the subdistributions in the usual fashion. Now let ϕ be a \mathcal{B} -formula. For a given multivariate distribution MD , we write $MD \Vdash \phi$, if the following hold:

1. $M \Vdash x_i \parallel x_j \mid \{x_{l_1}, \dots, x_{l_k}\}$,
if $MD(x_i, x_j \mid x_{l_1}, \dots, x_{l_k}) = M(x_i \mid x_{l_1}, \dots, x_{l_k})M(x_j \mid x_{l_1}, \dots, x_{l_k})$.
2. $M \Vdash x_i \not\parallel x_j \mid \{x_{l_1}, \dots, x_{l_k}\}$, if $M \not\equiv x_i \parallel x_j \mid \{x_{l_1}, \dots, x_{l_k}\}$.

Finally, we say that $M \Vdash Th(G)$, if for all $\phi \in Th(G)$, $M \Vdash \phi$; so we interpret the set on the right hand side as conjunction. One main result on Bayesian causality is the following (actually, we do not know where this is already stated in this form, but it can be easily put together from results in Pearl 2009, pp. 16–19, and Geiger and Pearl 1990):

Theorem 5 *Assume (D, E, V) is minimal, and $MD = \parallel(D, E, V)\parallel^{MD}$. Then we have $\phi \in Th(E, V)$, if and only if $MD \Vdash \phi$.*

This establishes the correspondence between edges in a graph and conditional (in)dependence, and allows to infer the latter from simple graph-theoretic properties. But note that if we define the set of distributions which satisfy a given set of clauses, say $Th(E, V)$, contrary to $\parallel BN \parallel^{MD}$, we do not exclude any larger models/distributions having additional variables, so we do not have uniqueness. This is firstly because \Vdash disregards any kind of numerical values of networks, and secondly because it does not say anything about unknown variables.

4.4 BN and Probabilistic Models

We now have to make clear how we put these tools to use for probabilistic models. We could go the full way over Bayesian networks, making the distribution we associate with each vertex of the underlying graph a random variable mapping a proposition (=set of possible worlds) to $\{0, 1\}$, such that each vertex determines whether a proposition holds. Because most interesting results on the connection of \mathcal{B} and Bayesian networks are well-established, we rather make a shortcut and directly depart from the language \mathcal{B} . This allows us to use all advantages we mentioned above, whereas

Theorem 5 ensures we remain able to reconstruct the notions in terms of Bayesian networks. We now slightly modify the language \mathcal{B} , and form the language \mathcal{BL} as follows: we take PL for granted, and the countable set of propositional variables var as fixed. The formation rules for \mathcal{BL} are simple:

1. If $\phi_1, \phi_2, \phi_3 \in PL$, then $(\phi_1 \parallel \phi_2 | \phi_3) \in \mathcal{BL}$;
2. if $\phi_1, \phi_2 \in PL$, then $(\phi_1 \parallel \phi_2) \in \mathcal{BL}$;
3. if $\phi_1, \phi_2, \phi_3 \in PL$, then $(\phi_1 \nparallel \phi_2 | \phi_3) \in \mathcal{BL}$;
4. if $\phi_1, \phi_2 \in PL$, then $(\phi_1 \nparallel \phi_2) \in \mathcal{BL}$;
5. nothing else is in \mathcal{BL} .

So *all* statements of \mathcal{BL} have a quite similar form. Let (\mathcal{M}, P, M) be a probabilistic model; we define the satisfaction \Vdash as follows:

1. $(\mathcal{M}, P, M) \Vdash (\phi_1 \parallel \phi_2 | \phi_3)$, iff $P(\sigma(\phi_1) \cap \sigma(\phi_2) | \sigma(\phi_3)) = P(\sigma(\phi_1) | \sigma(\phi_3))P(\sigma(\phi_2) | \sigma(\phi_3))$;
2. $(\mathcal{M}, P, M) \Vdash (\phi_1 \parallel \phi_2)$, iff $P(\sigma(\phi_1) \cap \sigma(\phi_2)) = P(\sigma(\phi_1))P(\sigma(\phi_2))$;
3. $(\mathcal{M}, P, M) \Vdash \phi_1 \nparallel \phi_2 | \phi_3$, iff $P(\sigma(\phi_1) \cap \sigma(\phi_2) | \sigma(\phi_3)) \neq P(\sigma(\phi_1) | \sigma(\phi_3))P(\sigma(\phi_2) | \sigma(\phi_3))$;
4. $(\mathcal{M}, P, M) \Vdash (\phi_1 \nparallel \phi_2)$, iff $P(\sigma(\phi_1) \cap \sigma(\phi_2)) \neq P(\sigma(\phi_1))P(\sigma(\phi_2))$.

Given a finite collection \mathbf{B} of \mathcal{BL} -clauses, we say that $(\mathcal{M}, P, M) \Vdash \mathbf{B}$, if $(\mathcal{M}, P, M) \Vdash \phi$ for all $\phi \in \mathbf{B}$.

It is not hard to see that we can also specify a BN, and get from this BN to both its (graph-)theory and a corresponding probability space. Taking this way, the central result would be: we have $(\mathcal{M}, P, M) \Vdash \mathbf{B}$, if and only if there is a minimal Bayesian network (D, E, V) , such that firstly $\mathbf{B} = Th(E, V)$, and secondly, (\mathcal{M}, P, M) is a model of (D, E, V) . However, it is the latter relation of a probability space being a model of a BN which is quite tedious to spell out, and the construction would only reproduce well-known results; so we omit this construction and go directly from \mathcal{BL} to probability spaces.

As we have said, the language of Bayesian networks is powerful enough to express that probabilistic causality is directional, even though there is no way to state this explicitly in \mathcal{BL} . This is for us a kind of touch-stone: we will devise two languages, which come closer to our intuition about causation being a binary predicate, and which subsume the language PL . Actually, the two languages $\mathcal{L}, \mathcal{L}'$ are syntactically isomorphic, but they differ considerably in their semantics: whereas in the first, \mathcal{L} , we will interpret causality as factual, entailing the truth of its arguments, in the latter, we will interpret it “intensionally”,⁵ as being disconnected from truth. As we will see, the factual interpretation is not expressive enough to encode knowledge about directionality, whereas the “intensional” interpretation is.

⁵We write this quotation mark as it does not really reflect the standard meaning of intension.

5 The Language \mathcal{L} and Its Semantics

We now introduce the language \mathcal{L} . It is an extension of propositional logic. Therefore, we start from the usual countable set of propositional variables var ; we assume two connectives, \wedge , \neg , and in addition C , which is binary. \mathcal{L} is now defined as follows:

1. If $p \in var$, then $p \in \mathcal{L}$;
2. if $\phi, \chi \in \mathcal{L}$, then $\phi \wedge \chi, \neg\phi, C(\phi, \chi) \in \mathcal{L}$;
3. nothing else is in \mathcal{L} .

We assume the usual possible world semantics, and construct a probabilistic model over the set of possible worlds. We first construct the classical logical satisfaction \models_l , defined as follows:

1. $(\mathcal{M}, P, M, m) \models_l p$, where $p \in var$, iff $p \in m$.
2. $(\mathcal{M}, P, M, m) \models_l \phi \wedge \chi$, iff $(\mathcal{M}, P, M, m) \models_l \phi$ and $(\mathcal{M}, P, M, m) \models_l \chi$.
3. $(\mathcal{M}, P, M, m) \models_l \neg\phi$, iff $(\mathcal{M}, P, M, m) \not\models_l \phi$.
4. $(\mathcal{M}, P, M, m) \models_l C(\phi, \chi)$, iff $(\mathcal{M}, P, M, m) \models_l \phi$ and $(\mathcal{M}, P, M, m) \models_l \chi$.

It is clear that all other classical connectives can be defined in terms of \wedge , \neg , so we use them freely. Note that C is logically interpreted in the same way as \wedge , and in this sense it is *factual*: $C(\phi, \chi)$ entails ϕ and χ . We define $\sigma : \mathcal{L} \rightarrow \wp(M)$ over the logical satisfaction, that is, $\sigma(\phi) := \{m : (\mathcal{M}, P, M, m) \models_l \phi\}$. Keep in mind that here and in what is to follow, σ refers to \models_l ! We now define a partly *probabilistic* satisfaction \models for \mathcal{L} . Assume $\phi \in \mathcal{L}$. We define $(\mathcal{M}, P, M, m) \models \phi$ by the following clauses:

1. $(\mathcal{M}, P, M, m) \models p$, iff $p \in m$;
2. $(\mathcal{M}, P, M, m) \models \phi \wedge \chi$, iff $(\mathcal{M}, P, M, m) \models \phi$ and $(\mathcal{M}, P, M, m) \models \chi$;
3. $(\mathcal{M}, P, M, m) \models \neg\phi$, iff $(\mathcal{M}, P, M, m) \not\models \phi$;
4. $(\mathcal{M}, P, M, m) \models C(\phi, \chi)$, iff $(\mathcal{M}, P, M, m) \models \phi$, $(\mathcal{M}, P, M, m) \models \chi$ and $P(\sigma(\chi)|\sigma(\phi)) > P(\sigma(\chi))$.

To denote validity, we write $(\mathcal{M}, P, M) \models \phi$, if for all $m \in M$, $(\mathcal{M}, P, M, m) \models \phi$.

The only difference of \models_l and \models is the probabilistic semantics of the connective C . Note that while this is extremely simplistic, it incorporates a number of surprising features: the probabilistic interpretation of C makes reference to a conditional probability, which we defined in the standard fashion. As in \models_l , also \models provides an entailment relation from $C(\phi, \chi)$ to $\phi \wedge \chi$, that is, causation entails truth. From the empirical point of view (departing from the intuitive meaning of *because*), this is fully justified. But it is exactly this choice that comes with considerable formal restrictions and commitments, which ultimately lead to the failure of \mathcal{L} to be sufficiently expressive.

There is usually a problem with using conditional probabilities in an inductive definition: we cannot “nest” conditional probabilities; as a matter of fact, this was one of the major obstacles to the interpretation of conditionals in terms of conditional probabilities, see Kaufmann (2009). For us, the problem does not arise at all:

for $C(\phi, \chi)$ only the *logical* properties of ϕ, χ matter, nothing else; so conditional probabilities do not embed. For the (probabilistic) truth of a C -clause, an embedded C clause just translates as logical conjunction; but for itself, the latter still has to be satisfied probabilistically. Note however that this simple solution is possible only due to our separation of logical and probabilistic satisfaction, and σ being based on the logical satisfaction! We now make this more precise: by $\phi[\chi]$ we mean that χ occurs as a **subformula** in ϕ , where the notion of subformula is defined as follows:

Definition 6 ϕ is a subformula of ψ , only if it can be derived according to the following rules:

1. For all $\phi \in \mathcal{L}$, ϕ is a subformula of ϕ ;
2. if ϕ is a subformula of χ , $\psi \in \mathcal{L}$, then ϕ is a subformula of $\neg\chi$, $\chi \wedge \psi$, $\psi \wedge \chi$, $C(\chi, \psi)$, $C(\psi, \chi)$.

Lemma 7 Let $\phi[C(\psi, \chi)]$ be a formula, where $C(\psi, \chi)$ does not occur within the scope of the connector \neg . Then $(\mathcal{M}, P, M, m) \models \phi[C(\psi, \chi)]$ if and only if $(\mathcal{M}, P, M, m) \models \phi[\psi \wedge \chi] \wedge C(\psi, \chi)$.

Here by $\phi[\psi \wedge \chi]$ we denote the formula resulting from $\phi[C(\psi, \chi)]$ and replacing the subformula as indicated. Note that we define \vee, \rightarrow in terms of \wedge, \neg , so they are not covered by this lemma!

Proof By induction on the formula:

Base case: assume $\phi[C(\psi, \chi)] = C(\psi, \chi)$; then the claim is obvious.

Induction hypothesis: Assume $(\mathcal{M}, P, M, m) \models \phi[C(\psi, \chi)]$ iff $(\mathcal{M}, P, M, m) \models \phi[\psi \wedge \chi] \wedge C(\psi, \chi)$. Then we also have

1. $(\mathcal{M}, P, M, m) \models \phi[C(\psi, \chi)] \wedge \theta$ iff $(\mathcal{M}, P, M, m) \models (\phi[\psi \wedge \chi] \wedge \theta) \wedge C(\psi, \chi)$ (associativity and commutativity of \wedge),
2. $(\mathcal{M}, P, M, m) \models \theta \wedge (\phi[C(\psi, \chi)])$ iff $(\mathcal{M}, P, M, m) \models (\theta \wedge (\phi[\psi \wedge \chi]) \wedge C(\psi, \chi))$ (associativity of \wedge);
3. $(\mathcal{M}, P, M, m) \models C(\phi[C(\psi, \chi)], \theta)$ iff $(\mathcal{M}, P, M, m) \models C(\phi[\psi \wedge \chi], \theta) \wedge C(\psi, \chi)$, because we have $(\mathcal{M}, P, M, m) \models C(\phi[C(\psi, \chi)], \theta)$ iff $(\mathcal{M}, P, M, m) \models \phi[C(\psi, \chi)] \wedge \theta$ and $P(\sigma(\chi)) < P(\sigma(\chi)|\sigma(\phi[C(\psi, \chi)]))$ iff $(\mathcal{M}, P, M, m) \models (\phi[\psi \wedge \chi] \wedge \theta) \wedge C(\psi, \chi)$ and $P(\sigma(\chi)) < P(\sigma(\chi)|\sigma(\phi[\psi \wedge \chi]))$ (cf. 1 and definition of σ) iff $(\mathcal{M}, P, M, m) \models C(\phi[\psi \wedge \chi], \theta) \wedge C(\psi, \chi)$,
4. $(\mathcal{M}, P, M, m) \models C(\theta, \phi[C(\psi, \chi)])$ iff $(\mathcal{M}, P, M, m) \models C(\theta, \phi[\psi \wedge \chi] \wedge C(\psi, \chi))$ by the same argument.

In the proof of 3. we have left implicit the transition from \wedge to the metalogical “and”, but the proof should be clear nonetheless. \square

We can call this strong property **causal modularity**: as long as no negation is involved, causal meaning can be neatly separated from logical meaning. Note that a negation changes things dramatically: $\neg C(\phi, \chi)$ can be true in (\mathcal{M}, P, M, m)

both for logical reasons, ϕ or χ being false at m , or for probabilistic reasons, P not satisfying the desired inequation. Does the property of causal modularity have an analog in natural language? By way of example, consider the following:

- (4) a. Because Susan did not get the job because she is a woman, she went to court.
 b. People do not get a job because they are women, and because Susan is a woman and did not get the job, she went to court.

We can interpret a. as meaning exactly b., and our interpretation of causality would be exactly this one. But obviously, there are also readings in which a. is true, b. is false: there is a reading (i) where the probability of not getting a job for woman is not higher at all, and it was just the particular interviewer (on that particular morning etc.) having a particular favor for employing men. In that case, b. would be wrong, a. true. But note that the reading (i) is exactly the one which causes problems for our treatment even in a non-embedded way: whenever our causality comes from particular attitudes rather than general correlations, the probabilistic notion of causality fails. So the problem of reading (i) really points to a general problem of the probabilistic interpretation of causality, rather than to our our treatment of embedding causal statements: probabilistic notions necessarily address some general case rather than some particular instance. So to sum up, we find the properties of C so far appealing from a formal point of view and acceptable from an empirical point of view.

6 Expressive Power of \mathcal{L}

Now we come to the crucial question regarding \mathcal{L} , namely its expressive power. By the expressive power of an interpreted language we refer to its capacity to *axiomatize* certain classes of models. Let ϕ be an \mathcal{L} -formula. We say ϕ **axiomatizes** a class C of probabilistic models, if $C := \{(\mathcal{M}, P, M) : (\mathcal{M}, P, M) \models \phi\}$. In particular, we are interested in whether we can express the notions of conditional (in)dependence as they can be expressed in \mathcal{BL} , which turn out crucial for inferring the directionality of causation from the model.

As we said above, in our most simple reading, the connective C is symmetric (commutative), that is, $(\mathcal{M}, P, M, m) \models C(\phi, \chi)$ if and only if $(\mathcal{M}, P, M, m) \models C(\chi, \phi)$. Their logical equivalence is obvious (C is logically equivalent to \wedge); what might need a short word of explanation is their probabilistic equivalence. If $P(\sigma(\chi)|\sigma(\chi)) > P(\sigma(\chi))$, then $\frac{P(\sigma(\chi)\cap\sigma(\phi))}{P(\sigma(\chi))} > P(\sigma(\phi))$ iff $\frac{P(\sigma(\chi)\cap\sigma(\phi))}{P(\sigma(\phi))} \frac{P(\sigma(\phi))}{P(\sigma(\chi))} > P(\sigma(\phi))$ iff $\frac{P(\sigma(\chi)\cap\sigma(\phi))}{P(\sigma(\phi)P(\sigma(\chi)))} > 1$ iff $\frac{P(\sigma(\chi)\cap\sigma(\phi))}{P(\sigma(\phi))} > P(\sigma(\chi))$.

This equivalence might at first sound problematic, but in itself it is not: not knowing anything else, we do not know whether a causal statement is epistemic or physical; it is the more complex interaction of several variables/events which allows us to infer the direction of “proper causality”. So the question is: is the language \mathcal{L} under the given semantics actually rich enough to encode information about conditional dependence and independence?

The answer is: only partially. We are able to express unconditional dependence and independence of events, but not the conditional counterparts. The underlying reason for the latter is that \mathcal{L} does not even allow us to define or speak of conditional probabilities. First, we show the positive result. We say two formulas $\phi, \chi \in PL$ are *logically independent*, if none of $\phi \rightarrow \chi, \chi \rightarrow \phi, \phi \rightarrow \neg\chi, \neg\chi \rightarrow \phi$ is a (classical) tautology. Note that this implies that ϕ, χ cannot be (classical) tautologies themselves; moreover, ϕ, χ are logically independent, if and only if for any subset of $\{\phi, \chi\}$, there is a possible world containing exactly this subset of $\{\phi, \chi\}$.

Lemma 8 *Assume $\phi, \chi \in PL$ are independent. Then for the propositions $\sigma(\phi), \sigma(\chi)$, there are formulas ψ, ψ' such that*

$(\mathcal{M}, P, M) \models \psi$ iff and only if $P(\sigma(\phi \wedge \chi)) = P(\sigma(\phi))P(\sigma(\chi))$; and
 $(\mathcal{M}, P, M) \models \psi'$ iff and only if $P(\sigma(\phi \wedge \chi)) \neq P(\sigma(\phi))P(\sigma(\chi))$.

Proof We only construct ψ ; ψ' is then obtained as $\neg\psi$. Note that we must take care that ψ is logical tautology; the “purely probabilistic meaning” of ψ is quite easy to express. We put

$$\psi := \neg(C(\phi, \chi)) \wedge \neg(C(\phi, \neg\chi))$$

To see why this formula works, assume $(\mathcal{M}, P, M) \models \psi$, which means that for all $m \in M$, $(\mathcal{M}, P, M, m) \models \psi$. Now take a model (\mathcal{M}, P, M) , and assume $P(\sigma(\phi \wedge \chi)) \neq P(\sigma(\phi))P(\sigma(\chi))$; then either $P(\sigma(\phi \wedge \chi)) > P(\sigma(\phi))P(\sigma(\chi))$ or $P(\sigma(\phi \wedge \chi)) < P(\sigma(\phi))P(\sigma(\chi))$. Assume the former. As ϕ, χ are logically independent, there is $m \in M$ such that $\phi \in m, \chi \notin m$. But then $(\mathcal{M}, P, M, m) \models C(\phi, \chi)$, and so there is an m such that $(\mathcal{M}, P, M, m) \not\models \psi$, and thus $(\mathcal{M}, P, M) \not\models \psi$. The same holds for the other case and the second conjunct.

Conversely, assume $P(\sigma(\phi \wedge \chi)) = P(\sigma(\phi))P(\sigma(\chi))$. Then obviously, for any $m \in M$, $(\mathcal{M}, P, M, m) \not\models (C(\phi, \chi))$ and $(\mathcal{M}, P, M, m) \not\models (C(\phi, \neg\chi))$, and thus $(\mathcal{M}, P, M, m) \models \neg(C(\phi, \chi)) \wedge \neg(C(\phi, \neg\chi))$. \square

Reading the proof carefully, one discovers some subtleties of our definition of \models ; it is exactly the behavior of C under negation which departs from our expectations of classical logic. Note that the premise of ϕ, χ being independent was necessary to make the question meaningful. If $\phi \rightarrow \chi$ is a tautology (or any of the other), then $P(\chi|\phi) = 1$, and all questions boil down to logical ones.

So we have seen that \mathcal{L} can express the notion of dependence and independence of events. We now show that it *cannot* express the notions of *conditional* dependence or independence, because we cannot even express the concept of a conditional probability. By this result, we are also unable to express the concepts which are necessary in order to infer the directionality of (probabilistic) causation, as these make crucial reference to conditional probabilities.

Lemma 9 *Let (\mathcal{M}, P, M) be a probabilistic model. There is no general way, given two propositions $\sigma(\phi), \sigma(\chi)$, to construct an \mathcal{L} -formula ψ , such that $P(\sigma(\psi)) = P(\sigma(\phi)|\sigma(\chi))$.*

Proof We show this by counterexample, inspired by the fair dye. Put $m_1 \supseteq \{p_1, \neg p_2, \dots, \neg p_6\}, \dots, m_6 \supseteq \{\neg p_1, \dots, \neg p_5, p_6\}$, with $P(\{m_1\}) = \dots = P(\{m_6\}) = \frac{1}{6}$, and for all $O \subseteq M$, $P(O) = \sum_{1 \leq i \leq 6, m_i \in O} P(\{m_i\})$. Now, the probability of the event $\sigma(p_1)$ given the event $\bigcup_{i=1}^5 \sigma(p_i)$, both definable by \mathcal{L} -formulas p_1 and $\neg(\neg(p_1) \wedge \dots \wedge \neg(p_5))$, is obviously $\frac{1}{5}$. But now it is easy to show that there is no \mathcal{L} formula ψ , such that $P(\sigma(\psi)) = \frac{1}{3}$: every probability of an \mathcal{L} -formula in this space has the form $\frac{n}{6}$, where $0 \leq n \leq 6$, as its probability only depends on whether it is true in m_1, \dots, m_6 . \square

Given this result, one might ask the question: given that we can define the connective \rightarrow in our language, what is $P(\sigma(p \rightarrow q))$? That is not very intuitive in terms of probability. By definition, we get $P(\sigma(\neg p \vee q))$, and in the above example, with $q := p_1$, $p := \neg(\neg(p_1) \wedge \dots \wedge \neg(p_5))$, $\sigma(\neg p \vee q) = \sigma(p_6) \cup \sigma(p_2)$, and so we get $P(p \rightarrow q) = \frac{1}{3}$ in this space. So whereas \wedge, \vee, \neg behave as generally expected in probability theory, the \rightarrow does not.

This is the first flaw of \mathcal{L} ; there is yet another. Assume we want to express a conditional (in)dependence. There does not seem to be an alternative to the scheme: $C(\phi_1 \wedge \phi_3, \phi_2 \wedge \phi_3)$, where some ϕ_3 figures on both sides—otherwise, the dependence is unconditional. But that does bring us any closer to the desired result, because it claims that $P(\sigma(\phi_2 \wedge \phi_3)) < P(\sigma(\phi_2 \wedge \phi_3) | \sigma(\phi_1 \wedge \phi_3))$. But this is way to weak a requirement, because on the left side of the inequation, the ϕ_3 is unconditional, and we have to “pay” for it in terms of probability; whereas on the right hand side, it is also among the given information, for which reason we can just take it out: $P(\sigma(\phi_2 \wedge \phi_3) | \sigma(\phi_1 \wedge \phi_3)) = P(\sigma(\phi_2) | \sigma(\phi_1 \wedge \phi_3))$. So the difference is not only that ϕ_1 is given, but also ϕ_3 . Again, there does not seem to be a way to get around this.

Now of course, if we make the connective C ternary, then this can be solved easily; we can also solve both problems by introducing a new binary predicate, and even by giving C two different meanings depending on its context. But these are things we are reluctant to do: intuitively, the relation of causality is binary and has a single meaning. We try to solve the problem thus by giving another semantics to an isomorphic language. As we will see shortly, this is possible, but it comes with rather deep changes in the interpretation of the logical connectives.

7 The Language \mathcal{L}' and Its Semantics

\mathcal{L}' is syntactically isomorphic to \mathcal{L} ; we leave logical connectives unchanged, but skip C and write C' in its stead; nothing else changes. So the crucial change concerns the semantics of \mathcal{L}' . C had the logical meaning of \wedge , and in this sense it was “factual”. This will change for C' in \mathcal{L}' . However, as concerns the probabilistic meaning of C' , we want to stick to C , as this was our most simplistic intuition about probabilistic causality. What we rather do is the following: we change the logical interpretation of propositions and connectives in a way which allows us to conditionalize probabilities

wrt. to arbitrary definable subspaces. This way, we conditionalize the probability of a proposition wrt. the assignments which make the proposition true. This requires us to make a deep change: we do not consider truth at single worlds, but at *sets* of worlds. Therefore we have to lift everything to the powerset. Let τ be defined as follows:

1. $\tau(p) = \{m : p \in m\}$ for p atomic;
2. $\tau(\phi \wedge \chi) = \tau(\phi) \cap \tau(\chi)$;
3. $\tau(\neg\phi) = M - \tau(\phi)$;
4. $\tau(C'(\phi, \chi)) = M$.

We then define *prop* by $prop := \{O : O = \tau(\phi) \text{ for some } \phi \in \mathcal{L}'\}$. We now lift this to the powerset:

1. $\bar{\tau}(p) = \{\tau(p)\}$ for p atomic;
2. $\bar{\tau}(\phi \wedge \chi) = \{O \cap O' : O \in \bar{\tau}(\phi), O' \in \bar{\tau}(\chi)\}$;
3. $\bar{\tau}(\neg\phi) = \{M - O : O \in \bar{\tau}(\phi)\}$;
4. $\bar{\tau}(C'(\phi, \chi)) = \{O \in prop : P(\tau(\chi)|O) < P(\tau(\phi), O)\}$.

It is easy to see (by induction) that for $\phi \in PL$, $\bar{\tau}(\phi) = \{\tau(\phi)\}$. Now we can define truth of a formula in a model in a very simple fashion.

$$(\mathcal{M}, P, M, O) \blacktriangleright \phi, \text{ iff } O \in \bar{\tau}(\phi).$$

There are plenty of things to note:

1. Satisfaction is defined with respect to a set of worlds. This is needed to make definitions involving conditional probability meaningful (otherwise, we get a bunch of non-sense probabilities, as worlds are not even propositions).
2. C' does not make any factual presupposition. This is not a matter of choice, but rather becomes untenable: we cannot conditionalize on the events the conditional probability of which we want to define!
3. A note on the embedding of C' in other connectives. C' “downgrades” $\bar{\tau}$ to τ in its inductive definition; this is a technical necessity, but entails that “embedded” C' -clauses are practically meaningless. We might in principle be able to remedy this situation, but that would make definitions much more involved, so for now, we leave it like this.

The crucial advantage we draw from these definitions is the following: for any definable $O \subseteq M$, we can conditionalize probabilities on O . We now give the definition of validity: $(\mathcal{M}, P, M) \blacktriangleright \phi$ iff $(\mathcal{M}, P, M, M) \blacktriangleright \phi$. It is easy to check this preserves validity of the classical tautologies; moreover, we have $(\mathcal{M}, P, M) \blacktriangleright \phi \wedge \chi$ iff $(\mathcal{M}, P, M) \blacktriangleright \phi$ and $(\mathcal{M}, P, M) \blacktriangleright \chi$. From this we can also easily derive why C' cannot make a truth presupposition: assume $\phi, \chi \in PL$. Then it is impossible to affirm both the truth of ϕ, χ and $C'(\phi, \chi)$:

$$(\mathcal{M}, P, M, O) \blacktriangleright \phi \wedge \chi \wedge C'(\phi, \chi)$$

can never be true. This is because then we have $O \in \bar{\tau}(\phi \wedge \chi)$, and we require that $P(\tau(\chi)|O) < P(\tau(\chi)|\tau(\phi), O)$. However, $O \subseteq \tau(\phi)$, therefore $O \cap \tau(\phi) = O$, so $P(\tau(\chi)|O) = P(\tau(\chi)|\tau(\phi), O)$ —contradiction. This reflects the intuition (see Sect. 1.4) that probability and thus causality really affect different objects (“generic propositions”) than truth (“situated propositions”)—only that our ontology does not allow to make this distinction!⁶ So C' only covers the portion of *because* which we have called “*intensional*”.

8 Expressive Power of \mathcal{L}'

Having said this, we can now have a look at the expressive power of \mathcal{L}' in terms of probabilistic models which can be axiomatized. Obviously, we can do what was impossible with \mathcal{L} , namely: we can talk about conditional probabilities. How can we put this to use? For example, consider the following clause:

$$p_1 \rightarrow C'(p_2, p_3) \tag{4}$$

How does a model look like which satisfies this clause? Assume $(\mathcal{M}, P, M) \models p_1 \rightarrow C'(p_2, p_3)$. By convention, we have

$$\begin{aligned} (\mathcal{M}, P, M) \models p_1 \rightarrow C'(p_2, p_3) &\text{ iff} \\ (\mathcal{M}, P, M) \models \neg p_1 \vee C'(p_2, p_3) &\text{ iff} \\ (\mathcal{M}, P, M) \models \neg(p_1 \wedge \neg C'(p_2, p_3)). & \end{aligned}$$

Now we can disentangle truth-conditions:

$$\begin{aligned} (\mathcal{M}, P, M) \models \neg(p_1 \wedge \neg C'(p_2, p_3)) &\text{ iff} \\ M \in \bar{\tau}(\neg(p_1 \wedge \neg C'(p_2, p_3))) &\text{ iff} \\ \emptyset \in \bar{\tau}(p_1 \wedge \neg C'(p_2, p_3)) &\text{ iff} \\ \tau(p_1) \notin \bar{\tau}(\neg C'(p_2, p_3)) &\text{ iff} \\ \tau(p_1) \in \bar{\tau}(C'(p_2, p_3)) &\text{ iff} \\ P(\tau(p_3)|\tau(p_1)) < P(\tau(p_3)|\tau(p_2), \tau(p_1)). & \end{aligned}$$

We have taken some shortcuts, but all bi-implications can be easily verified. Consequently, $(\mathcal{M}, P, M) \models p_1 \rightarrow C'(p_2, p_3)$ if and only if $P(\tau(p_3)|\tau(p_2)) < P(\tau(p_3)|\tau(p_2), \tau(p_1))$. We thus have defined conditional dependence of variables. This is exactly what was not possible in \mathcal{L} , and what we need in order to express the conditional (in)dependencies we were able to express in the language of \mathcal{BL} , and which allowed us to infer the direction of causation. The reader should keep this argument in mind, as we will not repeat it in the proof of the next theorem.

⁶Recall that in this reading, “because A, B” has to be analyzed as: in general, A increases probability of B, and we have the fact (A at a certain time, place etc.) and (B at a certain time, place etc.). Our interpretation of C' only covers the former portion, not the latter.

Our main theorem will now be the following: for any finite collection of \mathcal{BC} -formulas, there is an \mathcal{L}' formula ϕ , such that for any probabilistic model, it is a model of the one if and only if it is a model of the other.

Theorem 10 *For each finite $\mathbf{B} \subseteq \mathcal{BC}$, there is a $\phi \in \mathcal{L}'$, such that for every probabilistic model (\mathcal{M}, P, M) , $(\mathcal{M}, P, M) \blacktriangleright \phi$ if and only if $(\mathcal{M}, P, M) \Vdash \mathbf{B}$.*

Proof Let \mathbf{B} be a finite collection of \mathcal{BC} -clauses. There are two types of these clauses, the ones constructed from \parallel and those with $\parallel\!\!\!|$. As in \mathcal{L}' , we have negation, there is no problem in getting from one to the other; and as we have conjunction, there is no problem in getting from single clauses to finite collections. So the problem boils down to only two kinds of \mathcal{BC} -clauses we have to show we can translate into \mathcal{L}' : firstly for statements of the type:

$$(\phi_i \parallel\!\!\!| \phi_j | \phi_l).$$

By definition of \mathcal{BC} , we know that $\phi_i, \phi_j, \phi_l \in PL$, and thus in \mathcal{L}' . We can quite easily translate this into an \mathcal{L}' -formula as follows:

$$(\phi_i \parallel\!\!\!| \phi_j | \phi_l)^+ := (\phi_l) \rightarrow (C'(\phi_i, \phi_j) \vee C'(\phi_i, \neg\phi_j))$$

We can verify that this has the desired meaning; we shorten the proof because we already showed large part of it by the above considerations. Note however that it is crucial that $\phi_i, \phi_j, \phi_l \in PL$, such that we have $\bar{\tau}(\phi_i) = \{\tau(\phi_i)\}$ etc. Now assume that $(\mathcal{M}, P, M, O) \blacktriangleright \phi_l$. Then we must have either $P(\tau(\phi_j)|O) < P(\tau(\phi_j)|\tau(\phi_i), O)$ or $P(\tau(\neg\phi_j)|O) < P(\tau(\neg\phi_j)|\tau(\phi_i), O)$. Now, by definition of \neg and probability axioms, it is clear that if $P(\tau(\phi_j)|O) = P(\tau(\phi_j)|\tau(\phi_i), O)$, then both are false, so the translation works as required. As negation in \mathcal{L}' is fully classical, this is also sufficient for $(\phi_i \parallel\!\!\!| \phi_j | \phi_l)$, which is equivalent to $\neg(\phi_i \parallel\!\!\!| \phi_j | \phi_l)$.

Regarding \mathcal{BC} -formulas of the form $(\phi_i \parallel\!\!\!| \phi_j)$, we simply translate them by

$$(\phi_i \parallel\!\!\!| \phi_j)^+ := C'(\phi_i, \phi_j) \vee C'(\phi_i, \neg\phi_j)$$

Finally, to finish the proof, we simply observe that \mathbf{B} is a finite set of clauses, and $(\mathcal{M}, P, M) \Vdash \mathbf{B}$ if $(\mathcal{M}, P, M) \Vdash \phi$ for all $\phi \in \mathbf{B}$. Let $[-]^+$ be the translation from \mathcal{BC} to \mathcal{L}' . Then, by definition of \wedge , $(\mathcal{M}, P, M) \Vdash \mathbf{B}$ if and only if $(\mathcal{M}, P, M) \blacktriangleright \bigwedge_{\phi \in \mathbf{B}} \phi^+$, which is a finite conjunction and therefore an \mathcal{L}' -sentence. \square

Note that the inverse of Theorem 10 is obviously false: \mathcal{L}' is much more expressive than \mathcal{BC} for several reasons: we cannot specify a set $\mathbf{B} \subseteq \mathcal{BC}$ such that $(\mathcal{M}, P, M) \Vdash \mathbf{B}$ if and only if $P(E_i|E_j) > P(E_i)$; we can only talk about independence. Theorem 10 is extremely important for the following reason: in the same fashion as \mathcal{BC} gives us proper information about the directionality of causation—under suitable side assumptions—, the same holds for \mathcal{L}' : given a sufficiently rich source of \mathcal{L}' clauses, we can infer in which direction causality goes.

Note however that the question of inferences is a topic of its own right, and can be conceived of in several fashions: in the first conception, we just have an \mathcal{L}' -clause ϕ ; it might be that ϕ , as some knowledge on the world, encodes the directionality of

causation of two events E_1, E_2 . That does of course *not* mean that we can immediately see or verify this: to “decode” directionality of causation from ϕ might be a difficult problem, on which we have said nothing so far, and we think this will require an investigation on its own. The more obvious conception is the following: assume we have a given model (\mathcal{M}, P, M) , where P can be effectively computed (or approximated). As we have seen above, given two events $E_1, E_2 \subseteq M$, there are two \mathcal{BL} -clauses ϕ, ψ , such that if $(\mathcal{M}, P, M) \models \{\phi, \psi\}$, then there is a directional causal link from E_1 to E_2 . By Theorem 10, this means: we can also just check whether $(\mathcal{M}, P, M) \triangleright \phi^+ \wedge \psi^+$, and thereby establish a causal link. The \triangleright relation, as long as P is computable, can easily be shown to be decidable.

So for \mathcal{L}' as for all our languages we have to keep in mind: when we say some formula (or set of formulas) *encodes* the directionality of causation, we mean: we can infer this information from the formula, but it *cannot* be expressed in the language itself (not even in the broadest sense, because this would require some sort of quantification over propositions)! This situation does not seem undesirable, when we compare to the situation to natural language: there is no *syntactic* difference between proper and epistemic usage of *because* (at least not across the board, though there are preferences in certain constellations). Nonetheless, “semantically”, we have clear intuitions about the two.

9 Conclusion

We have presented several “languages of causality”, namely \mathcal{BL} , \mathcal{L} and \mathcal{L}' . Statements in \mathcal{BL} can be read as denoting the (un)conditional (in)dependence of events; this language is well-known as a meta-language for Bayesian networks, and can be used to uniquely encode the directionality of a link in a minimal Bayesian network modelling the distribution under discussion. We have introduced and investigated the languages $\mathcal{L}, \mathcal{L}'$ and their semantics, because they are closer to our intuition about causation being a binary relation, which can in turn be embedded arbitrarily in other connectives. The main question was: how expressive in terms of axiomatization of probability spaces are these logical languages? \mathcal{BL} served for us as a touch-stone: if any of the languages $\mathcal{L}, \mathcal{L}'$ can encode \mathcal{BL} -statements, it can encode the directionality of causation. We have shown that \mathcal{L} is not sufficiently expressive, in particular, it cannot even express the notion of a conditional probability. On the other hand, \mathcal{L}' turns out to be powerful enough to encode arbitrary (finite sets of) \mathcal{BL} statements.

Though we have settled these questions, there are many questions left open, among which most prominently: given $\mathcal{L}, \mathcal{L}'$ and their semantics, is there a purely syntactic, proof-theoretic way to determine consequence and validity? And in particular, a way to determine whether a given set of formulas encodes a particular kind of dependency? That seems an interesting, difficult problem. The next question is: let \models, \triangleright denote consequence relations between (sets of) formulas in the usual fashion, that is: $\Gamma \models \phi$ if for all $\gamma \in \Gamma$, $(\mathcal{M}, P, M) \models \gamma$, then $(\mathcal{M}, P, M) \models \phi$. What properties do

these relation have? Do they behave like normal, logical consequence relations, being monotonous, finitary, closed under substitution etc.? This also seems far from obvious. These seem to us the most natural questions to address next.

References

- Dehghani, M., Iliev, R., & Kaufmann, S. (2012). Causal explanation and fact mutability in counterfactual reasoning. *Mind & Language*, 27(1), 55–85.
- Ebbinghaus, H.-D., & Flum, J. (1995). *Finite Model Theory*. Perspectives in Mathematical Logic. Springer.
- Geiger, D., & Pearl, J. (1990). Logical and algorithmic properties of independence and their application to Bayesian networks. *Annals of Mathematics and Artificial Intelligence*, 2, 165–178.
- Kaufmann, S. (2009). Conditionals right and left: Probabilities for the whole family. *Journal of Philosophical Logic*. 38(1), 1–53.
- Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. Adaptive computation and machine learning. MIT Press, Cambridge.
- Pearl, J., (2009). *Causality. Models, Reasoning, and Inference*. Cambridge University Press, Cambridge.
- Schurz, G., & Leitgeb, H. (2008). Finitistic and frequentistic approximation of probability measures with or without σ -additivity. *Studia Logica*, 89(2), 257–283.
- Verma, T., & Pearl, J. (1990). Equivalence and synthesis of causal models. In P. P. Bonissone, M. Henrion, L. N. Kanal, & J. F. Lemmer, (Eds.), *UAI*, (pp. 255–270). Elsevier.
- Wachter, B., Zhang, L., & Hermanns, H. (2007). Probabilistic model checking modulo theories. In *QEST*, (pp. 129–140). IEEE Computer Society.
- Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.
- Woodward, J. (2006). Sensitive and insensitive causation. *The Philosophical Review*, 115, 1–50.

Shannon Versus Chomsky: Brain Potentials and the Syntax-Semantics Distinction

Mathias Winther Madsen

Abstract The N400 and the P600 are two patterns of electrical brain potentials which can sometimes be found when people read or hear unexpected words. They have been widely claimed to be the neurological correlates of semantic and syntactic anomalies, respectively, but evidence accumulated over the last decade has raised some serious doubts about that interpretation. In this paper, I first review some of this evidence and then present an alternative way to think about the issue. My suggestion is built on Shannon's concept of noisy-channel decoding by tables of typical sets, and it is thus fundamentally statistical in nature. I show that a proper application of Shannon's concepts to the reading process provides an interesting reinterpretation of our notion of "syntax," thus questioning some fundamental assumptions of linguistics.

Keywords N400 · P600 · Electroencephalography · Psycholinguistics · Information theory · Probabilistic inference · Noisy-channel models · Syntax-semantics interface · Autonomy of syntax

One of the key ideas that helped establish linguistics as a respectable science in the 1950s was the idea that the processing of syntax is carried out by an autonomous mental module that works independently of the processing of meaning. Noam Chomsky famously pushed this point by presenting sentence pairs that seemed to dissociate violations of syntactic expectations from violations of semantic expectations (Chomsky 1957, pp. 15–16):

- The child seems sleeping (syntactic anomaly)
- Today I saw a fragile whale (semantic anomaly)

M.W. Madsen (✉)
Institute for Logic, Language and Computation,
University of Amsterdam, Amsterdam, The Netherlands
e-mail: m.w.madsen@uva.nl

© Springer International Publishing Switzerland 2015
H. Zeevat and H.-C. Schmitz (eds.), *Bayesian Natural Language Semantics and Pragmatics*, Language, Cognition, and Mind 2,
DOI 10.1007/978-3-319-17064-0_6

Based on such examples, he concluded that violations of syntactic rules had nothing to do with the meaning of the words in the sentence, and thus

... we are forced to conclude that grammar is autonomous and independent of meaning, and that probabilistic models give no particular insight into some of the basic problems of syntactic structure (Chomsky 1957, p. 17).

While the introspective case from example sentences was Chomsky's strongest argument at the time, he did imagine hypothetical psychological experiments that could be used to corroborate his claim, such as memory or reading tests that could perhaps distinguish "ungrammatical" sentences from mere nonsense (Chomsky 1957, p. 16).

This paper is about one attempt to realize such a research program. As I shall explain shortly, the brain sciences found some new results in the 1990s which seemed to finally put Chomsky's claims on a solid scientific basis. For a period of about ten years, the consensus was accordingly that the anatomical and psychological reality of Chomsky's little syntactic engine was a scientific fact. Syntax had acquired the authority of being "brain stuff."

In one sense, this paper is an elaborate criticism of this idea. I will review a wealth of empirical evidence from the last decade which seriously complicates the picture of the brain phenomena that were originally used to vindicate the Chomskyan view of language. Even a casual look at this literature exposes an abundance of unruly little details that fit very badly into the old pigeonholes.

But this is also a constructive paper: I propose a precise computational model that may explain much of the psychological evidence. This model is based on a seemingly insignificant idea about statistical error used by Claude Shannon in the 1948 paper which founded the field of information theory (Shannon 1948).

This idea has some deep connections to the foundations of statistical inference and thus to the notion of rational decision-making in the most general sense. Presenting my model will thus require a detour around some basic concepts from information theory and statistics. Once I have introduced those concepts, I will focus on a more specific example of a decoding problem similar to the experimental situations used in the psychological literature. I will then discuss how that toy problem might be scaled up to give accurate, quantitative predictions.

1 The N400 and the P600

In this section, I will introduce the brain phenomena that are pivot of the entire paper. These phenomena are a couple of distinctive changes in the electrical brain potential that can be measured off the scalp of a person's head during reading, listening, or other activities. Only in the next section will I discuss in more detail how we should interpret these modulations of the brain potential.

1.1 The N400

Suppose I present you with a sentence by flashing each word on a computer screen at regular intervals:

- I ... spread ... the ... warm ... toast ... with ...

While you are reading the sentence, I am recording the electrical activity at different places on the top of your head. Since all nerves in your body change the electrical field around them when they are excited, this recording provides an estimate of how active the nerves inside your head are. By using the electrical activity in the head as a measure of cognitive effort, this experimental paradigm can thus reveal whether a sentence requires more effort than usual to read, and where in the sentence the problems begin.

This means that rather specific differences between sentences can be compared. I could, for instance, change a single word in a sentence and then compare the pattern of electrical activity that arises when you read the original sentence with the pattern associated with the manipulated one:

- I spread the warm toast with butter (original)
- I spread the warm toast with socks (manipulation)

This manipulation was the idea behind an experiment which was first carried out by Martha Kutas and Steven A. Hillyard in 1980, and which has since then been replicated many times. Their experiment involved a set of sentence pairs like the ones above: One completely unproblematic sentence and one that ended in a “semantically inappropriate (but syntactically correct)” word (Kutas and Hillyard 1980, p. 203).

By averaging over several subjects’ responses to the two versions of the sentences, Kutas and Hillyard found that the oddness of the manipulated sentences tended to evoke a characteristic electrical response characterized by an excess of negative electrical potential relative to the control condition (Fig. 1).

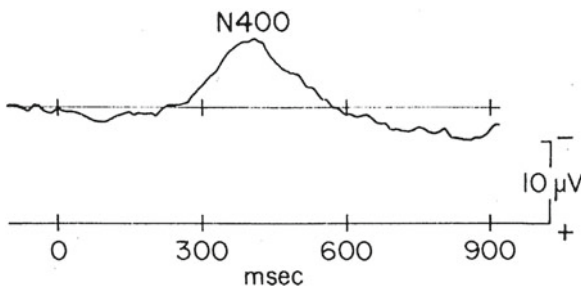


Fig. 1 A graph from Kutas and Hillyard (1980, p. 204) showing the averaged response to the inappropriate word (e.g., *socks*) minus the averaged response to the appropriate (e.g., *butter*). Both waveforms were recorded by an electrode at the back of the head (the medial parietal position). As is conventional in electrical engineering, negative charge is plotted as up

Because this bump on the waveform showed up around 400 ms after the unexpected word was presented, they called the effect the “N400.” They suggested that it might reflect a kind of “second look,” like a cognitive equivalent of rubbing your eyes in disbelief.

1.2 *The P600*

Kutas and Hillyard’s experiment was explicitly designed to manipulate “semantic congruity” while keeping syntactic factors constant. This naturally led to the question of whether syntactic manipulations would lead to different results than semantic manipulations.

This question was answered in 1992 by Lee Osterhout and Phillip J. Holcomb, who compared a different kind of sentence pairs in an otherwise virtually identical experimental paradigm:

- The swimmer decided to lose weight (control)
- The swimmer urged to lose weight (manipulation)

Although the difference in wording between these two sentences lies in the verb (*decided* versus *urged*), the difference in how odd they appear only later: The sentence fragment *The swimmer urged ...* can be completed in many perfectly fine ways, but the extension *The swimmer urged to ...* cannot. It is thus only when the word *to* is presented that it gets difficult to come up with a plausible structural interpretation of the sentence.

Consistent with this reasoning, Osterhout and Holcomb found a marked difference between the reaction to the word *to* in the two conditions, with the manipulated sentence provoking an increased positive potential. They found that this positive effect peaked at about 600 ms after the onset of the word and logically named it “P600.”

Having thus apparently found a neurological correlate of the mismatch between verb type and complement type, Osterhout and Holcomb speculated that “the P600 and N400 effects are elicited as a function of anomaly type (syntactic and semantic, respectively),” and that “the P600 co-occurs with syntactic anomaly” (Osterhout and Holcomb 1993, pp. 785 and 798).

Their results thus seemed to support the idea that syntax and semantics are distinct features of language—not only in the mind of the linguist, but also in the brain of the language user.

2 **The Frayed Ends of Syntax**

With the discovery of the P600, the brain sciences produced a strong neurological argument in favor of the Chomskyan view of language. Linguists could now with confidence postulate the “autonomy of syntax” and cite the authority of brain

science as support (see e.g. Fernández and Cairns 2010, Chap. 3). The issues raised by Chomsky from a purely introspective perspective finally seemed to be settled by empirical investigations.

The issue is, however, not quite as simple as this. Within the last decade, a number of experiments have documented P600 responses to sentences that are not syntactically anomalous in any usual sense of the word “syntax.” I will spend the remainder of this section illustrating these complications in some detail, drawing quite heavily on an excellent review paper by Gina Kuperberg (2007). For a more exhaustive review of the wide variety of contexts in which the N400 is found, including non-linguistic contexts, see Kutas and Federmeier (2011).

2.1 Word Shuffling and P600 Effects

One of the first indications that the P600 could not unproblematically be equated with broken syntax came from a Dutch study which manipulated the test sentences by changing a passive construction into an active, causing the logical subject and the logical object swap roles (Hoeks et al. 2004).

In Dutch, the main content verb is often placed at the very end of the sentence, as in *The boy has the ball kicked*. If you read such a verb-final sentence in a word-for-word presentation, a whole string of nouns and auxiliary verbs can thus be presented to you before you finally reach the main content verb (here *kicked*). This means that one can construct a Dutch sentence such that it creates a very strong expectation as to what the final verb will be, and those expectations can then be violated or respected.

This grammatical fact was exploited in the materials used by Hoeks et al. They constructed three different manipulated sentences from a control by substituting either the main content verb, an auxiliary verb, or both:

- De eed werd door de jonge artsen afgelegd.
 (“The oath was taken by the young doctors.”)
- De eed werd door de jonge artsen behouden.
 (“The oath was kept/retained by the young doctors.”)
- De eed heeft de jonge artsen afgelegd.
 (“The oath has taken the young doctors.”)
- De eed heeft de jonge artsen behouden.
 (“The oath has kept/retained the young doctors.”)

The manipulations of the main verb (e.g., substituting *retain an oath* for *take an oath*) tended to produce mere nonsense, and as expected, this manipulation produced an N400 effect. However, the manipulation of the auxiliary verb—which caused the subject and object to change roles—unexpectedly triggered an excess of late positive potential beginning around 600 ms after the final verb (cf. Fig. 2).

Although both of these manipulations resulted in a grammatical sentence, only one of them would produce an N400, and the other would instead produce a very

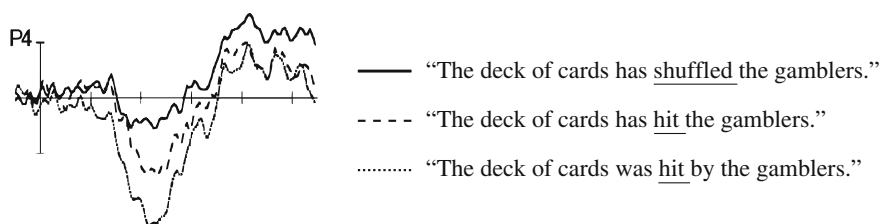


Fig. 2 Averaged single-electrode recordings from Hoeks et al. (2004, p. 68), with positive charge plotted as up. The three waves show the differences in voltage between the three manipulated conditions and the control condition (“The deck of cards was shuffled by the gamblers”). All three waveforms exhibited a significant P600 effect at this electrode

pronounced P600. This ran directly counter to the notion that the P600 exclusively tracks syntactic problems.

Similar effects were reported for English. In one 2003 experiment, Kuperberg et al. compared two different kinds of unexpected verbs and again found that only some of them provoked an N400 response:

- For breakfast, the boys would only eat ... (control)
- For breakfast, the boys would only bury ... (N400)
- For breakfast, the eggs would only eat ... (P600)

Kuperberg et al. suggested that this difference might be explained in terms of the thematic roles prescribed by a verb like *eat*, assuming that “the P600 is sensitive to violations in this thematic structure” (Kuperberg et al. 2003, p. 127).

This line of thought was also supported by another study which compared mere nonsense to bizarre subject-verb combinations (Kim and Osterhout 2005). This study, too, found marked differences in the kinds of response elicited by the two manipulations relative to a control condition of ordinary English:

- The library books had been borrowed by the graduate student. (control)
- The tragic mistake was borrowing in a hurry. (N400)
- The library books had been borrowing the graduate student. (P600)

Like Kuperberg et al., Kim and Osterhout suspected that the difference between the two target sentences had something to do with how nouns plug into verbs. They formulated this intuition in terms of a “semantic attraction to particular predicate–argument combinations” (Kim and Osterhout 2005, p. 215).

2.2 Animacy and Related Criteria

Looking at these examples, one might come to expect that the crucial issue here had something to do with animacy: *boys* are animate and therefore appropriate subjects for the verb *eat*, but *eggs* are not. However, as several authors have argued,

this distinction does not quite capture the differences in the empirical material (Kuperberg et al. 2003; Kuperberg 2007).

For instance, *library books* and *tragic mistakes* are both inanimate noun phrases, and both are inappropriate as subjects for the verb *borrow*. Still, the construction *the library books borrowed* elicits a P600, while *the tragic mistakes borrowed* elicits an N400, as mentioned above. Since these constructions are equivalent in terms of animacy violations, this dimension alone is thus an unreliable predictor of P600 effects.

It should also be kept in mind that many earlier experiments had used subject-verb combinations that were clearly illegitimate in terms of animacy and other selection criteria, but still produced N400 effects instead of P600 effects. One German study, for instance, used the following target sentences:

- Der Honig wurde ermordet.
 (“The honey was murdered.”)
- Der Ball hat geträumt.
 (“The ball has dreamt.”)

Both of these sentence types provoked a marked N400 response, but no P600 (Rösler et al. 1993). It would thus appear that inanimate objects can be used with verbs that require animate subjects without provoking a P600 (although this will probably produce an N400).

As for the opposite implication—P600 effects predicting animacy violations—other research groups have also documented P600 responses to sentences in which animate subjects do clearly animate things. One study by Kolk et al. (2003), for instance, found this in Dutch sentences like

- De vos die op de stropers joeg ...
 (“The fox that was hunting the poachers ...”)
- De kat die voor de muizen vluchtte ...
 (“The cat that was fleeing the mice ...”)

Of course, the roles are in some sense swapped in this sentence—but since both foxes and poachers are animate, the strangeness of these sentences cannot be described in terms of animacy alone.

In another experiment applying essentially the same materials, van Herten et al. (2005, p. 249) to their own surprise, also found P600 effects for sentences such as

- De bomen die in het park speelden ...
 (“The trees that played in the park ...”)

In a later paper, they hypothesized that this P600 effect was caused by the strong semantic relation between the words (one can *play* in a *park* full of *trees*) in combination with the oddness of the sentence as a whole (the *trees played*). One way of describing this pattern could be that it involves nouns and verbs that tend to occur in the same contexts, but rarely in subject-predicate combinations (van Herten et al. 2006).

This theory would explain differences such as the following:

- Jan zag dat de olifanten de bomen snoeiden en ...
 (“John saw that the elephants pruned the trees and ...”—P600)
- Jan zag dat de olifanten de bomen verwenden en ...
 (“John saw that the elephants spoiled the trees and ...”—N400)

Trees, elephants, and pruning can easily occur in the same text or even the same sentence, but it is rarely the elephants that are doing the pruning. On the other hand, *spoiling* (as in *spoiling a child*) is semantically unrelated to elephants and trees, and the word tends to occur in very different contexts.

Whether this exact formulation of the theory is correct or not, there are thus quite strong reasons to doubt that the conditions producing the P600 is a matter of plugging an inanimate noun into a verb frame that requires an animate argument, or *vice versa*. A wide variety of word relationships seem to be responsible for turning “semantic” anomalies into potentially “syntactic” ones.

2.3 *Grammatically Ungrammatical*

Given all of this data, it seems that there is no clear and straightforward relationship between the P600 and the arguments that can be plugged into a verb. The issue does not seem to be syntactic well-formedness, nor does it seem to be about animacy or thematic structure. It might still be the case, however, that some other grammatical parameter—say, aspect, mood, gender, number—is the key. But even this vague suggestion has some empirical evidence against it.

One such piece of evidence comes from a study that recycled some sentences which had already previously been shown to produce strong and reliable N400 effects. In this experiment, an elaborate context was prepended to the test sentences, and the brain potentials were recorded in the same manner as in the previous study (Nieuwland and Van Berkum 2005).

One of stories used in this experiment read as follows, in translation from the original Dutch:

- A tourist wanted to bring his huge suitcase onto the airplane. However, because the suitcase was so heavy, the woman behind the check-in counter decided to charge the tourist extra. In response, the tourist opened his suitcase and threw some stuff out. So now, the suitcase of the resourceful tourist weighed less than the maximum twenty kilos. Next, the women told the suitcase that she thought he looked really trendy. ...

Surprisingly, this long build-up to the crucial unexpected word completely canceled the N400 effect found in the original study and instead produced a strong P600 effect. Thus, the manipulation of the context which turned the suitcase into a prop in the narrative dramatically changed the way the subjects read the nonsense sentence *the woman told the suitcase*.

But perhaps the deepest problem with the idea that the P600 can be described in grammatical terms is that it can be provoked by sentences without any syntactic anomalies at all. This was already shown in a study from 1999 (Weckerly and Kutas 1999) which compared sentences of the following form:

- The novelist that the movie inspired praised the director for ...
- The editor that the poetry depressed recognized the publisher of the ...

Strictly speaking, there is nothing semantically or syntactically anomalous about movies that inspire novelists, or about poetry that depresses editors. Still, these constructions tended to produce a P600 response.

Weckerly and Kutas used these findings in order to make a point about “the use of animacy information in the processing of a very difficult syntactic structure” (Weckerly and Kutas 1999, p. 569). Their findings may also have more specific causes, though, such as the strong, conventional expectations carried by word clusters like *novelist*, *movie*, and *inspire*. However, since their experimental materials were not reprinted in the paper, it is hard to assess how strong the case for such an explanation might be.

However, regardless of these points of interpretation, it seems clear that there are some serious problems with seeing the P600 as the exhaust pipe of a syntactical processor in the brain: There are several kinds of semantic oddness that are quite strongly related to the P600; it can be quite heavily modulated by discursive context; and it can even be elicited by grammatical sentences. The standard tools from the toolbox of theoretical linguistics thus seem to have some problems getting to grips with this phenomenon.

I take this as a sign that we need to back up a bit and take a fresh look at our assumptions. In the next section, I will consequently suggest that in order to understand the brain processes behind the N400 and the P600, we need to go all the way back to Chomsky, and then go a little further back.

3 The Statistics of Decoding: Errors of Type I and II

In the preceding section, I have presented a number of empirical phenomena that seriously challenges the traditional identification of the N400 and the P600 with semantic and syntactic anomaly, respectively. As the examples have shown, the P600 in particular crops up in a number of different circumstances that cannot be equated with broken syntax without bending our notion of syntax completely out of shape. This raises two questions.

1. We might wonder whether there is any system at all to where the P600 is present. Indeed, after looking at the wide variety of examples in the previous section, one might get the impression that brain potentials just aren't the kind of phenomena that can be adequately predicted. It is an open question whether we can even describe the phenomenon in intuitive terms, and whether we can look at a sentence and guess what kind of brain response it will evoke.

2. Assuming that there is some kind of system, the question is what formalism would be most suited to articulate it. It is tempting to pull down the grammatical vocabulary from the shelf, since we are after all talking about language; but it is not a given that a theory of the N400 and the P600 should come in such a form, or that it should involve any concepts from Chomskyan linguistics.

In this section, I will suggest that there is indeed a way of understanding the two brain components of the reading-related brain potentials, but that they have little to do with language as such. Instead, I will draw on some classical statistical insights from information theory (Shannon 1948) and approach the activity of reading as a kind of decoding problem. The two brain responses will then show up as correlates of two particular kinds of decoding error.

This suggestion is consistent with an intuition shared by many researchers in the field, namely, that the P600 is some kind of “error-correction signal” (Gibson et al. 2013, p. 8055) or the electrical counterpart of a “monitoring process triggered by a conflict between two or more incompatible representations” (van Herten et al. 2006, p. 1195; see also Frenzel et al. 2011).

Similarly, Kolk et al. (2003) note that both the N400 and the P600 are triggered by unexpected events, and they add:

The problem with such events is that they can have two sources. They can be real, in the sense that an unexpected event has indeed occurred. On the other hand, they can also stem from a processing error (Kolk et al. 2003, p. 31).

To paraphrase this in information-theoretic terms, strange things can either happen in the source or in the channel. I agree completely with this intuition, and the purpose of this section is to spell it out in some more mathematical detail. The model that I will present is very similar to the one discussed by Gibson et al. (2013, p. 8055), both in its general philosophy and in the mathematical detail. I will, however, be a bit more explicit in my claims about the N400 and the P600 (which they only mention in passing), in particular when it comes to identifying their exact computational-level counterparts.

I will claim that the two electrical responses correspond to two different kinds of decoding error: The P600 is present when there are too many candidate causes that match an observed effect, and the N400 is present when there are none. This hypothesis draws on an idea that Shannon used in the proof of his discrete channel coding theorem (Shannon 1948, Theorem 11) and which was later made more explicit by others (Cover 1975; Cover and Thomas 1991, Chap. 7.7). Shannon operationalized the notion of a “match” between a cause and an effect in terms of a concept called the “typical set.” This set contains, as I shall explain in more detail below, the events that are about as probable as one would expect an event to be (Shannon 1948, Theorem 3).

Before I can formulate my hypothesis about the difference between the N400 and the P600, I will first have to discuss this concept and a few other basic notions from information theory. I will then illustrate the idea behind my theory by showing how it works in a computationally simple case, and then hint at what it would look like in a more realistic model.

3.1 Typical Sets and Most Probable Sets

One of building blocks of information theory is the concept of a “typical set” (Shannon 1948, Theorem 3). The typical set associated with an experiment is the set of outcomes whose logarithmic probabilities are close to the expected logarithmic probability.

To put this differently, suppose we identify the experiment in question with a random variable X and that we let $p(x)$ be the probability of the event $X = x$. As suggested by Samson (1951), we can then think of the number $-\log p(x)$ as a quantification of how “surprising” the observation $X = x$ is. This “surprisal” is a number between 0 and ∞ , with less probable events having a higher surprisal values (cf. Fig. 3).

A random variable X which can take different values with different probabilities thus always corresponds implicitly to a “surprisal” variable $-\log p(X)$ which ranges over the set of surprisal values. The expected value of this surprisal variable is $H = H(X)$, the entropy of X .

Using these concepts, we can then reformulate the definition of the typical set: It is the set of events $X = x$ whose surprisal value is close to the average surprisal. For a fixed $\varepsilon > 0$, the ε -typical set of a random variable X thus contains the x for which

$$\left| \log \frac{1}{p(x)} - H \right| \leq \varepsilon.$$

The ε -typical set associated with an experiment does not necessarily include the most probable outcome of the experiment. However, as mentioned by Shannon (1948, Theorem 3) and explained in more detail by Cover and Thomas (1991, Chap. 3.3), the set of outcomes that are less surprising than H usually differs very little from the typical set, since any two high-probability sets must have a high-probability overlap. Specifically, if the values of the random variable X are long sequences from an

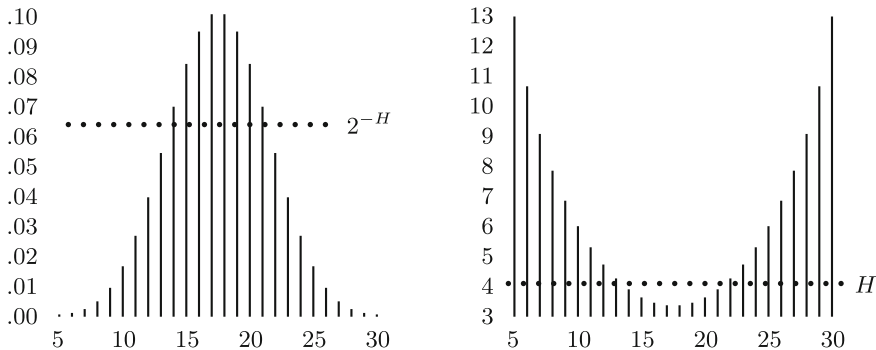


Fig. 3 Left The probability of getting a sum of $\Sigma = 5, 6, 7, \dots, 30$ in 5 dice throws, with the typical probability 2^{-H} shown as a dotted line. Right The same, plotted on an inverse logarithmic scale

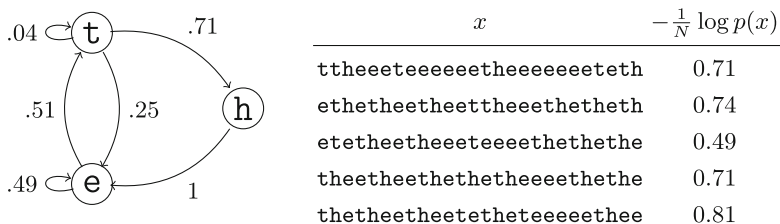


Fig. 4 Sequences of length $N = 25$ from an ergodic source. The entropy rate of the source is $H = 0.80$ bits per symbol

ergodic random process (such as strings of letters or words; cf. Fig. 4), then including the most probable outcomes will only change the size and total probability of the set slightly. In many important respects, it thus makes little difference whether we define the typical set by the bound $-\log p(x) \leq H + \epsilon$ or the symmetric condition $H - \epsilon \leq -\log p(x) \leq H + \epsilon$.

3.2 Decoding by Jointly Typical Sets

The conditions discussed above define the typical set for a single random variable. In the context of information transmission, however, we are rarely interested in a single, isolated variable, but much more often in the relationship between two variables X and Y that model a cause and an effect, or a transmitted message and a received signal.

In order to study such relationships, it is useful to consider the typical sets associated with the joint stochastic variable $X \times Y$. This set consists of pairs (x, y) that are typical with respect to the joint probability distribution on $X \times Y$, and it is called the jointly typical set.

In an information transmission context, a jointly typical pair (x, y) consists of a high-probability message x and a noisy signal y which has a high probability given the message. The set of all such pairs collectively describe the properties of the communication channel modeled by $X \times Y$. A model of written English, for instance, might include *(forty, fourty)* as a typical pair, since *forty* is a common word, and *fourty* a common misspelling of it. Other channels will have other characteristic types of confusability (cf. Fig. 5).

This way of looking at the issue suggests that a table of the typical pairs could be used for decoding: When you receive a message y , you just skim through the table, find a pair of the form (x, y) , and then return x as your decoded message. Although this hypothetical algorithm is usually not feasible in its naive form, it captures many of the conceptual ideas about channel coding, and its complexity problems are largely irrelevant to the point I want to make here.

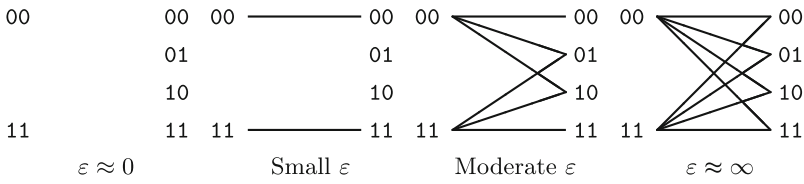


Fig. 5 One possible structure of a jointly typical set for increasing tolerance thresholds and thus decreasing lower bounds on the joint probability $p(x, y)$

3.3 Two Types of Decoding Error

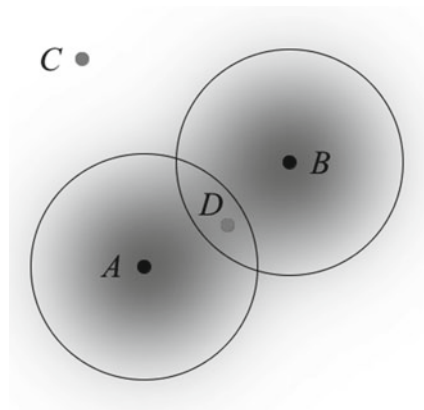
As is apparent from the description of the hypothetical brute-force algorithm for decoding by jointly typical sets, two distinct problems can arise for a decoder:

1. There is no x for which (x, y) is typical
2. There is more than one x for which (x, y) is typical

As an illustration of these two types of error, suppose that the message is one of two points in the plane, A or B , and that the received signal is equal to the message plus some Gaussian noise (cf. Fig. 6). The ϵ -typical set for this communication channel will then contain the input-output pair (A, C) for every C on a disc centered on A , and the pair (B, C) for every C on a disc centered on B . The sizes of the discs are determined by ϵ , the variance of the noise, and the prior probabilities of A and B .

In a situation like this, the two types of error correspond to different regions of the plane: Errors of the first type correspond to everything outside the two discs; errors of the second type correspond to their overlapping parts. Outside the discs, both of the events $X = A$ and $X = B$ have extremely low posterior probability, even if one of them is more probable than the other. In the overlap between the two discs, on the other hand, the two explanations have roughly equal posterior probability.

Fig. 6 A inference situation with two possible causes (A and B), a type I error (C), and a type II error (D)



We could reduce the risk of an error of the first kind by choosing a bigger ε (i.e., making the discs larger), but this would also expand the overlap between the discs, and that would in turn increase the probability of the other kind of error. There is thus a trade-off between accuracy and coverage.

While it is important to notice the similarities between this picture and the frequentist notion of type I and type II errors (Neyman and Pearson 1928), there are also a couple of differences that are worth keeping in mind. Shannon's analysis recommends that we accept a cause x as an explanation for the observed effect y when

$$p(x | y) p(x) \geq 2^{-H(X \times Y) - \varepsilon}.$$

On this analysis, the region of acceptance depends both on the prior probabilities of the inputs, $p(x)$, and the noisiness of the channel, $H(X \times Y)$. The Neyman-Pearson analysis, by contrast, disregards both of these statistics. Instead, it uses a fixed significance level α and accepts x as an explanation of y when

$$p(y | x) \geq 1 - \alpha.$$

The difference between these two decision rules reflects a difference in philosophy. The purpose of Shannon's analysis is ultimately to assign one cause to each effect; the purpose of the Neyman-Pearson analysis is to delineate the set of causes that can, in isolation, safely be regarded as implausible explanations of the data. The two methods thus answer different questions.

4 An Application to Locally Noisy Strings

While the previous section considered some rather abstract statistical concepts, I would now like to turn to a more concrete example which is highly relevant to the general topic of this paper. Nominally, the example is a statistical model of misspellings, but it applies to any kind of communication channel that can distort a sequence through local changes, such as reversing the order of two adjacent symbols.

Such models have been studied extensively in computer science (Levenshtein 1966; Damerau 1964), and many textbooks contain discussions of related ideas. The theory that underlies such models has roots in classical ideas from computer science, most notably, dynamic programming (Bellman 1952; Viterbi 1967).

4.1 A Generative Story for Misspellings

There are many complicated reasons why people misspell words, but in order to get a better mathematical grip on the situation, I will present a very schematic model of the process here. The model takes the form of a cartoonish generative story which

depicts people as a certain kind of machines that experience random glitches during the writing process. These glitches cause them to incorrectly externalize the contents of their internal memory, and someone reading their output thus needs to reconstruct their original intentions based on the corrupted output signal.

More specifically, let us assume that a writer chooses a word x with a probability close to the frequency of the word in written English. Next, the writer consumes the word, letter for letter, in a left-to-right manner (cf. Fig. 7). While traversing the input string in this way, the writer also writes an output string y which more or less faithfully reflects the input. How exactly the writer traverses the word x and how he or she chooses which letters to write is decided stochastically as summarized in Table 1. Note that this channel model explicitly includes local permutations such as *percieve* instead of *perceive* (here modeled as an operation called “Reverse”).

As the table shows, the model contains several hyperparameters that be can set to arbitrary values, depending on what we think best reflects actual behavior. In the examples discussed below, I have used the values $\alpha = 0.96$ and $\beta = \gamma = \delta = \eta = 0.01$, meaning that an average of 24 out of 25 letters are expected to be faithfully reproduced.

Thanks to classic computer science trick, the transmission likelihoods $p(y | x)$ for this channel can be computed quite easily. The way a writer consumes a string x and outputs a string y can be identified with a path through a matrix. Each step of the

Fig. 7 Consuming the word *stirs* by walking through positions 0, 1, 2, 4, 4, and 5

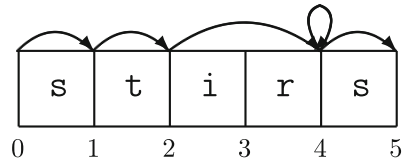
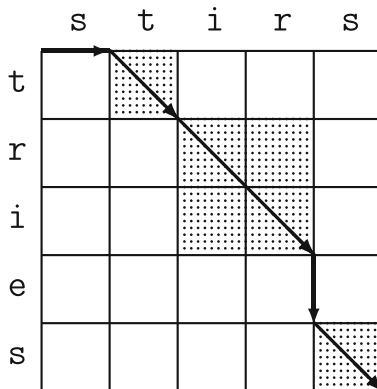


Table 1 Generative model for a spelling channel

State	Action	Probability	Output	Next state
Start	Choose x	$\Pr(x)$	–	0
$i < x $	Echo	α	x_i	$i + 1$
$i < x $	Change	$\beta/25$	$c \neq x_i$	$i + 1$
$i < x $	Insert	$\gamma/26$	c	i
$i < x $	Delete	δ	–	$i + 1$
$i < x $	Reverse	η	$x_{i+1}x_i$	$i + 2$
$i = x $	Echo	α	a_i	$i + 1$
$i = x $	Change	$\beta/25Z$	$c \neq x_i$	$i + 1$
$i = x $	Insert	$\gamma/26Z$	c	i
$i = x $	Delete	δ	–	$i + 1$
$i = x + 1$	Insert	$\gamma/26$	c	i
$i = x + 1$	Halt	$1 - \gamma$	–	–

The rows labeled “Change” and “Insert” should be read as abbreviated forms of several distinct entries, one per possible output letter. Z is the normalizing constant $(\beta + \gamma)/(\beta + \gamma + \eta)$

Fig. 8 A path through (*stirs*, *tries*) containing a deletion, an echo, a reversal, a spurious insertion, and another echo



path corresponds to a particular editing operation like deletion or reversal (cf. Fig. 8). The problem of finding the likelihood $p(y | x)$ is thus equivalent to the problem of summing up the probabilities of all paths through this matrix. The probability of any specific path can be found by multiplying the probability of the individual editing operations, since the operations are independent.

The crucial fact that allows us to sum up the probability of these paths is that the transmission errors only have local effects. This means that the only matrices for which we need to count all the possible paths are the ones of size 2×2 . For all larger sizes (e.g., 3×4), we can ease the computation by using the partial results we have already computed (e.g., the sums for the matrices of size 3×3). Thus, by working upwards from size 2×2 and keeping careful score of our partial results, we can add up the probability of every path through the matrix without actually looking at each one individually.

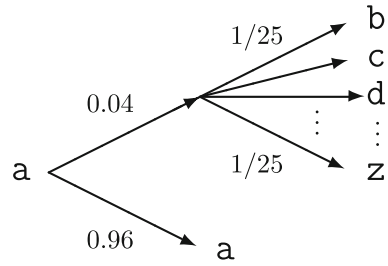
This style of a bottom-up recursion is known as dynamic programming (Bellman 1952). I will tacitly use it in all the examples discussed below.

4.2 Estimating Joint Surprisal Values

Analyzing the joint entropy of the channel $X \times Y$ is not completely trivial for the model used here, and the details are not terribly important for my present purposes. It will, however, be very illustrative to see a few representative examples, so I will now provide a rough back-of-the-envelope calculation of some descriptive statistics associated with the spelling channel.

Let us first assume that an input word x is given and then have a look at what our uncertainty about the output string y is. We can think roughly of the consumption of a single input letter as a process requiring two choices to be made: Deciding whether to make a mistake, and if so, deciding which mistake to make (Fig. 9). The first choice involves two options with probabilities 0.96 and 0.04, respectively, so it contains

Fig. 9 The simplified channel model



$$0.96 \log \frac{1}{0.96} + 0.04 \log \frac{1}{0.04} = 0.17 \text{ bits of uncertainty.}$$

The second choice is a random selection from a set of 25 letters, so it amounts to $\log_2 25 = 4.64$ bits of uncertainty. However, since this choice only comes up in the four percent of the cases, the grand total is

$$H(Y|X) = 0.17 + 0.04 \cdot 4.64 = 0.36 \text{ bits of uncertainty per letter.}$$

An input word with N letters is thus associated with about $2^{0.36N}$ typical output strings, according to this simplified calculation. This accounts for the channel entropy $H(Y | X)$. The source entropy $H(X)$, on the other hand, only involves choosing a word from the dictionary. Using the frequencies in the Brown corpus (Francis and Kucera 1967) as estimates of the word probabilities, we arrive at an entropy of about $H(X) = 10.54$ bits. Words like *know*, *while*, *last*, *us*, *might*, *great*, and *old* have surprisal values close to this average. The most frequent word, *the*, has a surprisal value of 3.84.

To show more concretely what these numbers mean, consider the word *great*. This word consists of five letters, and as we have seen, each letter contributes about 0.36 bits of surprisal. We can thus expect an average conditional surprisal about the output, $H(Y | X = \textit{great})$, in the ballpark of

$$5 \cdot 0.36 = 1.80 \text{ bits.}$$

In order to find the prior surprisal associated with the input, we can look up the probability of the word *great*. This turns out to be about $6.6 \cdot 10^{-4}$, corresponding to a surprisal value of 10.56 bits. Adding this prior surprisal to the average conditional surprisal, we find an average joint surprisal for pairs of the form (\textit{great}, y) . On average, this should be about

$$10.56 + 1.80 = 12.36 \text{ bits.}$$

In fact, we can pick apart this average even further: Suppose for instance that the transmission does not introduce any errors, so that the input-output pair is $(x, y) = (\textit{great}, \textit{great})$. Then the joint surprisal is

Table 2 Examples of joint surprisals

x	y	$-\log p(x, y)$
<i>great</i>	<i>great</i>	10.84
<i>great</i>	<i>graet</i>	17.32
<i>great</i>	<i>grate</i>	24.00
<i>great</i>	<i>grxqz</i>	30.42

$$10.56 - \log 0.96^5 = 10.85 \text{ bits,}$$

or in fact slightly less, since the output *great* can either be produced by a faithful reproduction of x (with probability 0.96^5), or an unlikely combination of errors that cancel each other out. If a single error is introduced, on the other hand, the joint surprisal for the pair (x, y) will be about

$$10.56 - \log 0.96^4 - \log 0.01 = 17.44 \text{ bits,}$$

depending a bit on what the error is. Table 2 gives some direct computations that corroborate these estimates.

4.3 Letter-for-Letter Surprisal Values for Competing Hypotheses

All of the preceding discussion assume that we selected an entire word from our dictionary and then transmitted it through the spelling channel in a single transmission. How does the situation change if we suppose that the letters of the word are transmitted one by one, and that the receiver is allowed to iteratively revise his or her predictions in the light of the incoming data?

The probability of observing two events, $p(x_1, x_2)$, is the same as the probability of observing the first event and then the second, $p(x_1) p(x_2 | x_1)$. Since the logarithm turns products into sums, this means that if you observe a string of events, the sum of your individual surprisals equals the bulk surprisal you would experience from observing the whole sequence at once. For instance,

$$\log \frac{1}{p(x_1, x_2, x_3)} = \log \frac{1}{p(x_3 | x_1, x_2)} + \log \frac{1}{p(x_2 | x_1)} + \log \frac{1}{p(x_1)}.$$

As an illustration of what this means in the context of the spelling channel, suppose that you have forecasted, rather arbitrarily, that you are about to receive the message $x = \textit{fall}$. In fact, however, the actual output that you will see is $y = \textit{flat}$, so at some point along the way, your expectations will be violated. As you add up the surprisal values associated with each individual letter, you will gradually accumulate a total of 25.05 bits, the surprisal of the pair $(\textit{fall}, \textit{flat})$.

Table 3 Cumulative letter-for-letter surprisals at the string *flat* from the perspective of various hypotheses

<i>x</i>	–	<i>f</i>	<i>fl</i>	<i>fla</i>	<i>flat</i>
<i>for</i>	6.73	6.77	13.38	19.03	25.05
<i>large</i>	11.44	18.04	18.11	18.17	24.78
<i>fall</i>	12.73	12.77	19.39	19.39	25.05
<i>flat</i>	13.85	13.89	13.95	14.01	14.07

The surprisal need not be evenly distributed, though. When you have only seen the letter *f*, for instance, the hypothesis $x = fall$ is still perfectly consistent with the data, and your surprisal is moderate. It is only when the next letter turns out to be an *l* that your surprisal jumps upwards. More examples are given Table 3.

As can be seen from the last row of the table, the hypothesis *flat* terminates at a moderate 14.07 bits, which is not too far from the roughly 12 bits that would expected for a word of this length. For an ϵ between roughly 2 and 10, this hypothesis would thus be accepted, and the others rejected.

Notice also that the table contains occasional cases of reanalysis. In particular, when the *a* in *flat* is revealed, the hypothesis $x = fall$ hardly changes surprisal level. This is because the segment *fla* is consistent with a reversal hypothesis under which the *l* and the *a* were written in the reverse order. Observing the *a* thus in a sense explains away the unexpected *l* that would otherwise have to be explained as a spurious insertion.

4.4 Decoding Error in the Spelling Channel

These observations bring me back to the main topic of this section, the two types of decoding error in a noisy channel. As explained in Sect. 3.3, decoding by typical sets can either break when you have too many or too few candidates that can explain a received signal.

In the context of the spelling channel, an error of type I behaves largely as we would expect it to: Unrepairable nonsense words like *srxzk* do not look like any known words, and any hypothesis about the underlying message will thus keep accumulating surprisal as more letters are revealed (cf. Fig. 10). The most probable hypothesis for the string *srxzk*, the word *size*, eventually accumulates 31.78 bits of surprisal, well above the level expected for a word of this length. The string *srxzk* could thus lead to a decoding error of type I, since a decoder might simply give up trying to reconstruct the original word given this unexpectedly high surprisal score.

This situation should be compared to that of the non-word *flide* (cf. Fig. 11). As can be seen from the graph, there is a whole clutter of hypotheses which are roughly equally far away from this word. The hypothesis *slide* is the best candidate, but the

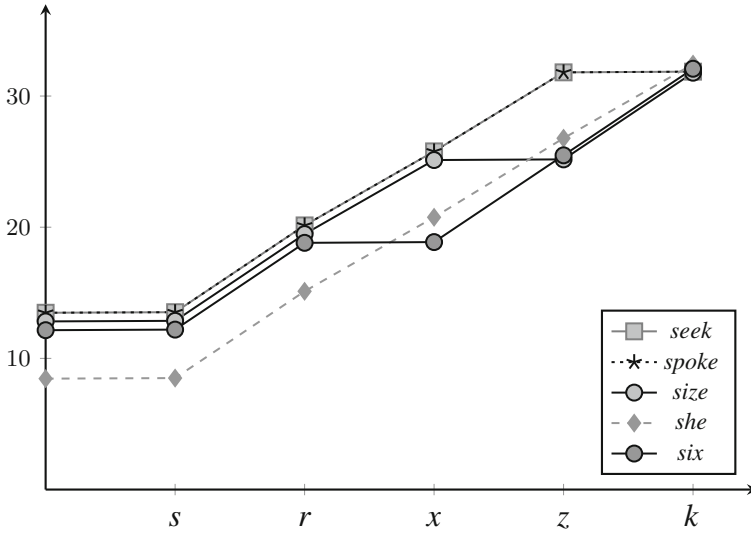


Fig. 10 Letter-for-letter surprisal values for various decodings of the string *srxzk*

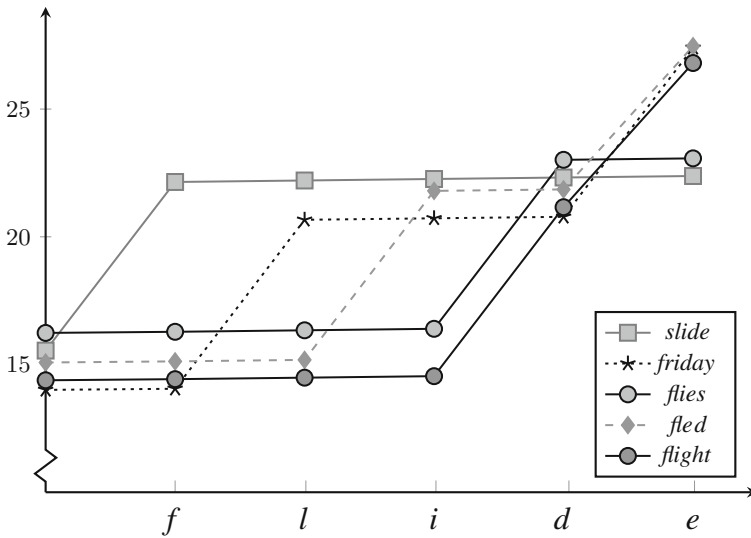


Fig. 11 Letter-for-letter surprisal values for various decodings of the string *flide*

difference up to the second best, *flies*, is only 0.70 bits. For many choices of ϵ , the two hypotheses *slide* and *flies* would thus be simultaneously rejected or simultaneously accepted. Both of those decisions would result in a decoding error.

From the present perspective, it is also interesting that the graph for the various hypotheses cross each other so frequently: After the fragment *fli* is presented, the

two hypotheses *slide* and *fled* are practically equiprobable. After the fragment *flid* is presented, the hypothesis *flies* is temporarily more probable than the eventual winner, *slide*. Again, this flip-flopping between roughly equiprobable hypotheses means that a decoder with a not too cautious choice of ε would be in a high danger of committing an error of type II.

This oscillation between competing hypotheses also means that the online decoding comes with its own “garden path” phenomenon: Hypotheses with high prior probability attract early attention which may in retrospect turn out to be unwarranted. This is, of course, particularly true if the output word contains errors or is otherwise abnormal so that ordinary expectations do not apply. Regardless of the cause, however, it should be emphasized that this phenomenon occurs here in a situation that does not involve any kind of “syntax” except the possibility of swapping two neighboring letters.

We have thus seen two kinds of statistical error related to two kinds of manipulated strings. Figuratively speaking, one abruptly pushes the receiver away, while the other keeps pushing the receiver back and forth between a portfolio of hypotheses that all seem to fit somewhat, but not quite. The first is associated with errors of type I, and the other with errors of type II.

5 Discussion

In the previous section, I gave a detailed description of a particular channel model and the decoding errors that were characteristic of it. This channel took English words as inputs and returned possibly misspelled versions of those strings as outputs. In this section, I will discuss the possibility of designing a similar model that can handle sentence-level decoding. My conclusion from this discussion is that it involves additional hard problems, but that it need not be impossible.

5.1 Two Challenges

I chose to illustrate my hypothesis about the N400 and the P600 by means of the quite simplistic spelling channel because I wanted to sidestep some technical complications that were irrelevant to my conceptual point. The word decoding problem is considerably easier than the sentence decoding problem that I really want to study.

Two facts in particular make words more manageable than sentences:

1. For most practical purposes, the vocabulary of English can be considered as a finite set, and this means that decoding can be done through exhaustive search. This property sets it apart from many other naturally occurring inference tasks which have NP-complete decoding problems, such as the most common versions of statistical machine translation (Knight 1999). Thus, a more realistic model of the English language would probably require a commitment to a specific search heuristic.
2. In spite of half a century of development, we still do not have a good computational theory of English. There is no shortage of formalisms and theories, but anybody who has done any serious attempt to apply these theories beyond a few toy sentences will know that we are nowhere near human-level predictive accuracy. Our language models will thus continue to be a bottleneck in the prediction of psycholinguistic phenomena, since we have no reliable and fully automatic way of computing the conditional probability of a sentence continuation given its context (let alone making categorical decisions about “grammaticality”).

Accurate predictions of the occurrence of N400 and P600 effects thus hinge on unsolved and perhaps even unsolvable problems, so perfect prediction is probably out of the question. However, there is a difference between bad and worse, and good approximations to English might still capture much of the statistical structure relevant to the prediction of these phenomena. In the next two subsections, I will give an example of one possibility for how we can approach the issue.

5.2 A Swap Model and a Trigram Model

Many of the linguistic phenomena associated with the P600 have to do with order effects, and it would thus be useful to have a descriptive statistic which measured to what extent a given English sentence respects the conventional order of words in the language. The crudest possible way of quantifying this level of conventionality is to produce a list of all the word pairs in the sentence, ignoring intervening words, and then check how frequently those word pairs occur in the same order in a reference corpus.

For instance, the sentence *There is still time* is associated with the pairs

(There, is), *(There, still)*, *(There, time)*, *(is, still)*, *(is, time)*, *(still, time)*.

Using the Brown corpus as a reference, we find that these word pairs occur with the frequencies

$1.1 \cdot 10^{-5}$, $5.3 \cdot 10^{-7}$, $6.8 \cdot 10^{-6}$, $2.9 \cdot 10^{-6}$, $7.9 \cdot 10^{-6}$, $1.3 \cdot 10^{-6}$.

Their corresponding surprisal values are consequently

$$16.47, 20.86, 17.16, 18.4, 16.95, 19.54.$$

On average, the surprisal for this string is thus 18.22 bits per word pair, corresponding to an “average probability” of $2^{-18.22} = 3.2 \cdot 10^{-6}$ (where the “average” is a geometric average).

This order model captures some long-distance relationships between words, and it can detect certain fishy-looking word orders. However, it contains no concept of long and short distances, only of “before” and “after.” It consequently has poor grasp of the strong, short-distance relations between neighboring words (e.g., the extremely low probability of the phrase *the of*).

These constraints are captured better by a Markov chain model (i.e., a table of conditional probabilities). Such a model can assign probabilities to the next word in a sentence given the immediate context, e.g., $p(\textit{still} \mid \textit{There is})$. Again using the Brown corpus, and assuming a context width of two words, such a model assigns an average surprisal of 18.72 bits to the string *There is still time*. This corresponds to a “average probability” of $2.3 \cdot 10^{-6}$.

In this case, the two models thus agree very much in their assessment of the input string, but this need not always be the case. In general, it can be advisable to use a mixture of several two complementary models in order to iron out some of the idiosyncrasies that each model has on its own. In the following, I have thus used an equal mixture of the order model and the trigram model, that is, a model of the form $p(x) = 0.5 p_1(x) + 0.5 p_2(x)$.

5.3 Neighborhood Decoding

An imperfect but very simple decoding method for channels like the spelling channel is to construct the set of source strings that are less than one editing operation away from the observed output, and then find the best candidate in that smaller set.

For instance, for if the output of a spelling channel was $y = \textit{tup}$, then the set of neighboring strings includes

- *tip, top, tap, tp, up, tups, tpu, upt, ...*

and many others. We can then decode $y = \textit{tup}$ by selecting the x on this list for which the joint probability $p(x, y)$ is the highest, hoping that the true solution was not more than one operation away. Implementing this decoding algorithm on the sentence level, using no other editing operations than swapping two words, we obtain decoding candidates such as those shown in Table 4.

As the table shows, even this simple statistical model, which has no concept of hierarchical structure or syntax, can discover and correct seemingly “high-level” errors. For instance, the decoder correctly pinpoints the words *police* and *prisoner* as a potential problem in the first example even though there is no semantic representation

Table 4 Potential decodings of an output string

Sentence (y)	$-\frac{1}{N} \log p(y)$
The police was trying to escape the prisoner	19.00
The prisoner was trying to escape the police	18.44
I'm afraid so	19.71
I'm so afraid	19.71
The patient looked at his doctor	18.84
The doctor looked at his patient	18.68
The doctor looked at his patient	18.68
The at looked doctor his patient	17.90
The house was parked in front of the car	10.84
The car was parked in front of the house	10.73
The water went straight through the boat	18.14
The boat went straight through the water	17.87
He walked straight to the room of the middle	11.09
He walked straight to the middle of the room	10.45
I heard a flash and saw thunder	19.17
I saw a flash and heard thunder	18.97
The hat put on his man	18.15
The man put on his hat	11.38
The mouse was chasing the cat	20.36
The the was chasing mouse cat	19.20

In each double row, the first sentence is the output string, and the second row is the a posteriori most probable input sentence that can be obtained by swapping two words. The surprisal values are only based on the source probabilities, since there is no explicit channel model

whatsoever in the model. In this and several other examples, corrections that seem to require sophisticated knowledge of syntax or semantics are thus achieved on the basis of nothing but raw word order.

Not all examples are equally convincing, though. The sentence *The mouse was chasing the cat* is, despite its simplicity, not decoded correctly. A more careful analysis of the corpus might reveal why this is the case, but at the time of this writing, I have no good explanations.

I should emphasize again here that the decoded sentence in each pair is the single most probable sentence in the neighborhood of the output string. The second sentence in each row was identified and selected automatically, based on the probabilities assigned by the mixture model. The table should thus not be read as a comparison between two hand-selected sentences.

Notice also that the channel probabilities—that is, the probability of the various distortions of a sentence—are not taken into account here; this will make a difference if these probabilities are very low. The string *I'm afraid so* will for instance not be corrected to *I'm so afraid* unless the cost of swapping two neighboring words (in terms of surprisal) is practically nil.

5.4 *Scaling Up?*

The toy model that I have proposed in this subsection is not a perfect model of English, and it could be improved in many ways. I have included it here to indicate that the principles discussed in the previous section pertain to any kind of noisy-channel model and not just the spelling channel.

Even this very simplistic model, however, was able to detect a number of different distortions of a sentence. Many of these distortions and their corresponding corrections superficially appeared language-specific or even syntax-specific in nature, but in fact, they were detected and corrected based only on very weak and general assumptions about the order of tokens in a sequence.

So while the neighborhood decoder for this channel model is quite crude and brittle, it does demonstrate that there is nothing inherently unrealistic about extending the basic statistical idea of this paper to more realistic channel models. Each such model comes with its own version of the two concepts of statistical error, and my claims about the N400 and the P600 will provide quantitative predictions in each such case—from sentence to sentence, and from word to word. The more difficult part of the future work thus consists in deciding what kind of channel speech and writing really is.

6 Conclusion

With the discovery of the P600, it seemed for a while as if the brain sciences had found a physiological counterpart of the distinction between semantics and syntax. Over the course of the last decade, this interpretation of the early results has been considerably complicated by further experiments which have produced P600 effects in a large variety of conditions, many of which are very difficult to explain in terms of syntax as such.

These discoveries have led a number of researchers to stipulate “parallel stream” explanations of the data: For instance, Bornkessel-Schlesewsky and Schlewsky (2008, p. 67) propose that the P600 occurs when cues from “animacy, voice and other language-specific information types” clash with the “plausibility” of the assignment of logical roles that these cues suggest. Kuperberg (2007, p. 44) argues that it is due to conflicts between “a semantic memory-based analysis and possibly a semantically-driven combinatorial thematic analysis.” Kolk et al. (2003, pp. 29–32) suggest that the cause is a conflict between an interpretation supported by a “conceptual bias” and another one supported by a “syntactic bias”.

These hypotheses are all very interesting and have a lot of empirical merit. However, I think they are formulated in too linguistic terms and on a too algorithmic level, and hence that they miss the larger conceptual picture. The working philosophy I have been applying in this paper is that we should take a principled perspective on the inference task our subjects are trying to solve (Anderson 1991). Instead of imme-

diately committing to a decoding algorithm in all its gory detail, we should choose a model of the communication situation and analyze its statistical properties before we postulate any specific heuristics or neurological implementations. This way, we can let machinery of statistics decide what counts as a “cue,” a “surprise,” or a “conflict” instead of inventing a new version of those concepts for every new experiment. This approach will promote models that are more modular, more transparent in their assumptions, and easier to simulate on individual words or sentences.

To illustrate what this approach entails, this paper has presented a very conventional noisy-channel model of spelling mistakes and looked at some of its properties. The decoding problem for this channel was simple enough to be solved by exhaustive enumeration, and this liberated me to focus on the statistical aspects of the situations. As the example showed, the exhaustive-search decoder for this simple model could run into two distinct kinds of statistical error: Words like *srxkz* produced an excess of surprisal and had no plausible reconstructions, while words like *flide* had several competing reconstructions, all of which had a surprisal level within the expected range. The decoding process could thus derail either because there were too few plausible decoding hypotheses, or because there were too many.

My claim in this paper is that these two kinds of decoding error—too few and too many plausible decodings—map onto the N400 and P600, respectively. The P600 is thus not produced by a conflict between two “streams” that handle specific types of information (such as animacy versus word order). Rather, it is caused by more general issues that arise whenever we try to make sense of ambiguous stimuli in an uncertain world. And while it is tempting to describe these issues in a highly task-specific vocabulary, I would argue that the process of decoding noisy signals provides a fruitful framework for understanding N400 and P600 effects.

More broadly, this statistical perspective on reading and listening could have some rather profound consequences for linguistic theory. One of the foundational assumptions of the Chomskyan program for linguistics is that sentences can be “ungrammatical,” a property that can allegedly be separated cleanly from the property of being nonsensical. From the statistical perspective, by contrast, the notion of “making sense” would not be assigned to a dark corner, but rather be the centerpiece of the whole theory. Noisy-channel coding is, at its core, a formalization of the process of guesswork and sense-making.

These speculations are, however, part of a larger discussion about the foundations of linguistics that cannot be settled by anything in this paper. I have proposed a specific explanation of the N400 and P600 effects which made no reference to grammatical and language-specific concepts, but instead explained these phenomena in statistical terms. This defuses one of the arguments in favor of linguistic nativism and related philosophical ideas, but it is still an open question how far the style of analysis can be pushed.

References

- Anderson, J. R. (1991). The place of cognitive architectures in a rational analysis. In K. VanLehn (Ed.), *Architectures for intelligence* (pp. 1–24). New York: Psychology Press.
- Bellman, R. (1952). On the theory of dynamic programming. *Proceedings of the National Academy of Sciences of the United States of America*, 38(8), 716.
- Bornkessel-Schlesewsky, I., & Schlewsky, M. (2008). ProminencePlausibility: An alternative perspective on semantic P600 effects in language comprehension. *Brain Research Reviews*, 5(1), 55–73.
- Chomsky, N. (1957). *Syntactic structures*. The Hague: Walter de Gruyter.
- Cover, T. (1975). An achievable rate region for the broadcast channel. *IEEE Transactions on Information Theory*, 21(4), 399–404.
- Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: Wiley-Interscience.
- Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3), 171–176.
- Fernández, E. M., & Cairns, H. S. (2010). *Fundamentals of psycholinguistics*. Malden: Wiley.
- Francis, W. N., & Kucera, H. (1967). *Computational analysis of present-day American English*. Providence: Brown University Press.
- Frenzel, S., Schlewsky, M., & Bornkessel-Schlewsky, I. (2011). Conflicts in language processing: A new perspective on the N400–P600 distinction. *Neuropsychologia*, 49(3), 574–579.
- Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20), 8051–8056.
- Hoeks, J. C. J., Stowe, L. A., & Doedens, G. (2004). Seeing words in context: The interaction of lexical and sentence level information during reading. *Cognitive Brain Research*, 19(1), 59–73.
- Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language*, 52(2), 205–225.
- Knight, K. (1999). Decoding complexity in word-replacement translation models. *Computational Linguistics*, 25(4), 607–615.
- Kolk, H. H. J., Chwilla, D. J., van Herten, M., & Oor, P. J. W. (2003). Structure and limited capacity in verbal working memory: A study with event-related potentials. *Brain and Language*, 85(1), 1–36.
- Kuperberg, G. R. (2007). Neural mechanisms of language comprehension. *Brain Research*, 1146, 23–49.
- Kuperberg, G. R., Sitnikova, T., Caplan, D., & Holcomb, P. J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cognitive Brain Research*, 17(1), 117–129.
- Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, 62, 621–647.
- Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427), 203–205.
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. In *Soviet Physics Doklady*, 10, 707.
- Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part i. *Biometrika*, 20A(1/2), 175–240.
- Nieuwland, M. S., & Van Berkum, J. J. A. (2005). Testing the limits of the semantic illusion phenomenon: ERPs reveal temporary semantic change deafness in discourse comprehension. *Cognitive Brain Research*, 24(3), 691–701.
- Osterhout, L., & Holcomb, P. J. (1993). Event-related potentials and syntactic anomaly: Evidence of anomaly detection during the perception of continuous speech. *Language and Cognitive Processes*, 8(4), 413–437.

- Rösler, F., Pütz, P., Friederici, A., & Hahne, A. (1993). Event-related brain potentials while encountering semantic and syntactic constraint violations. *Journal of Cognitive Neuroscience*, *5*(3), 345–362.
- Samson, E. W. (1951). Fundamental natural concepts of information theory. Technical report E 5079, Air Force Cambridge Research Center
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.
- van Herten, M., Kolk, H. H. J., & Chwilla, D. J. (2005). An ERP study of P600 effects elicited by semantic anomalies. *Cognitive Brain Research*, *22*(2), 241–255.
- van Herten, M., Chwilla, D. J., & Kolk, H. H. J. (2006). When heuristics clash with parsing routines: ERP evidence for conflict monitoring in sentence perception. *Journal of Cognitive Neuroscience*, *18*(7), 1181–1197.
- Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, *13*(2), 260–269.
- Weckerly, J., & Kutas, M. (1999). An electrophysiological analysis of animacy effects in the processing of object relative sentences. *Psychophysiology*, *36*(5), 559–570.

Orthogonality and Presuppositions: A Bayesian Perspective

Jacques Jayez

Abstract This work focuses on the reasons why the *projection* properties of presuppositional elements are not uniform. Presupposition projection designates the fact that operators that suspend the truth of a proposition have no effect on the presupposition it is associated with. For instance, with *Paul did not stop smoking*, the main content (that Paul does not smoke) is canceled by the negation whereas the presupposition that Paul has been smoking is left untouched. It has been observed that projection is not uniform across the different expressions that convey a presupposition (its *triggers*), and a distinction between *weak* and *strong* triggers has emerged. It seems that this distinction correlates with an intuition of (non-)orthogonality, that is, of (non-)independence between the main content and the presupposition. I argue that this intuition is correct to some extent but has to be clarified. To this aim, I propose a Bayesian approach, which provides a more precise rendering of orthogonality.

Keywords Presupposition projection · Strong triggers · Bayesian approach · Bayesian relevance · Layered meaning

1 Introduction

A significant part of the large literature on presuppositions (PP) has been concerned with the phenomenon of *projection*. Projection denotes the fact that operators like negation, interrogation or *if*, which normally cancel or suspend the truth of a proposition, have no effect on the PP associated with the proposition. For instance, in *Did Paul stop smoking?*, the truth of the proposition that Paul smokes is suspended, whereas the PP that Paul has been smoking remains active. The most natural reading

I am grateful to my two anonymous reviewers for convincing me with excellent arguments to rewrite the initial version of this paper. They are not responsible for the remaining shortcomings.

J. Jayez (✉)
ENS de Lyon and L2C2, CNRS, Lyon, France
e-mail: jjayez@isc.cnrs.fr

of the question is ‘Given that Paul has been smoking, is it true or false that he does not smoke now?’. However, in some cases, the PP itself can be suspended and it seems that the possibility of suspending the PP varies with the different expressions that convey it (its *triggers*). Recently, this has led to a distinction between *weak* triggers (or *soft* triggers), those that allow for suspending the PP, and *strong* triggers (or *hard* triggers), which do not. Although this distinction has some empirical support, it is far from transparent. Why should some triggers be ‘stronger’ than others? It seems that the difference depends on the (degree of) *orthogonality* of the PP and the main content (MC). For instance, in *Paul stopped smoking*, the MC (Paul does not smoke) and the PP (he has been smoking) seem to be somehow ‘entangled’, whereas with *Paul hit the target too*, the MC (Paul hit the target) and the PP (someone else did) seem to be independent (‘orthogonal’). In this paper, I show that orthogonality is not exactly what we need and that the difference between weak and strong triggers is amenable to a general principle regulating discourse attachment in a Bayesian framework. In Sect. 2, I recall the main notions and observations that are at the origin of the projection problem. In Sect. 3, I present introspective (Sect. 3.2) and experimental (Sect. 3.3) data that seem to support the distinction between weak and strong triggers. I discuss the interpretation of the experimental data in Sect. 3.4, arguing that they are actually based on a metalinguistic interpretation. I turn to orthogonality in Sect. 4. In Sect. 4.1, I present Abbott’s take on the weak/strong distinction and argues that her analysis in terms of *detachability* is misleading. However, I defend her intuition in Sect. 4.2 and show how it can be expressed in a Bayesian framework. Finally, in Sect. 5, I address two remaining problems. Admittedly, the empirical scope of this paper is limited because I do not study the full range of triggers (see Sect. 3.1 for a discussion of this problem). However, one has to start somewhere and I think that is not inconsequential to show that the distinction between weak and strong triggers, which has gained great currency in semantics, has to be reanalyzed.

2 Presuppositions and Projection

2.1 Basic Notions

Natural languages offer to speakers the opportunity to convey different types of information in a single utterance. The MC concerns the most basic truth-conditional contribution of the utterance, whereas the non-MC concerns, for instance, intellectual or emotional attitudes of the speaker, or what she presupposes to be true. So, (1a) informs the hearer that some friend of the speaker found a solution (MC), but also that this fact was unexpected (intellectual attitude) and that the speaker does not hold the mentioned friend in high regard (emotional attitude). (1b) asserts that Charles has talked to someone (MC) and presupposes that it is his sister.

- (1) a. Unexpectedly, my stupid friend found the solution.
- b. Charles has talked to his sister.

A quick test to separate MC and non-MC is to consider what the natural target of a direct refutation is. In most cases, people interpret the target as being the MC, see (2) and (3).

- (2) A—Unexpectedly, my stupid friend found the solution.
 B—It's false/You're wrong/You're mistaken.
 ~↗ Your friend did not find the solution
 ↗↗ It's not unexpected or/and your friend is not stupid
- (3) A—Charles has talked to his sister.
 B—It's false/You're wrong/You're mistaken.
 ~↗ Charles did not talk to his sister
 ↗↗ Charles has no sister

Among the expressions conveying a non-MC, *presuppositions* (PP) have been the subject of intensive research. Intuitively, a PP is a piece of information which is taken for granted or presented as such (Stalnaker 1974) or must be true in every context where the sentence conveying it can be asserted without contradiction (Beaver 2001; Heim 1983). A hallmark of the non-MC is its tendency to *project*, that is, to remain unaffected by operators that negate or suspend the MC, such as negation, interrogation or *if*. For instance, whereas the proposition that my friend found the solution does not receive a definite truth-value in (4), the non-MC (my friend is stupid) is endorsed by the speaker exactly as in (1a). A similar difference holds for negation and conditional.

- (4) Did my stupid friend find the solution?
 ↗↗ My friend found the solution
 ~↗ My friend is stupid

2.2 Ideal and Actual Projection

In an ideally simple world, projection would be uniform across all types of triggers. What is the situation in the real world? Non-MC tends to project rather neatly for *conventional implicatures* (Potts 2005) such as evaluative adverbs (*unexpectedly*) or expressives (Determiner + *stupid* + N, interjections, etc.). What happens for PP? There is clearly a tendency to project, see Chierchia and McConnell-Ginet (1990), Geurts (1999), Beaver and Geurts (2013) for different surveys. For instance, there is a strong intuition that all the forms in (5) presuppose that Paul has been smoking.

- (5) a. Paul stopped smoking.
 b. Paul didn't stop smoking.
 c. Did Paul stop smoking?
 d. It is possible that Paul stopped smoking.
 e. If Paul stopped smoking, he must feel better.

Concerning the difference with conventional implicatures, conditional structures that suspend a PP seem to have no effect on implicatures, as suggested by the difference between (6a) and (6b), which sounds odd or heavily metalinguistic.

- (6) a. If Paul has been smoking, he has stopped.
 ↗ Paul has been smoking
 b. ? If my friend is stupid, then my stupid friend found the solution.

However, this is not true for every type of conventional implicature, in particular for appositive relatives, see (7). In a context where Paul is John's friend, the speaker does not necessarily endorse the view that Paul is stupid.

- (7) If Paul is stupid, John can expect that his friend, who is not very clever, won't find the solution.

If projection for PP was uniform, the difference between conventional implicatures and PP would be essentially a matter of domain. For instance, conventional implicatures could perhaps be conceived as concerning all the domains where evaluation, whether emotional or intellectual, plays a central role. However, it seems that, whereas implicatures project in a systematic way, it is not the case for PP. For instance, Karttunen (1971) introduced a difference between *full* factives like *regret*, whose PP always project, and *semi-factives* like *discover*, whose PP may be suspended, as in (8). See Beaver (2004) for a thorough discussion.

- (8) a. If I discover that I was mistaken, I will admit publicly I was wrong.
 ↗ I was mistaken
 b. If I regret that I was mistaken, I will admit publicly I was wrong.
 ↘ I was mistaken

Other ways to suspend the PP include the following ones.

1. Ignorance about the PP (Geurts 1995), as in (9).
2. Negation of the PP in answers (Cummins et al. 2013), as in (10), where the A1-B1 exchange is judged by speakers to be better than the A2-B2 exchange.
3. Abusch (2002, 2010) suspension test, as in (11)

- (9) [Context: Paul has never seen or heard about Mary smoking, but she seems very nervous and restless. He wonders whether Mary has been smoking and is trying to stop.]

Is Mary quitting smoking?
 ↗ Mary has been smoking

- (10) A1—Has John stopped smoking?
 B1—No, because he never smoked before.
 A2—Has John smoked again?
 B2—No, because he never smoked before.

- (11) a. I have no idea whether Paul participated in the Road Race yesterday.
 But if he won it, then he has more victories than anyone else in history.
 (=Abusch's 2010 example 3d)
- b. ?? I don't know if Paul participated in the race, but if Mary participated too, they probably ran together.

These and similar examples suggest that (i) PP projection has a certain degree of variability and (ii) that it is probably not just a matter of context or pragmatic considerations. Stalnaker (1974) defended the idea that projection is a default or preferred behavior, which can be more or less plausible in different contexts. For instance, he explained the non-projection in (8a) by noting that, if the speaker of (8a) presupposes that she has missed the point, she necessarily realizes that she did so. So, by contraposition, if she does not realize she has missed the point, she cannot presuppose that she did, hence the lack of commitment to the PP (absence of projection). So, a simple reasoning about the plausibility of the PP leads to abandoning it. Unfortunately, it is difficult to understand along the same lines the difference between *too* or *again*, which are robustly projective, and *stop* or *win*, for which the PP can be suspended more easily.

3 Weak and Strong Triggers

Examples (10) and (11) suggest that there are at least two classes of triggers. *Strong* triggers (ST) always give rise to projection whereas *weak* triggers don't. This distinction is supported both by introspective and experimental data. Before presenting these data, however, I need to clarify what the ambition and limits of trigger taxonomies can be. This is done in the next section.

3.1 A Disclaimer in Form of Clarification

In the previous section I have mentioned some well-known cases. A reviewer complains that I have confined my attention in this paper to "a rather limited number of presupposition triggers". This is quite true but can hardly be avoided. First, as evidenced by the (partial) list in Beaver and Geurts (2013), there is a huge set of presumed PP triggers. The goal of the present paper is not to study the variability of triggers *per se* but only with respect to projection. There are a lot of important or minute differences that are not captured by a distinction in terms of projection and I did not intend to address them for a simple reason: I am unable to offer a 'complete' theory of PP triggers, meaning a theory in which each trigger would be described completely and correctly. I suspect that such a theory would require a detailed lexical analysis of each trigger. I am not aware of the existence of such a theory in the literature and I do not feel committed to proposing one.

Second, it is almost impossible to have clear-cut introspective intuitions on a very wide range of PP triggers, and it might be impossible to have clear intuitions on some triggers except in well-known (and generally trivial) cases. The strategy of the paper is, accordingly, to focus on a number of items that are considered as representative of some major categories: aspectual verbs like *begin*, emotive factives like *regret*, clefts, ‘strong’ adverbial triggers like *too*. Although their number is indeed limited, they span a not insignificant number of *categories* of PP triggers. I assume that the categories addressed in the paper are sufficiently interesting to be worth studying (I might be wrong).

My proposal is based to a large extent on experimental observations. The interest of an experimental approach is twofold. First, by going beyond the intuitions of isolated speakers, there is more chance to detect general tendencies, if any, and to compensate for the variability of introspective judgments. Second, an experimental perspective forces one to remain aware of the limits of large scale comparisons. Measuring the ‘projectability’ of triggers with comparison tests raises the question of the interpretation context. When we are given sentences in isolation or with a very limited or very specific context, we are not necessarily able to quickly scan a sufficiently large set of contexts in order to find a ‘good’ one, that is, a context that makes the sentence natural. So, a naturalness signal¹ can reflect the degree of context accessibility/construction. To use a generic term, I will call the task of inferring a context that makes a linguistic stimulus plausible *context abduction*. To my best knowledge, there is at the moment no established technique for measuring context abduction. However, when a context can be provided which makes a linguistic stimulus natural, whereas the same stimulus is massively perceived as not natural in isolation, one can reasonably suppose that the licensing context cannot be easily abducted. This is the main idea behind the experiment reported in Sect. 5.2. There, two triggers considered as strong are considered, *regret* and clefts. It is shown that, in suitable contexts, the PP of these triggers can be explicitly suspended without creating an impression of contradiction (it does not project). However, the contexts that make non-projection possible are not the same for the two triggers. Therefore, *comparing* the two triggers would be more risky because it would amount to comparing the behavior of two items in different contexts. Clearly, this is a general problem. Since PP triggers are very different, it is difficult and perhaps impossible to show in a rigorous way how exactly they compare to each other in terms of *degree* of projection. It is for instance possible that triggers that allow non-projection and are thereby similar differ nonetheless in terms of degree of projection strength but that the difference is impossible to evaluate. Once the context is given, this possible difference might disappear. When isolated sentences are considered, observed differences might be attributed to context abduction rather than resistance to projection. If the same contexts could be used, we would of course be in a better position to adopt a truly comparative perspective.

¹Whether this signal is an explicit judgment or a different manifestation such as a reading time, fixation or ERP measure does not change the basic problem.

The upshot of these remarks is that, whereas it would be extremely interesting to compare triggers pairwise, it is also a difficult task, which seems to be beyond our reach for the time being.

3.2 Introspective Data

The idea that PP triggers differ in their projection behavior is mentioned in several places in the linguistic and philosophical literature on PP. However, there is in general no systematic attempt to lay the foundations of a classification. Three notable exceptions are Abbott (2006), Abusch (2002, 2010) and Zeevat (1992). Zeevat's proposal is discussed in Sect. 3.3. Its relation to the WT/ST distinction is less clear than in the works of Abbott and Abusch, which also offer the advantage of capitalizing on more recent contributions. In this section, I adopt Abbott's and Abusch's point of view and focus on the three main points that are to be found in their papers.

First, Horn (1972) identified a possibility of suspending a PP by means of a special *if*-clause or a *may not* construction (12a). According to him, the first sentence must be negative (12b)

- (12) a. John does not realize that Sue loves him, if indeed she does.
 b. # If John realizes that Sue loves him (and in fact she may not), then he'll tell us.

One can observe that *regret* or *again*, for instance, are less felicitous in environments comparable to (12a).

- (13) a. ? John does not regret that Sue loves him, if indeed she does.
 b. ? John has not studied the problem again, if indeed he has studied it before.

Second, there is the *discover/regret* contrast mentioned by Karttunen, whose examples can be adapted to avoid the first-person effect mentioned by Stalnaker (see Beaver (2004)). Various similar observations suggest that epistemic factives like *discover* or *know* are WT, whereas emotive factives like *regret* or *be glad/sad/happy/surprised* are ST.

- (14) a. If someone discovers that I was mistaken, I will admit publicly I was wrong.
 ↗ the speaker was mistaken
 b. If someone regrets that I was mistaken, I will admit publicly I was wrong.
 ↘ the speaker was mistaken

Finally, there is Abusch's test, used in (11). The structure of the test sentences is: $?p'$, *but if* p, q , where p presupposes p' . What is tested is the possibility of accommodating the missing PP (whose truth has been suspended by a preliminary ignorance statement) in the local context corresponding to the antecedent of the *if* p, q conditional.

Table 1 Weak and strong triggers

Category	Items	Horn	Abusch
Temporal markers	<i>after, since</i>	Weak (?)	Weak (?)
Epistemic factive verbs	<i>discover</i>	Weak	Weak
Emotive factive verbs	<i>regret, be surprised</i>	Strong	Strong
Aspectual verbs	<i>begin, stop</i>	Weak	Weak
Exclusives	<i>only</i>	Weak	Weak
Scalar additives	<i>even</i>	Strong	Strong
Discourse markers	<i>too, again</i>	Strong	Strong
Sortal restrictors	<i>bachelor</i>	(?)	(?)
Clefts	<i>It's ...who</i>	Weak	Weak
Quantifiers	<i>each, every</i>	Weak	Weak (?)
Definite descriptions	<i>the, her</i>	Weak	Weak
Names	...	Weak	Weak

A sample of items

Summarizing, there are reasons for distinguishing between WT and ST on a purely observational basis. However, the various observations that have been proposed give only a patchy picture of the weak/strong contrast. Table 1 gathers a number of introspective guesses about the category of different triggers and I provide illustrative examples in Table 2.² The inventory of triggers follows Beaver and Geurts (2013) with one exception. I did not consider manner adverbs such as *quickly* or *clumsily*, because they are part of the MC in most cases. In fact, if we were to accept that manner adverbs are PP triggers because they presuppose some event, I don't see why we should not include adjectives and other modifiers, on the account that they presuppose the existence of the entity to which they apply.

The pseudo-Karttunen and Horn tests use negation and *if* as suspension operators. Horn's test is easier to use because it does not appeal to a conditional structure and I have focused on it for the two tables. I did not include examples for epistemic and emotive factives because they are discussed in the paper. For focus particles, I limited myself to mentioning exclusives. The presuppositional status of other focus particles has to be established. For instance, as explained in (Beaver and Clark 2008, pp. 70–72), scalar additives have a very different semantics than exclusives. In some cases, I realized that I had very shaky intuitions and added a "(?)" in Table 1. I didn't use diacritics (?, #, *) in Table 2, in order to let the reader develop her own intuition.

²I can't but repeat my cautionary statement of Sect. 3.1: introspection can be misleading, especially, as here, when comparing very different items.

Table 2 Weak and strong triggers. Examples of use

Category	Test type	Example
Temporal markers	Horn	<i>Paul didn't change his mind since he read the review, if indeed he read it</i>
	Abusch	<i>I don't know whether Paul read the review, but, if he changed his mind since, the final version of the paper is going to be pretty different</i>
Aspectual verbs	Horn	<i>Paul has not stopped smoking, if indeed he has been smoking</i>
	Abusch	<i>I don't know whether the victim has been smoking before, but, if he has stopped smoking, we should find traces of tobacco tar in his lungs</i>
Exclusives	Horn	<i>I doubt that only Paul came, if indeed he came</i>
	Abusch	<i>I don't know whether Paul came, but, if only him came, it must have been disappointing</i>
Scalar additives	Horn	<i>I doubt that even Paul missed the target, if indeed it would be THAT surprising</i>
	Abusch	<i>I don't know whether Paul missing the target would be very surprising, but, if even him missed the target, we'll see what the others will do</i>
Discourse markers	Horn	<i>I doubt that Paul missed the target again, if indeed he missed it before</i>
	Abusch	<i>I don't know whether Paul missed the target before, but, if he missed it again, really, it's a pity</i>
Sortal restrictors	Horn	<i>This Jordan is not a bachelor, it seems, if indeed Jordan is a male</i>
	Abusch	<i>I don't know whether this super rich Jordan is a male, but if he is a bachelor, he certainly has a lot of women chasing after him</i>
Clefts	Horn	<i>It's not Paul who solved the problem, if indeed someone solved it</i>
	Abusch	<i>I don't know whether someone solved the problem, but, if it's Paul, he must be very proud</i>
Quantifiers	Horn	<i>Since he settled only very recently, I doubt that every professional nurse in the town met the new doctor, if indeed there are professional nurses in such a little town</i>
	Abusch	<i>I don't know if any student cheated, but, if every student who cheated has been expelled, we'll see if some students are missing at the beginning of the next term</i>
Definite descriptions	Horn	<i>Paul didn't find the key, if indeed there is a key for this door</i>
	Abusch	<i>I don't know if there is a key for this door, but, if Paul found the key, we are lost</i>
Names	Horn	<i>I never met this Mr Slush, if indeed there is someone with this name</i>
	Abusch	<i>I don't know whether there is a Mr Slush, but, if Dora met him she is in trouble</i>

3.3 Experimental Data

Certain—but not all—introspective observations are confirmed by the experiment reported in (Cummins et al. 2013). Cummins et al. draw their initial motivation from Zeevat’s (1992) distinction between three different types of triggers, a distinction which I summarize in the next paragraph.

Resolution triggers form the first category. They are strongly anaphoric, that is, they demand that the entities they refer to be present in the environment (discourse or situation) at the moment the PP trigger is used. Zeevat mentions definite descriptions, *when-* and *after-*clauses, as well as—tentatively—clefts. *Lexical* triggers correspond to “concepts with the applicability conditions”, which means that certain conditions—corresponding to the PP—must be met for the concept to be applicable. An example provided by Zeevat is *regret*. At this stage, there is a significant complication. Zeevat assumes that a trigger can trigger several PP. For lexical triggers like verbs and nouns, the sortal information—whether we refer to an event or to an individual—is *lexical* in nature, whereas the *descriptive* information—what is predicated of the event or the individual—can be *resolutive*. For instance, Zeevat analyses an example like *John believes that the king is bald* as divided into a lexical presupposition, a resolution presupposition and a content (the MC of the present paper), see (15).

- (15) John believes that the king is bald.
 Lexical PP : x (x has the type of individuals)
 Resolution PP: $x \wedge king(x)$
 Content : $B_j bald(x)$

Zeevat also considers triggers like *again*, *also*, *another* and *too*. They are characterized by their ability to involve parts of the context that are not normally accessible, as illustrated by (16).

- (16) John believes that Mary was in Egypt. Sue was there too

Cummins et al. intend to compare the experimental correlates of lexical and resolution triggers. In fact, the set of triggers they choose contains (i) lexical triggers (*regret*, *continue*, *stop*), *only*, and comparatives, (ii) two triggers of the third type (*again* and *too*) and (iii) *still*. This is a little confusing but I assume that their goal is essentially to compare triggers which look for an antecedent (*again*, *too* and *still*) and triggers which express their PP directly in their lexical content. I will call the former *anaphoric* and the latter *lexical*. Concerning comparatives, I am skeptical as to their presuppositional status, since the alleged PP is not robust in a plain assertion: *Mary is more clever than Paul, who is really rather stupid*.

In their main experiment, Cummins et al. use an evaluation task where subjects have to rate some dialogs on a 5-point scale. The distribution of possibilities is as follows, for the trigger *again*.³

³As noted by a reviewer, the bottom right case is not very clear. The answer *No, because he didn't lose his wallet before* could be interpreted as Endorsing MC + Refuting PP. I won't try to take this

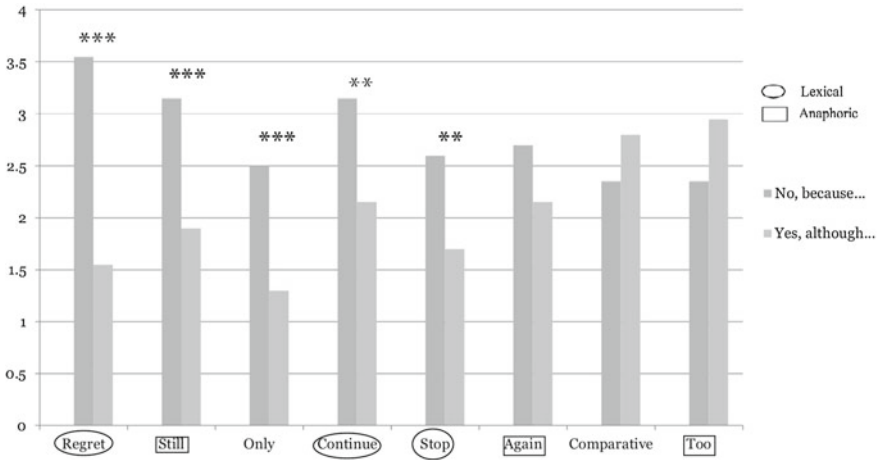


Fig. 1 Results of the second Cummins et al. experiment

The initial expectations of the authors are that the combination of endorsing the MC and refuting the PP are better for anaphoric triggers than for lexical triggers and that the combination of refuting the MC *and* the PP is better for lexical triggers than for anaphoric triggers. The results are summarized in Fig. 1. The stars indicate significance in the difference between the scores for the *No, because ...* (left bar) and *Yes, although ...* (right bar) answers.

Based on the introspective data of the previous section, one would expect that *regret, continue, stop* and *only* are WT, whereas *still,*⁴ *again* and *too* are ST. As indicated, I leave aside the case of comparatives. There is no significant difference for *again* and *too*, but they don't pattern alike. *Still* does not align with *too* and *again*, although it is considered to be a ST. Similarly, there is a difference for the (presumably) ST *regret*. So, the results delivered by the experiment do not coincide with introspection and call for an analysis of the task itself.

The main question that experiments of this type raise is to determine how the subjects interpret the dialogs they have to evaluate. Consider a set of pairs like those in Table 3. When a subject sees the answer *Yes, although he didn't lose his wallet before*, what does she understand? It seems pretty obvious that the answerer means that she considers the proposition that Brian lost his wallet as true. But what is the meaning of *although*? It is unlikely that *although* signals an incompatibility between the fact of losing one's wallet and the proposition that one has not lost one's wallet before, because such a connection would be very obscure. It is likely, in contrast,

(Footnote 3 continued)

interpretation into account, but its existence is an additional symptom of the difficulty of determining what the target of *Yes/No* and *although/because* is exactly in this type of experiment.

⁴In my intuition, *still* does not behave differently from *too* or *again* with respect to the Horn and Abusch tests.

Table 3 The four possibilities in Cummins et al. design

Did Brian lose his wallet again?		
PP \ MC	Endorsing	Refuting
Endorsing	Yes, he did lose his wallet again	Yes, although he didn't lose his wallet before
Refuting	No, he didn't lose his wallet this time	No, because he didn't lose his wallet before

that *although* is ‘metalinguistic’, that is, it indicates that the answerer has not used an appropriate expression. Is it also the case with *because* in other examples? What are the most reasonable predictions one can make, given the interpretations assigned to the different micro-dialogs presented in the experiment? Do these predictions correspond to what is observed by the authors? In the next section, I try to clarify the import of Cummins et al. experiment, before reconsidering the WT/ST distinction in the light of introspective and experimental observations in Sects. 4 and 5.

3.4 Discussion of the Experiment

Let us suppose that the *yes* and *no* particles in the various answers are free to target the MC alone, the PP alone or the MC+PP combination. I will call this set of configurations the *free attachment scenario*. Under this scenario, there is no *a priori* plausibility in favor of one specific attachment, in comparison to the others. Clearly, this scenario is only a theoretical and rather implausible possibility. Nevertheless, it is useful to enumerate all the possibilities and sort out those that stand a chance.

Figure 2 shows the different possibilities of the free attachment scenario. Table 4 compares what is predicted under the free attachment scenario to what is observed by Cummins et al.

Let us first describe the predictions of Table 4. With Yes-1, that is, when *Yes* targets the MC alone, we predict that the perceived quality of the *Yes, although ...* answer should depend on the interpretation of the *although* discourse marker. If

Fig. 2 The *a priori* possibilities for attachment

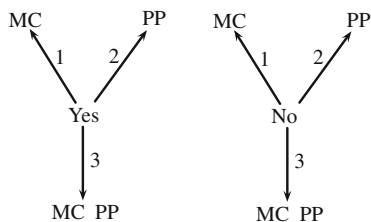


Table 4 Predictions and observations

Category	Diagnostic	Prediction for ST and WT	Observed
Yes-1	Depends on <i>although</i>	WT = ST	WT \neq ST
Yes-2	Contradictory (PP and not PP)	WT = ST	WT \neq ST
Yes-3	Contradictory (PP and not PP)	WT = ST	WT \neq ST
No-1	Explanation relation obscure	WT = ST	WT \neq ST
No-2	Redundant explanation (X <i>because</i> X)	WT = ST	WT \neq ST
No-3	?OK	WT = ST	WT \neq ST

it is interpreted as introducing only a correction, the answer should be judged as correct because its meaning can be paraphrased by “what you say_{MC} is true, but your question presupposes something which is false”. *Although* might also be interpreted as introducing an opposition. In that case, a form ‘A *although* B’ would mean ‘A is true but B, others things being equal, makes the probability of A decrease’. Under the opposition interpretation, the negation of the PP would be seen as making the probability of the MC decrease. It is not clear that this interpretation corresponds to our intuition in the general case. However, we might imagine contexts in which this dependency holds. For example, sailing across the Atlantic (as a pilot) probably demands some experience. So, one might think that, if you never crossed the Atlantic before and if you are not an experienced sailor, it is rather unlikely that you can succeed. If Paul is an unexperienced sailor, one might imagine a dialog like (17). In the case of (17), B’s answer should sound natural, since it means something like “Yes, he has crossed the Atlantic, although it was his first experience”.

- (17) A—Has Paul crossed the Atlantic again?
 B—Yes, although he never crossed it before.

Whatever the interpretation for *although* is, the prediction for an exclusive MC attachment is that ST and WT should not be different. Let us review more briefly the other attachment configurations. For Yes-2, the target of the answer is the PP, which makes the answer self-defeating, since the PP is endorsed and rejected at the same time. A similar diagnosis has to be made for Yes-3, where the MC and the PP are the targets. In the latter two cases, no difference between ST and WT is predicted. All in all, for *Yes*, the free attachment scenario does not predict any difference between WT and ST.

With No-1, the answer presents the negation of the PP as a reason to believe that the MC is false. The causal relation behind this reasoning is rather obscure in the general case. It might perhaps be clearer in particular contexts, like with *although* in

the case of Yes-1. With No-2 (the PP is the target), *because* introduces a redundant explanation ($\neg p$ *because* $\neg p$). The last case might be more felicitous. *No* targets the MC and the PP and the answer might be interpreted as ‘it is not true that A_{MC} and B_{PP} because it is not true that B_{PP} ’. Under the three *No* versions, like in the case of *Yes*, the free attachment scenario does not predict any difference between WT and ST.

What are the observations? Generally speaking, the predictions of the abstract theory (free attachment) are not borne out. There *are* differences between triggers. How can we make sense of what Cummins et al. observe? The authors report two different findings. First the *Yes, although ...* answer is judged to be globally worse than the *No, because ...* one. This can be explained straightforwardly if we assume that the answers target the *conjunction* of the MC and the PP. In that case, *Yes, although ...* is predicted to be anomalous for reasons we just saw with the comment of Yes-3 in Table 4. The speaker endorses A *and* B and rejects B immediately after, which is a plain contradiction.

The second finding is the difference between WT and ST—or, more precisely, between certain items classified as ST and certain items classified as ST. Unfortunately, it turns out that the results of the experiment are difficult to interpret, given that there is a potential confound created by *because*. Under a *descriptive* (causal or logical) interpretation, ‘A *because* B’ presents B as a cause of or a reason for A. Under a *metalinguistic* interpretation, ‘A *because* B’, where A has a negative form $\neg C$, presents B as a reason for not using the form of words associated with C.

Horn (2001, Chap. 6) exposes all the intricacies of metalinguistic negation. The type of metalinguistic interpretation that is relevant in the context of the present paper can be described as in (18).

- (18) A metalinguistic interpretation consists in assuming the existence of an eventuality while denying that it can be described correctly by a certain form of words.

To illustrate, consider (19). B1’s answer is a plain negative answer. There was no event of Paul tidying up his room. This answer does not commit B1 to something happening at all. For all we know, Paul might have stayed quietly in his armchair. B2’s answer, where the first syllable of *tidy up* is stressed, conveys the idea that something happened which cannot be properly called tidying up. It is well-known that marking focus with a stress, as here, is conducive to a contrast interpretation, in which several alternatives are competing (Beaver and Clark 2008; Erteschik-Shir 2007; Rooth 1992).

- (19) A—Did Paul tidy up his room?
 B1—He didn’t.
 B2—He didn’t TIdy up his room.

In our case, the relation between the metalinguistic interpretation and the PP is relatively clear: the failure of the PP makes its trigger linguistically inappropriate. Note, however, that a metalinguistic interpretation can make use of the MC. For

instance, in (20), B underlines the fact that the MC (Paul does not smoke) is not an exact description of the actual situation, although it bears some resemblance to it.

(20) A—Paul stop smoking?

B—He didn't STOP, but he is smoking less and less.

In order to interpret some of the results of the Cummins et al. experiment, one has to determine, as far as possible, whether the subjects see certain dialogs with a *No, because ...* answer as metalinguistic or descriptive. I won't consider the MC-addressing metalinguistic interpretations in discussing Cummins et al. experiment, because it is clear that the MC is not directly rejected in the *No, because ...* answers. Consider dialog (21). A MC-addressing metalinguistic interpretation would amount to paraphrasing B's answer by 'It is not appropriate to say that Brian lost his wallet because he did not lose it before', which hardly makes sense.

(21) A—Did Brian lose his wallet again?

B—No, because he did not lose his wallet before.

However, the MC might be indirectly affected by the rejection of the PP, and I will have to take this possibility into account. It corresponds to what I have called the *descriptive* interpretation, in which negating the PP of the question ('Brian did not lose his wallet before') could make the probability of the negation of the MC ('Brian did not lose his wallet') increase.

How could we discriminate between the descriptive and the (PP-addressing) metalinguistic interpretation in our case? If the interpretation is descriptive, we have two possibilities. (i) The negation of the MC is caused or justified by the PP. This is very implausible in the general case. For instance, the *No, because ...* answer to a test sentence like *Did Ben regret arguing with his boss?* would be interpreted as "Ben has no negative feeling about arguing with his boss because he didn't argue with his boss". If this paraphrase means that Ben does not regret having argued with his boss because he didn't argue, what is explained or justified is not, in fact, the MC alone but the MC+PP compound. If the paraphrase means that Ben has no negative feelings with the very *idea* of arguing with his boss because that did not happen, the meaning is at best unclear since the descriptive (causal/logical) link is not particularly obvious. More importantly, interpreting the *No, because ...* answer as a '–MC *because* –PP' pattern, is counterintuitive. If it was a natural answer, the following dialog should sound felicitous, which is not the case.

(22) A—Did Ben regret arguing with his boss?

B—Ben has no problem with the idea because he didn't argue with his boss.

(ii) The second possibility for the descriptive interpretation is that the *No, because ...* answer rejects a logical *and*-conjunction because one of its terms is false. For instance, the following dialog (23) would receive the interpretation in (24). Similarly, the answer on the MC would correspond to (25).

(23) A—Does Toby continue to watch films?

B—No, because he didn't watch films before.

(24) A—Does Toby watch films and did he watch films before?

B—No, because he didn't watch films before.

(25) A—Does Toby watch films and did he watch films before?

B—No, because he does not watch films.

Given (24) and (25), one would not expect much difference between a *No, because* answer refuting the PP (24) and a *No, because ... answer* refuting the MC (25). This is precisely what we *don't* observe. In a first experiment, Cummins et al. (2013) show very clearly that speakers prefer an attachment to the MC, which echoes previous similar experimental findings in (Jayez 2010).

Summarizing, the most plausible candidate for the *No, because ... answer* is the PP-addressing metalinguistic interpretation, where a speaker underlines the inappropriateness of a linguistic form by denying the PP it triggers. Moreover, this interpretation is expected to remain rather marginal, because it creates a garden-path effect. The *No, because* part triggers an expectation that the answerer addresses the MC (descriptive interpretation) whereas the rest of the sequence only makes sense under a metalinguistic interpretation. Given that both *Yes, although ...* and *No, because ...* answers are most naturally interpreted as metalinguistic, we can account for the observed difference between the scores for the *Yes, although ...* and the *No, because ...* answers. Moreover, we expect that, when the PP is difficult to suspend (ST), even the *Yes, although ...* answer will not be perceived as natural, so the difference between the *Yes, although ...* and the *No, because ...* will be smaller than when the PP is easier to suspend (WT).

We can conclude that Cummins et al. experiment confirms that there exists a difference in the robustness of the PP, in agreement with the general feeling that some triggers (the ST) are more robust than others (the WT). However, the results leave us with two questions.

(1) The correspondence with the introspective data is imperfect. First, *still* and *regret* seem to be ST, whereas the experiment classifies them as WT. Second, the reason why *again* and *too* have inverse profiles is unclear, even if the difference between the scores for the two answer types is not significant. Third, how is it that *too* has a markedly better score for the *Yes, although ...* answer than *regret, only, continue* or *stop*?

(2) The nature of the distinction between WT and ST remains a bit mysterious. Clearly, not all triggers behave the same. Is this sufficient to conclude that they fall into 'natural' classes, e.g. anaphoric versus lexical or strong versus weak? Should the difference be conceived instead on an item-per-item basis, or should we consider an intermediate taxonomy, with several small classes, possibly originating from a set of parameters?

4 Separation and M-Relevance

4.1 Abbott on Detachability

Abbott (2006) offers a precious cue. She notes an apparent correlation between the status of triggers and the relation between MC and PP. The PP of ST correspond to pieces of information which are orthogonal to the MC, as evidenced by the *nondetachability* criterion, in the sense of (Grice 1989, Chaps. 1–2). Grice discusses nondetachability in the context of conversational implicatures. A conversational implicature is non detachable whenever it is not possible to ‘say’ the same thing, that is, to convey the same MC, without triggering the same conversational implicature, unless some feature of the alternative expression is relevant to the determination of the original implicature. Conversational implicatures are in general (but not always) nondetachable, whereas conventional implicatures are in general detachable. Interestingly, (Grice 1989, p. 43) mentions the implicatures of PP triggers as nondetachable. In *Paul stopped smoking*, the proposition that Paul has been smoking will survive any attempt to reformulate ‘what is said’ in Grice’s sense.

Let us assume for the moment that Grice is right. Clearly, with ST, it is possible to ‘say’ the same thing without the PP. So ST would be analogous to conventional implicatures, which, moreover, are reputedly hard to suspend (Potts 2005), like ST. Abbott explains the common resistance of ST and conventional implicatures to suspension by resorting to a mechanism of intention attribution. A speaker has a choice between communicating the PP of a ST or not. Since the PP and the MC are informationally independent, and the speaker *chose* to communicate the PP, the latter must be part of her communicative intention. For instance, in using *Mary hit the target again* instead of *Mary hit the target*, a speaker lets us think that drawing our attention to the PP is an essential part of her message, hence the incompatibility with suspension, which would result into an incoherent conversational plan.

It seems that Abbott’s account suffers from the same problem as Grice’s. It is just technically incorrect to claim that PP of verbs like *stop*, or of WT in general, are nondetachable. There are perfectly good ways to ‘say’ the same thing without conveying the same PP. First, there is the possibility of asserting the MC alone, like *Paul does not smoke* with respect to *Paul stopped smoking*. One might object that, when using *Paul does not smoke*, we do not convey the same content because the transition from an activity of smoking to its negation gets lost. At the root of this objection, there is a frequent cognitive illusion about the MC. A plausible logical formula corresponding to *Paul stopped smoking* is shown in (26.2). t is a time point (or a small incompressible time interval), I_1 and I_2 are time intervals, \triangleleft (\triangleright) are the relations of right- (left)-abutting. $t \triangleleft I$ iff t (or its smallest upper bound, if t is an interval) is the greatest lower bound of I . $I \triangleright t$ is defined symmetrically. (26.2) says that there is a past time t such that Paul has been smoking just before t and didn’t smoke for some time just after t . The MC is shown in (26.3). It says that John didn’t smoke for some time, just after some time t . We might express the MC simply by saying *Paul didn’t smoke* or *Paul has not been smoking*, or, if we are sure that Paul

didn't start smoking again, by *Paul does not smoke*. Admittedly, we do not thereby convey the same information as with *Paul stopped smoking*, but the information we lose depends on the PP, not on the MC. We are just asserting that there is some past time point after which Paul didn't smoke, during some time interval which extends or not to the present, and this is indeed the 'substance' of the MC.

(26) 1. *Paul stopped smoking*

2. $\exists t, I_1, I_2(\text{past}(t) \ \& \ I_1 \triangleright t \ \& \ \text{smoke}(\text{Paul}, I_1) \ \& \ t \triangleleft I_2 \ \& \ \neg \text{smoke}(\text{Paul}, I_2))$

3. $\exists t, \quad I_2(\text{past}(t) \ \& \quad t \triangleleft I_2 \ \& \ \neg \text{smoke}(\text{Paul}, I_2))$

Interestingly, what I presented as a cognitive illusion is also to be found in theories of conventional implicatures. Potts (2005) sees implicatures as orthogonal to the rest of the sentence, in the sense that their truth or falsity does not affect the truth or falsity of the rest of the sentence. For example, in (1a), whether my friend finding the solution is unexpected or not or whether he is stupid or not does not affect the truth or falsity of my friend finding the solution. In contrast, with a sentence like *Charles has talked to his sister*, a PP failure—assume that Charles has in fact no sister—makes the whole event of speaking to Charles's sister impossible. In the present approach, the point is that the fact that Charles has no sister does not affect the *main content* that Charles has talked to someone, who happens, by virtue of the PP, to be his sister. The difference between a conventional implicature trigger and a PP weak trigger is that the MC and the implicature are clearly separated.

One might object that, in (26), the MC and the PP share the temporal variable t . If we assign different attitudes to presupposing and asserting, we can reformulate (26.2) as $\exists t$ (the speaker presupposes $\phi(t)$ & the speaker asserts $\psi(t)$). A similar remark applies to factives, where the speaker presupposes a proposition and asserts that some agent entertains a certain attitude (knowing, discovering, regretting, etc.) with respect to the same proposition. The presence of a shared variable might appear to be the core of non-detachability. Unfortunately, a ST like *again* exhibits the same property. In saying *Mary hit the target again*, a speaker asserts that Mary hit the target at some time t in the past and presupposes that Mary had hit the target at some point t' before t . So, the t variable is shared. More generally, sharing a temporal variable is unavoidable in every situation where two events are declared and located with respect to each other.

Moreover, there is another formulation that does not convey the PP, the 'flat' one, in which the MC and the PP are juxtaposed within a conjunction, for instance, *Paul has been smoking until Monday and didn't smoke (for some time) after Monday*, instead of *Paul stopped smoking on Monday*. So, I don't see how to reconcile the MC versus PP distinction with the idea that the PP is a nondetachable element. Of course, one could choose to reject the distinction and see WT as holistic bundles, but, under this view, it seems that nothing is left of the initial problem addressed by Abbott. If WT are not 'triggers', but rather co-entail the MC and the PP, what is the whole discussion concerning the WT/ST difference about?

A reviewer mentions the characterization of soft triggers (=WT) proposed by Abrusán (2011) and wonders what relation it might have to Abbott's idea.⁵ Abrusán considers only verbal WT and offers a *predictive* theory. Given a verbal WT, she proposes criteria to determine what the PP is. According to her definition 32 (p. 509), a PP is an entailment that obeys one of the two following conditions.

1. The entailment can be expressed by a sentence that is not necessarily about the event time of the matrix predicate, or,
2. the entailment can be expressed by a sentence that is not necessarily about the event time of the sentence that expresses the most direct answer to the background question determined by grammatical marking. (1)

Condition 1 can be illustrated by *Paul knows that Mary solved the problem*. Clearly, the event of Mary solving the problem can span a different time interval than the interval corresponding to Paul's epistemic state. Condition 2 takes care of focus problems. For instance if the complement clause of a factive like *discover* is in focus, it is grammatically marked as answering a background question. The complement clause is then interpreted as expressing a secondary main point (in addition to the primary main point expressed by the matrix verb). Since a PP has to be temporally independent of the time span of every main point in the sentence, this predicts that, in some cases there is no room left for a PP. In example (27), where the embedded clause is in focus, there is no mention of an event that would be temporally disconnected from either the event expressed by the main verb or the event described by the embedded clause. So, nothing is presupposed, or, in the traditional terminology, the PP that Paul lied to the committee is suspended.

(27) If someone discovers that [Paul lied to the committee]_F it will be a scandal

Two remarks are in order. First, Abrusán's approach predicts that the PP of *discover*-like verbs cannot be suspended unless the embedded clause is in focus. I wonder whether this is not too strong. Imagine the following dialog (28), in a context where Paul, an assistant of the President, has just testified before an official committee in a bribe case. The final conditional sentence does not commit the speaker to the belief that Paul lied. Must this sentence be in focus? It is not necessary that there be an intonational focus. One might argue that the clause is 'indirectly' in focus because it is an answer to a question that is relevant to the main question about Paul's situation. This certainly makes sense but could be taken to be just the joint effect of two aspects: (i) the context does not establish that Paul lied (otherwise, suspension would be impossible) and (ii), in general, answers to questions must not mention irrelevant information. If we apply Abrusán's analysis, *S' = he has lied to the committee* is not in focus, since there is not grammatical marking, and we get the following configuration: (i) there is a sentence *S* (*someone discovers that S'*) with an entailment *S'*, (ii) *S'* is not necessarily about the event time of the main verb

⁵For time reasons, I was not able to properly analyze recent work by Romoli, which seems relevant to the issue addressed in this paper.

(*discover*), (iii) there is no background question determined by grammatical marking with respect to which S' could constitute an answer. So the PP projects, contrary to intuition in certain cases. We might drop the grammatical marking requirement but, in that case, we would have to say that every piece of information that might contribute to answer the main question will be 'in focus', or rather, in the terms of Simons et al. (2011), *at-issue*. According to Simons et al. (2011), only those elements of interpretation that are not at-issue can project, or, equivalently, at-issue elements cannot project. Their proposal would account for non-projection in (28), at least if one considers that the embedded clause *he has lied to the committee* is at-issue (=is relevant to a question under discussion). Another—conservative—option is to consider that the context makes one of the suspension/non-suspension interpretations more plausible than the other.

(28) A—What is exactly the situation of Paul?

B—I don't know for sure. At the moment, he seems to have the favors of the President, but, if someone discovers that he has lied to the committee, he is going to run into trouble.

The second remark to be made on Abrusán's approach is that, although she occasionally touches on the issue of ST, the goal of her paper is not to propose a theory of the distinction between WT and ST. This makes it hard to compare to Abbott's perspective and creates a more serious problem. She considers WT to be verbal, probably because she needs an event structure. It's not necessarily true (see Tables 1 and 2) and makes it difficult to extract from her analysis of WT elements that could be directly recycled to address the WT/ST distinction.

4.2 *The Separation Intuition*

In spite of the nondetachability problem, there is an extremely useful insight in Abbott's approach, namely the idea that conventional implicatures and ST are somehow akin. Their relationship can be expressed very simply: for conventional implicatures as well as ST, the two components of the content, that is the non-MC and the MC, are *separated* in the message. The part of the message that conveys the MC does not convey the non-MC and reciprocally. As a result, with a ST, the PP-addressing metalinguistic interpretation, which depends on the rejection of the PP, cannot concern the part of the message that conveys the MC since this part *is* appropriate. So, it must concern the trigger only. If one assumes with, among others, Partee (1991), that PP are unfocused by default—they normally constitute or contribute to the *topic*—they alone cannot be the target of a refutation unless some additional cue is provided, like a special stress, as in (29).

(29) Paul didn't hit the target aGAIN, since he missed it at his first try.

It follows from the separation property that ST are expected to be very poor targets under a PP-addressing metalinguistic interpretation, which is compatible with some of Cummins et al. observations.

However, some other observations seem difficult to reconcile with the separation property. Although *still*, *again* and *too* would be classified as ST according to Abusch's suspension test (see example 31), they don't behave similarly according to the Cummins et al. test. In fact, a closer look reveals that *too* is the only item for which the *Yes, although ...* answer gets a better score than the *No, because ...* one. I conjecture that the main reason is that, in the setting of the experiment, the PP of *too* remains vague. The four stimuli used by Cummins et al. are of the type *Did Ian win a prize too?*, which would ideally require that we know *who* won a prize, apart from Ian. Although the standard theory has it that the PP would be *someone else won a prize*, it has been repeatedly noted that this PP is in most cases trivially satisfied and that the standard usage is that the actual PP involves a particular individual, different from Ian, winning a prize (see Kripke 2009). I won't be concerned here with the exact presuppositional structure of *too*, which is the matter of some debate, but it is clear that a dialog like (30) is extremely difficult to interpret. In addition to the fact that the general preference for addressing the MC is defeated, a specific PP, of the form *x won a prize* cannot be construed from the first sentence and the rejected PP in the *No, because ...* answer is a consequence of an (absent) specific PP and not a genuine PP. The configuration is not logically impossible, but requires probably some extra processing, which results in a stronger impression of incoherence, when compared to triggers like *again* or *still*.

- (30) A—Did Ian win a prize too?
 B—No, because no one else won a prize.

Summarizing, the separation property provides a convenient explanation for the metalinguistic interpretations and, since I have argued that the most plausible interpretation is metalinguistic in the case of the Cummins et al. experiment., this property is a good candidate for explaining their observations.

However, I have not discussed Abusch's suspension test, exemplified in Table 2 and in (31) below. Abusch's idea is that ST do not permit PP (local) accommodation, in particular in the antecedent of a conditional, as in (31). One can observe that, more generally, ST like *still*, *again* or *too* cannot be easily suspended. In contrast to (9), repeated below, (32) and (33) are infelicitous.

- (31) a. ?? I don't know if Paul participated in the race, but if Mary participated too, they probably ran together.
 b. ?? I don't know if Paul participated in the race before, but if he participated again, he must be very proud.
 c. ?? I don't know if Paul was in the garden, but if he is still there, ask him to bring back the spade.

- (9) [Context: Paul has never seen Mary smoking, but she seems very nervous and restless.]

Is Mary quitting smoking?

↗ Mary has been smoking

- (32) [Context: Paul has never seen or heard about Mary smoking, but she seems very nervous and restless.]
 a. Is Mary still smoking?
 b. Is Mary smoking again?
- (33) [Context: Paul has no evidence that Mary has a particular problem, for instance personal difficulties.]
 Does Mary have professional problems too?

Examples of this kind do not seem to have much to do with a metalinguistic interpretation. They are not isolated cases. Consider sentences of the type illustrated in (34). Although (34a) and (34b) refer to the same set of situations, the latter is less natural.

- (34) a. I saw that Paul smokes, but I don't know whether he is starting or just continuing.
 b. ? I saw that Paul smokes, but I don't know whether he is starting or still smoking.

The contrast between prefixes and adverbs expressing repetition provides similar observations. In many Romance languages the *re-* prefix is used to express repetition (Sletsjøe 1979). For example, in French, there are two alternative forms for expressing repetition. One can use a ST like *encore*, *à/de nouveau* (again), *une fois de plus* (once more), or, in some cases, prefix a verb with the iterative marker *re-*. Imagine that Paul climbed the Matterhorn (*Cervin* in French) twice. One can express the repetition as in (35d) or (35e), where the repetition prefix is underlined.

- (35) a. Paul climbed the Matterhorn
 b. Paul a fait le Cervin
 c. Paul climbed the Matterhorn again
 d. Paul a encore fait le Cervin
 e. Paul a refait le Cervin

Now, suppose that Paul is climbing the Matterhorn with Mary and that she is impressed by the easiness with which he orients himself, finding his way through the different possible paths and avoiding dangerous areas. She tells herself (36). The intended interpretation is that Mary suspects that Paul climbed the Matterhorn before and that he didn't let her know, maybe for showing off. Under this interpretation, (36b) is not felicitous, like its English counterpart (36c). In contrast, (36a) is much better.

- (36) a. C'est étonnant! Peut-être qu'il le refait et qu'il n'a pas voulu me le dire.
 b. ?? C'est étonnant! Peut-être qu'il le fait encore et qu'il n'a pas voulu le dire.
 c. ?? Amazing! Maybe he climbs it again and he didn't wish to tell me.

Table 5 MC/PP Integration

Type of trigger	Mode of integration
Temporal markers	PP = complement
Factives	PP = complement
Aspectual verbs	PP conveyed by complement
Exclusives	Convey a part of the MC
Sortal restrictors	PP conveyed by a complement of the MC verb
Clefts	PP expressed by the complement
Quantifiers	PP expressed by an argument of the main verb
Determiners	PP expressed by an argument of the main verb
Proper nouns	PP expressed by an argument of the main verb

Mary observes the efficiency of Paul (the ‘data’) and she makes an hypothesis in order to explain it. In other terms, she selects a dependency corresponding to an acceptable likelihood. In Bayesian terms, the appropriate expression is a conditional probability, the probability of observing the (efficient) behavior of Paul assuming both the MC (he is climbing the Matterhorn) and the PP (he has done that before), in symbols ($\Pr(\text{efficiency}|\text{MC}\&\text{PP})$). Note that the MC *is* involved and cannot be taken out of the probabilistic dependency. Suppose that we are in a context where Paul is not climbing the Matterhorn, Mary could not observe that he is particularly efficient. The difference between (36a) and (36b) is due to the lexical packaging of an information piece which is otherwise perfectly identical in the two cases.

Examples (32), (33), (34), (35) and (36) suggest that a PP associated with a ST cannot be in focus. The infelicitous questions or suppositions in the various examples try to address the PP of triggers like *again* or *still*, that is, triggers that express the PP separately from the MC. With other triggers, the MC and the PP are integrated. The PP is related to an argument of the MC verb. A notable exception is *only*, which conveys a part of the MC as well as a part of the PP. For instance, as a an NP-modifier, *only* acts on a NP and a VP, assembling the content of both to produce the PP and the MC. So, there is no separation between the two parts of the content, but, rather, an entanglement determined by *only* (Table 5).⁶

The resistance of ST to focusing does not explain Abusch’s observations. In (11b), repeated below, the PP triggered by *too* is not in focus alone, unlike in the previous examples. However, there is a common feature between the two cases, the fact that an operator (question, *if*) cannot access a part of the message that is not involved in the expression of the MC. E.g., in (34a), (36b,c) and (11b), the operators have

⁶When it modifies a subject NP, *only* can be described categorially as in (a), where s , s' and s'' are string variables, a form $s : x, y$ denotes the fact that string s has the semantic type x and the syntactic type y , u is the semantic type of individuals, \mathcal{P} the semantic type of properties and t the semantic type of truth-values.

(a) 1. syntax: $(s'' : t, S/s' : \mathcal{P}, VP)/s : u, NP$
 2. $PP = \mathcal{P}(u)$, $MC = \forall x((x \neq u) \Rightarrow \neg \mathcal{P}(x))$.

to access the ST trigger, to apply a focus or conditional operator. With (11b), the desired interpretation is ‘If Paul participated (PP) and if Mary participated too’, that is, a local accommodation configuration in which the PP is in the scope of the conditional operator. As noted by Abusch, such a configuration is problematic. It seems, in fact, that the operator can’t but ignore the PP triggered by a ST, exactly like the focus operator has to ignore the PP of ST. In the next section, I try to account for this particular behavior of operators with ST.

- (11) a. I have no idea whether Paul participated in the Road Race yesterday.
 But if he won it, then he has more victories than anyone else in history.
 (=Abusch’s 2010, example 3d)
- b. ?? I don’t know if Paul participated in the race, but if Mary participated too, they probably ran together.

4.3 Integrating M-Relevance

In this section, I show that, in order to exploit the separation property, we need another ingredient (called *M-relevance*, for *monological relevance*) which allows us to understand why separation has the effect it has. Unless otherwise indicated, I consider only non-metalinguistic interpretations.

Introducing M-Relevance Jayez (2010), elaborating on an earlier proposal by Ducrot (1972) shows experimentally that it is practically impossible to construct a discourse attachment to the non-MC exclusively, at least for a subset of discourse relations that involve Bayesian relevance, à la Merin (1999, 2003), see definition (37).⁷ In this paper, relevance is restricted to monologues, and I will use the label M-relevance to make this restriction more apparent.

- (37) A proposition p is positively (negatively) relevant to another proposition q in a belief state $\mathcal{B} =_{\text{df}} \log[\text{Pr}_{\mathcal{B}}(p|q)/\text{Pr}_{\mathcal{B}}(p|\neg q)]$ is > 0 (< 0). So, relevance is just the difference between the log-likelihoods. Intuitively, p is positively (negatively) relevant to q iff updating the current context (belief state) \mathcal{B} with q makes the probability of p increase (decrease).

Causal relations are a case in point. With ‘A *because* B’ patterns, A is presented as having positive M-relevance to B or, equivalently, B is presented as making the probability of A increase. (38.a) exemplifies an attachment to the MC: it is because smoking is unhealthy that Paul does not smoke. Whatever (38.b) means, it *cannot*

⁷I follow Merin and the standard practice of using log of quotients instead of quotients. This ensures that relevance is null whenever $\text{Pr}(p|q) = \text{Pr}(p|\neg q)$. I am not convinced that the present notion of relevance is necessarily the best but I don’t know of any proposal about how to compare different notions.

Table 6 M-Relevance relations

<i>Paul stopped smoking</i>	<i>because it's unhealthy</i>
$\log[\Pr(\textit{bad-for-health} \neg\textit{smoke})/\Pr(\textit{bad-for-health} \textit{smoke})]$	
<i>Mary might be quitting smoking</i>	<i>because she is nervous</i>
$\log[\Pr(\textit{nervous} \textit{quit-smoking})/\Pr(\textit{nervous} \neg\textit{quit-smoking})]$	

mean that Paul smoked because he liked that,⁸ that is, any attachment to the PP alone is precluded. Jayez (2010) shows that this restriction holds for conventional implicatures as well. With abductive relations, where the speaker deduces possible causes from observations, the MC can correspond to the hypothetical cause. For (38a) and (9), we have the M-relevance relations in Table 6.

- (38) a. Paul stopped smoking because it's unhealthy.
 b. Paul stopped smoking because he liked it.
- (9) [Context: Paul has never seen or heard about Mary smoking, but she seems very nervous and restless. He wonders whether Mary has been smoking and is trying to stop]
 Is Mary quitting smoking?
 ↗ Mary has been smoking

The obligatory character of an attachment to the MC for M-relevant constituents is expressed in (39). The interaction between M-relevance and attachment corresponds to the following intuition. When a speaker makes it manifest that she introduces in the discourse some information that potentially makes the probability of a certain proposition increase or decrease, this information must concern the part of the message that conveys the MC, although, as we will see below, it can also concern other parts. It is not surprising that the MC should be involved, given that the MC corresponds to the foregrounded information in linguistic communication.⁹ It would be much more surprising that a speaker bothers to mark a certain content as foregrounded (the MC) and forgets it altogether in the rest of the discourse (see Jayez (2014) for a detailed discussion). This makes sense for monologues. For dialogs the situation is more complex, because an addressee can reject implicitly or explicitly the foreground/background divide of the speaker. I won't discuss the case of dialogs in this paper and will focus only on M-relevance.

The notion of M-relevance extends to suspension operators (negation, interrogation, *if*). When a speaker applies such an operator to a proposition, the MC must be involved for the same reason as with M-relevance attachments: why would the speaker foreground some information if it was to be ignored?¹⁰

⁸A reviewer notes that (38.b) might mean that the reason for Paul stopping smoking is that he liked smoking. This is true, but, in that case, the attachment is to the MC and the plausible interpretations are hard to get (maybe Paul is wary of addiction?).

⁹This is a widely shared intuition, see for instance Grice 1989, pp. 121–122 and Abrusán 2011.

¹⁰A reminder: I am not talking about the metalinguistic interpretation in the present section.

(39) **M-relevance and MC**

A M-relevance-based attachment must exploit the MC.

(39) leaves open the possibility that the MC acts as a cause or a consequence, in patterns of the form $MC \xrightarrow{\text{cause}} X$, $X \xrightarrow{\text{cause}} MC$, or $MC \xrightarrow{\text{cause}} \neg X$ and $X \xrightarrow{\text{cause}} \neg MC$ in the case of opposition relations. Importantly, (39) must not be interpreted as implying that the probability of the MC is modified. For instance, in a sentence like *Paul lost the race although he had a hard training period*, the probability of Paul losing the race is maximal for the speaker. Relevance concerns *potential* probability modifications. p is potentially relevant to q in \mathcal{B} if, in any context (belief state) \mathcal{B}' minimally different from \mathcal{B} and such that \mathcal{B}' does not satisfy p and does not satisfy q , updating \mathcal{B}' with q makes the probability of p increase or decrease. Constructing a proper definition of ‘minimally different’ is a complex problem and I assume that most definitions proposed in the literature are compatible with the definition of potential relevance (see van Benthem et al. (2009) for a recent discussion of the problem from the point of view of modal logic).

Attachments to *both* the MC and the non-MC are not disallowed by M-relevance. For instance, in *Paul has a strong will since he stopped smoking*, attributing a strong will to Paul is motivated by the *conjunction* of the fact that he was a smoker up to some point in the past and that he has not been smoking after that point.¹¹ This is exactly the type of attachment we have in (9). The possibility that Mary has been smoking and is no longer smoking is motivated by the observation that she is somewhat agitated. There are similar observations for ST, e.g. *Paul has hit the target. If he hits it again, it will be a real feat*. If the PP of a ST can be taken into account in an attachment, what does the distinction between WT and ST consist of exactly?

Connecting M-Relevance and Separation In Sect. 4.2, we have seen that focus- and *if*-operators cannot access the PP unless it is conveyed by a part of the linguistic message that also conveys elements of the MC. M-relevance demands that a relevance-based discourse relation in a monologue make use of the MC. Is there a relation between the two constraints?

An operator must apply to (access) a part of the linguistic message. This part is not necessarily the whole sequence of words on the right of the operator, as evidenced by focus-, question- and *if*-operators, which shunt ST triggers. A discourse relation is much more ‘abstract’ or ‘non-local’. It can exploit the propositional information in general, no matter what its linguistic realization or its degree of implicitness are. Thus, it is not surprising that discourse relations can access the PP of ST, since they are part of the general discourse information. This difference granted, the MC is an obligatory ingredient for operators and discourse relations and this common constraint interacts with the nature of operators and discourse relations to produce two different but related effects. With relevance discourse relations, the MC is necessarily used, but other pieces of information, including the PP can contribute to support the relation. With operators, again, the MC is a necessary ingredient, but, in contrast with

¹¹A reviewer proposes *Paul has a strong will since he isn't smoking* as an example where only the MC is present.

discourse relations, the operator cannot exploit several additional sources, because it cannot abandon its argument to jump to another part of the sentence, which entails that it can apply only to a PP if this PP is conveyed by the same part as the MC.

I register the constraint of obligatory attachment to a MC-part in (38).

(40) **M-relevance and separation**

A (non-metalinguistic) M-relevance-based discourse attachment in a monologue cannot target a part of the message that does not convey elements of the MC.

A consequence of (40) is that, for ST, the PP will tend to be ignored, except under the metalinguistic interpretation with a special focus marking on the PP, typically, some focal stress as in (29). This accounts for the observation that, in general, the PP of ST cannot be suspended. However, one may wonder what happens if the context somehow makes the PP salient. In Sect. 5.1, I report experimental findings which show that suspending the PP of *too* (a ST) is possible when the process of local accommodation is ‘boosted’ by the interpretation process. For WT, (40) predicts that a relevance-based attachment will use the MC and *can* also use the PP if we have a form of words that assembles the PP of a ST with its MC, instead of putting them into two different slots of the message. However, emotive factives are classified as ST, which does not square well with my analysis. In Sect. 5.2, I show that the apparent counterexample of factives can be dealt with in terms of context abduction and is not, in fact, an exception to constraint (40).

5 Loose Ends

5.1 Accommodation

If we think of the way in which we *learn* to use presuppositional terms, it is highly likely that in most contexts, the PP is satisfied. For instance, it is repeatedly noted in the literature on first language acquisition that *again* or *more* are part of the young (≤ 2 years) child lexicon, see Kauschke and Hofmeister (2002) for *wieder* (*again*). For English, the consultation of <http://childfreq.sumsar.net/> suggests that *more* and *again* are more frequent than *stop* and *start*.¹² This does not necessarily indicate that ST are learned before WT *because* they are strong. They could simply correspond to situations that a young child is likely to encounter and/or to categorize. It is nonetheless interesting that all those triggers are learned at a very early stage in language development, since they involve very robust PP, or in a more linguistic terminology, quasi-automatic projection. This is in agreement with our intuition that PP triggers would lose any presuppositional character if they were learned in

¹²Interestingly, as observed by a reviewer, one might argue that *stop* is more similar to *more* and *again* than *start*.

just random contexts, including those in which the PP is not satisfied.¹³ The fact that presuppositional terms are learned preferably or exclusively in contexts where there is no PP failure lends support to Lassiter’s analysis in terms of a probability threshold (Lassiter 2012). According to Lassiter, whenever p is an atomic sentence which carries the semantic PP \underline{p} , “a speaker should not utter p unless $\Pr(\underline{p})$ meets or exceeds a high threshold θ according to her epistemic state, and she believes that her audience also assigns \underline{p} at least probability θ .” (Lassiter 2012, Definition 9). This constraint might be too strong and reproduce, in a probabilistic setting, the difficulties which affect the notion of common ground. Accordingly, I will use only the first part of the constraint, that is, I will assume that a speaker communicates that the PP is highly probable in her own epistemic state.

When a PP has not been explicitly endorsed by the speaker or is not easily recoverable from the context, there is always the possibility of *accommodating* the PP, that is, of assuming, at least provisionally, that it is true. The notion of accommodation has a long history within the research field of PP and I won’t try to comment on the different approaches. Suffice it to say that it is to be expected that ST are resistant to accommodation, as noted by Abusch (2002, 2010). Accommodation is used when the PP is associated with the MC, since the entry point of a relevance-based interpretation is the MC carrier. With ST, the interpretation does not take the PP into account (it does not have to). As a result, if the PP does not have a strong probability at start, it will not be accommodated, which may result in a conflict. Let S be a sentence containing a trigger T and C a context for this sentence. The appropriateness¹⁴ of a WT in the sentence, given the context, is jointly controlled by the MC of the sentence in the context and the PP of the trigger in the contextualized sentence. This is very easy to formulate in Bayesian terms, by means of the conditional probability that the appropriateness of the trigger reaches a certain value, say x , given that the probabilities for the MC and the PP reach certain values, say y and z .¹⁵

$$(41) \Pr(\text{appropriate}(T, S, C) = x | \Pr(\text{MC}(S, C)) = y \ \& \ \Pr(\text{PP}(T, S, C) = z))$$

Equivalently, in a traditional causal network representation Pearl (2009), the appropriateness of a WT is represented as a *collider* assembling the MC and the PP, whereas a ST is appropriate only in virtue of the PP it expresses (Fig. 3). Colliders are graphs where a dependent variable is influenced by several factors simultaneously, which amounts to having a terminal node (the dependent variable) connected to the various factors by a bunch of edges.

¹³As noted by a reviewer, it is possible—and probable in fact—that children do not master the whole spectrum of usages of a given trigger. However, it is difficult to imagine that they would use presuppositional terms like *again* or *more* with a PP failure.

¹⁴The notion of appropriateness is to be kept distinct from that of truth. Obviously, a trigger can be used felicitously with respect to the belief state of the speaker and be inadequate with respect to the actual state of affairs.

¹⁵In contrast to what is usually done in Bayesian networks, the factors that influence the dependent variable (the appropriateness) are not sets of values of a random variable, but sets of values of a probability.

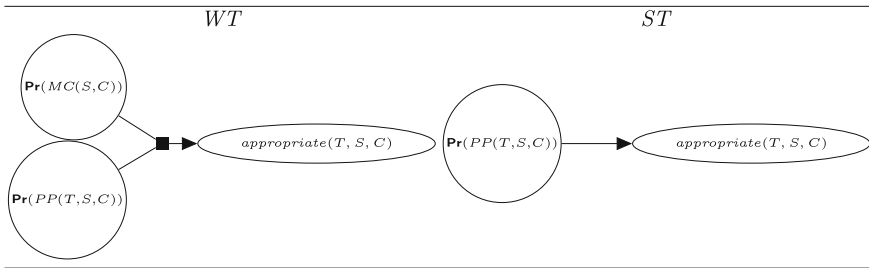


Fig. 3 Influence for a WT and a ST

By backward reasoning, when the appropriateness and the MC probability of a WT are determined, the probability of the PP can be adjusted, and possibly raised to a sufficient level if it is not determined (accommodation). Nothing similar can happen for ST since the probability adjustments are relativized to those parts of the network that contain PP nodes. However, it is possible that independent material makes the PP of a ST highly probable. In that case, ST are rescued. For instance Jayez and Mongelli (2013) and Jayez et al. (2013) show that the French counterparts of stimuli similar to (42) are positively evaluated. More interestingly, there is no significant difference between (42a) and (42b) patterns, which suggests that the hypothetical missing PP (Paul goes to the party) is accommodated *independently* of the presence of *too*. This is most probably due the discourse structure, which favors a sort of counterfactual reasoning attributed to Paul. When *but* is substituted for *because*, the subjects rate the resulting sentences as very poor. The contrast *but/because* is illustrated in Fig. 4. So, the missing PP is not really ‘accommodated’, in the sense of activated to satisfy

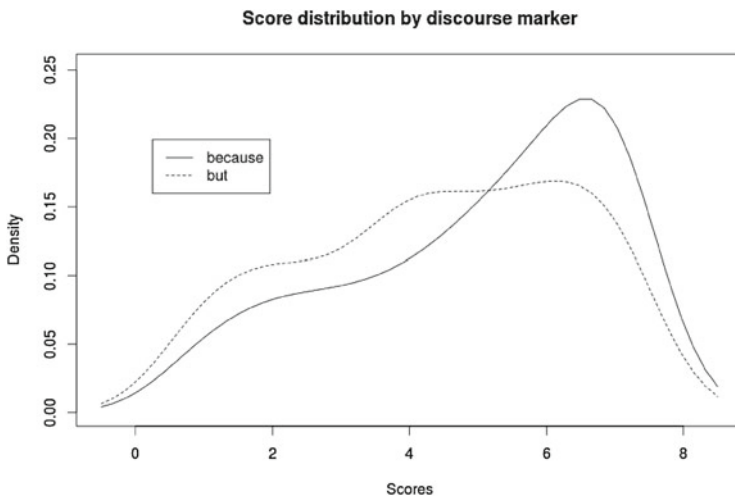


Fig. 4 Accommodation-like effect with *because* and *but*

the requirements of a trigger, but rather independently activated and retrieved by the trigger.

- (42) **Context:** Paul has been invited to a party. He is on very bad terms with Mary and they would prefer not to come across each other. Paul thinks that Mary might have been invited as well.
- a. I don't know whether Paul will go to the party because/??but, if Mary goes, it will be embarrassing.
 - b. I don't know whether Paul will go to the party because/??but, if Mary goes too, it will be embarrassing.

5.2 Are *Factives* a Problem?

Abbott (2006) notes that the existence of verbs like *regret* is a problem for her idea that the difference between WT and ST stems from nondetachability. *Regret* is classified as a ST by Abusch, whom Abbott follows, but is nonetheless detachable. I won't try to assess the nondetachability of *regret* since the concept is slippery. Rather, I turn to recent experimental results in Jayez et al. (2013), which suggest that *regret* is not a ST, even when one sticks to the framework of Abusch. Jayez et al. use French stimuli on the model of (43), which is the English translation of one of the control-target pairs. In the control condition, the speaker considers the PP as probable whereas she is agnostic about it in the target condition. Subjects must rate the sentences on a 7 point scale (1 = extremely obscure, 7 = totally natural). The goal of the experiment is to detect a significant difference between the control and the target conditions. If *regret* is a ST, one expects the scores for the target condition to be significantly worse than those for the control condition. A similar experiment was designed for clefts on the model of (44).

- (43) **Context:** Véronique is wondering whether she will change her current car for a bigger one.
- a. I think that Véronique bought a bigger model. If she regrets it later, it will be difficult to change again. [**control**]
 - b. I wonder whether Véronique bought a bigger model, but, if she regrets it later, it will be difficult to change again. [**target**]
- (44) **Context:** An employee cannot log in on his computer.
- a. I think that someone changed the password. If it was my colleague, I just have to wait to ask him. [**control**]
 - b. I don't know whether someone changed the password but, if it was my colleague, I just have to wait to ask him. [**target**]

Figure 5 shows the (smoothed) density plots of the scores in the control and target conditions for *regret* and *clefts*. The curves for the two conditions are extremely

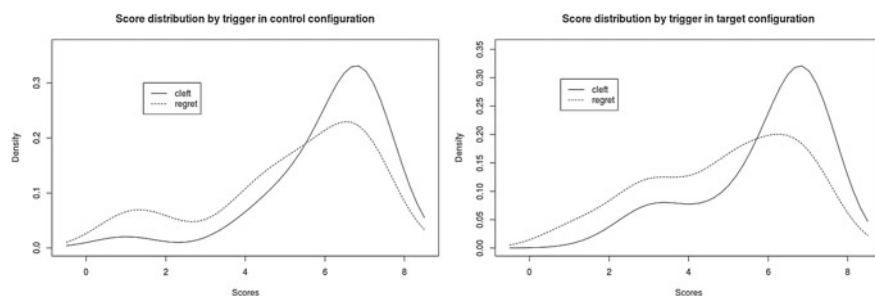


Fig. 5 Density plot by trigger under control and target conditions

similar. The statistical analysis reported in (Jayez et al. 2013, Sect. 1.3.2) failed to detect any difference between the two conditions for *regret* and for *clefts*. Note that clefts are considered by Abusch to be ST. Therefore, the observations are doubly problematic.

These results suggest that *regret* is not intrinsically a ST. Rather, the information it conveys makes it relatively difficult to find a context in which the PP it triggers is suspended. But that does not mean that it is impossible. In other words, with *regret* and other similar triggers, the problem is that of context abduction, the possibility of constructing contexts in which an expression is felicitous. Context abduction depends on the lexical content of the expression and probably the existence of alternatives, as suggested by Abbott. Abbott considers that *regret* and similar emotive factives are nondetachable. Yet, like in the case of *stop*, there are relatively natural paraphrases of the attitude expressed by *regret* that lack factivity. To regret that *p* is to prefer a counterfactual situation in which *p* did not take place. To prefer a situation in which *p* is false, irrespective of the truth-conditional status of *p*, amounts to entertaining the same attitude as that expressed by *regret*, minus the factivity. (45b) does not commit the speaker to a belief that Paul got a Porsche. It is indeed difficult to build very quickly a context in which one would use (45a) instead of (45b) without having the intention to communicate that the PP is true.

- (45) a. Paul regrets to have bought a Porsche
 b. Paul does not like the idea of having a Porsche

Concerning semi-factives like *discover*, Simons (2007) argues that verbs like *discover* or *realize* imply accessing some evidence that the PP is true. It follows that semi-factives are similar to examples like (46). In such cases, we observe exactly the same behavior as with *discover* and its mates, namely: (i) in positive assertions (46a), the conveyed information entails that Paul missed the point and it is not possible to cancel the latter proposition (46b), and (ii), in suspension environments, there is no longer an entailment (46c). With a negation, we have a choice between two interpretations: either we deny the existence of a proof, which amounts to suspending the PP in the case of *discover*, or we deny that Paul is aware of the truth, which amounts to preserving the PP.

- (46) a. Paul has a proof/conclusive evidence that he missed the point.
 b. Paul has a proof/conclusive evidence that he missed the point but ?? he didn't miss it.
 c. If Paul has a proof/conclusive evidence that he missed the point, ...

Generalizing, when the MC entails the PP or makes it highly probable, the very mention of the trigger is sufficient to activate the PP. This accounts for the *win* versus *too* contrast re-illustrated in (47). Winning implies normally participating, so the PP is a consequence of the MC. This also accounts for the case of clefts. A cleft like *It's Paul who solved the problem* entails that Paul solved the problem (MC) and presupposes that someone did. The PP is thus entailed by the MC. In sum, semi-factives, *win* and clefts exhibit plain non-orthogonality and do not call for a special theory of ST or WT.

- (47) a. I don't know if Paul participated in the race, but, if he won, he must be very proud.
 b. ?? I don't know if Paul participated in the race, but if Mary participated too, they probably ran together.

6 Conclusion

In this paper I have used introspective and experimental observations to show that the distinction between strong and weak triggers can be based on the interaction between the lexicon (the separation property, inspired by Abbott) and discourse planning (M-relevance). This interaction can be expressed in a straightforward way in an elementary Bayesian framework. The (rather metaphorical) notion of (non-) orthogonality is replaced by a basic dependency network, which is more in line with a number of observations. Further experimental work, of a different nature, is needed to construct a theory of the time course and the different scenarios of activation/accommodation, in particular in the case of strong triggers.

References

- Abbott, B. (2006). Where are some of the presuppositions gone? In B. J. Birner & G. Ward (Eds.), *Drawing the boundaries of meaning* (pp. 1–20). Amsterdam: John Benjamins.
- Abrusán, M. (2011). Predicting the presuppositions of soft triggers. *Linguistics and Philosophy*, 34, 491–535.
- Abusch, D. (2002). Lexical alternatives as a source of pragmatic presuppositions. In B. Jackson (Ed.), *Proceedings of Semantics And Linguistic Theory XII*. Ithaca: CLC Publications, Cornell University.
- Abusch, D. (2010). Presupposition triggering from alternatives. *Journal of Semantics*, 27, 37–80.
- Beaver, D. (2001). *Assertion and presupposition in dynamic semantics*. Stanford: CSLI Publications.

- Beaver, D. (2004). Have you noticed that your belly button lint colour is related to the colour of your clothing? In R. Bauerle, U. Reyle, and T. E. Zimmerman (Eds.), *Presuppositions and discourse. Essays offered to Hans Kamp* (pp. 65–100). Bingley: Emerald Group Publishing Limited.
- Beaver, D. I., & Clark, B. Z. (2008). *Sense and sensitivity. How focus determines meaning*. Chichester: Wiley-Blackwell.
- Beaver, D. I., & Geurts, B. (2013). Presupposition. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (Fall 2013 Edition)*. <http://plato.stanford.edu/archives/fall2013/entries/presupposition/>.
- Chierchia, G., & McConnell-Ginet, S. (1990). *Meaning and grammar: An introduction to semantics*. Cambridge: MIT Press.
- Cummins, C., Amaral, P., & Katsos, N. (2013). Backgrounding and accommodation of presuppositions: an experimental approach. In E. Chemla, V. Homer, & G. Winterstein (Eds.), *Proceedings of Sinn und Bedeutung 17* (pp. 201–218). <http://semanticsarchive.net/sub2012/>.
- Ducrot, O. (1972). *Dire et ne pas dire*. Paris: Hermann.
- Erteschik-Shir, N. (2007). *Information structure. The syntax-discourse interface*. Oxford: Oxford University Press.
- Geurts, B. (1995). *Presupposing*. Doctoral dissertation, University of Stuttgart.
- Geurts, B. (1999). *Presuppositions and pronouns*. Amsterdam: Elsevier.
- Geurts, B., & van der Sandt, R. (2004). Interpreting focus. *Theoretical Linguistics*, 30, 1–44.
- Grice, P. (1989). *Studies in the way of worlds*. Cambridge: Harvard University Press.
- Gutzmann, D. (2013). Expressives and beyond. An introduction to varieties of use-conditional meaning. In D. Gutzmann & H.-M. Gärtner (Eds.), *Beyond expressives: Explorations in use-conditional meaning* (pp. 1–58). Leiden: Brill.
- Heim, I. (1983). On the projection problem for presuppositions. In M. Barlow, D. Flickinger, & M. Westcoat (Eds.), *Second Annual West Coast Conference On Formal Linguistics* (pp. 114–126).
- Horn, L. R. (1972). *On the semantic properties of logical operators in English*. UCLA. Ph.D. dissertation.
- Horn, L. R. (2001). *A natural history of negation*. Stanford: CSLI Publications.
- Jayez, J. (2010). Projective meaning and attachment. In M. Aloni, H. Bastiaanse, T. de Jager, & K. Schulz (Eds.), *Logic, language and meaning. Revised selected papers of the 17th Amsterdam colloquium, Amsterdam 2009. LNAI*, Vol. 6042 (pp. 325–334). Berlin: Springer.
- Jayez, J. (2014). A note on presuppositions, discourse attachment and relevance. In J. Blochowiak, C. Griset, S. Durrlemann-Tame, & C. Laenzlinger (Eds.), *Collection of papers dedicated to Jacques Moeschler*. <http://www.unige.ch/lettres/linguistique/moeschler/Festschrift/Jayez2014.pdf>.
- Jayez, J., & Mongelli, V. (2013). How hard are hard triggers? In E. Chemla, V. Homer, & G. Winterstein (Eds.), *Proceedings of Sinn und Bedeutung 17* (pp. 307–324). <http://semanticsarchive.net/sub2012/>.
- Jayez, J., Mongelli, V., Reboul, A., & van der Henst, J.-B. (2013). Weak and strong triggers. In F. Schwarz (Ed.), *Experimental perspectives on presuppositions. Studies in Theoretical Psycholinguistics Series*. Springer (in press). <http://perso.ens-lyon.fr/jacques.jayez/doc/weak-strong-triggers.pdf>.
- Karttunen, L. (1971). Some observations on factivity. *Papers in Linguistics*, 4, 55–69.
- Kauschke, C., & Hofmeister, C. (2002). Early lexical development in German: A study on vocabulary growth and vocabulary composition during the second and third year of life. *Journal of Child Language*, 29, 735–757.
- Kripke, S. A. (2009). Presuppositions and anaphora: Remarks on the formulation of the projection problem. *Linguistic Inquiry*, 40(3), 367–386.
- Lassiter, D. (2012). Presuppositions, provisos and probabilities. *Semantics and Pragmatics*, 5(2), 1–37.
- Merin, A. (1999). Information, relevance, and social decisionmaking: Some principles and results of decision-theoretic semantics. In L. S. Moss, J. Ginzburg, & M. de Rijke (Eds.), *Logic, language and computation* (Vol. 2, pp. 179–221). Stanford: CSLI Publications.

- Merin, A. (2003). Presuppositions and practical reason. A study in decision-theoretical semantics. *Forschungsberichte der DFG-Forscherguppe in der Philosophie* 114. University of Konstanz.
- Partee, B. (1991). Topic, focus and quantification. In *Proceedings of SALT I* (pp. 257–280).
- Pearl, J. (2009). *Causality. Models, reasoning and inference*. Cambridge: Cambridge University Press.
- Potts, C. (2005). *The logic of conventional implicature*. Oxford: Oxford University Press.
- Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics*, 1, 75–116.
- Simons, M. (2007). Observations on embedding verbs, evidentiality, and presuppositions. *Lingua*, 117, 1034–1056.
- Simons, M., Tonhauser, J., Beaver, D., & Roberts, C. (2011). What projects and why. In N. Li & D. Lutz (Eds.), *Semantics and Linguistic Theory (SALT) 20*, (pp. 309–327) eLanguage.
- Sletsjøe, L. (1979). Le préfixe RE- en latin et dans les langues romanes occidentales. *Studia Neophilologica*, 51, 85–113.
- Stalnaker, R. (1974). Pragmatic presuppositions. In M. K. Munitz & P. Unger (Eds.), *Semantics and philosophy* (pp. 197–214). New York: New York University Press.
- van Benthem, J., Girard, P., & Roy, O. (2009). Everything else being equal: A modal logic for ceteris paribus preferences. *Journal of Philosophical Logic*, 38, 83–125.
- Zeevat, H. (1992). Presupposition and accommodation in update semantics. *Journal of Semantics*, 9, 379–412.

Layered Meanings and Bayesian Argumentation: The Case of Exclusives

Grégoire Winterstein

Abstract This work deals with the notion of argumentation and the way it is treated in a Bayesian perspective on reasoning and natural language interpretation. I am especially interested in the way the linguistic structure of utterances affect the way they are interpreted. Specifically, I argue that the sum of meanings conveyed by an utterance cannot be fed into such models indiscriminately and that the at-issue content of an utterance prevails in its argumentative interpretation. I support my claim by focusing on the case of the adverb *only* and by presenting the results of experiments which show that utterances that have similar informational contents do not behave the same way when they are argumentatively evaluated.

Keywords Argumentation · Bayesian reasoning · Exclusives · Argument from ignorance

This work investigates how the way information is conveyed affects the way speakers reason about natural language utterances. More specifically, I am interested in the way the argumentative properties of a sentence depend on the way information is packaged in an argument. To carry out this project, I focus on the case of exclusive particles, and more specifically on the case of the adverb *only*.

My study is rooted in Bayesian models of natural language interpretation and reasoning that come both from linguistics and the psychology of reasoning. My main claim is that the information which is used for the conditional update used in these approaches ignores what is not the main, or at-issue, content of the argument. I back-up this claim by experimental means, showing how the use of *only* changes the properties of an argument, even though its informational content is unchanged.

This research was supported in part by the Erasmus Mundus Action 2 program MULTI of the European Union, grant agreement number 2010-5094-7.

G. Winterstein (✉)
Aix Marseille Université, Laboratoire Parole Et Langage and Nanyang
Technological University, Singapore, Singapore
e-mail: gregoire.winterstein@linguist.univ-paris-diderot.fr

© Springer International Publishing Switzerland 2015
H. Zeevat and H.-C. Schmitz (eds.), *Bayesian Natural Language Semantics
and Pragmatics*, Language, Cognition, and Mind 2,
DOI 10.1007/978-3-319-17064-0_8

This also serves as a way to establish what the best paraphrase of *only* is, at least in argumentative terms.

I begin by presenting the framework of argumentation, focusing on its Bayesian interpretation (Sect. 1). In Sect. 2 the case of exclusives is presented. I introduce their semantics, and show that Bayesian approaches to reasoning make predictions that conflict with intuition. I test those predictions in Sect. 3 and discuss the results and potential implications in Sect. 3.3.

1 Argumentation

Argumentation has been defined in various ways according to the discipline in which it has been studied, be it in philosophy, beginning with the works of Aristotle, in computer science and artificial intelligence (Besnard and Hunter 2008), in cognitive science (Hahn and Oaksford 2006; Zenker 2013) or in linguistics (Anscombe and Ducrot 1983). Nevertheless, all these accounts attribute some common core properties to argumentation:

- Argumentation is an activity aimed at *persuasion*.
- Argumentation starts with a set of premises and uses some mechanism (e.g. deduction) to arrive at a *conclusion* (sometimes called *goal*).
- Argumentation is *defeasible*: an argument can be undercut, superseded, rebutted etc. by another, more persuasive argument.

One point of dissension for the various takes on argumentation is the question of what should be considered as a good argumentation. Some accounts focus on defining *validity* criteria, following the trend fostered by Aristotle. Others consider that an argument is good as soon as it yields some persuasive power, even if it is logically invalid. This approach is especially adopted in the cognitive studies of argumentation.

In this work, I will mainly consider argumentation from the cognitive point of view, more specifically adopting the Bayesian view on argumentation as proposed by Oaksford and Hahn (2004); Hahn and Oaksford (2006, 2007). I will also relate this perspective on argumentation to the linguistic perspective fostered by Anscombe and Ducrot (1983) and later developed in Bayesian terms by Merin (1999). Section 1.1 gives an introduction to the view of argumentation in the Bayesian reasoning perspective, and Sect. 1.2 presents the argumentative approach in linguistics.

1.1 Bayesian Reasoning and Argumentation

Bayesian models of reasoning have been successful for some time in the domain of the psychology of reasoning (Oaksford and Chater 2007). This framework can be rather straightforwardly applied to the case of argumentation. This has notoriously

been done by Hahn and Oaksford (2006, 2007) (henceforth H&O, for a review of recent works in that perspective, see also Zenker 2013).

The basic tenet of this approach is to consider that argumentation is linked to degrees of belief. A good argument is one that increases the degree of belief in the targeted conclusion, while a counter argument is one that decreases this belief. Thus, the strength of an argument is measured by its ability to affect a degree of belief. In the traditional Bayesian fashion, degrees of belief are equated with probabilities, and the effect of an argument is measured by looking at the conditional update of the probability function based on the contents of the argument.

For example, in (1), the targeted conclusion is $T = \textit{Ted is a priest}$ and the argument is $C = \textit{he wears a cassock}$.

(1) I think Ted is a priest because he wears a cassock.

Supposing that the addressee had no prior conception on Ted being a priest, then his degree of belief in T is rather low because priests are not very numerous (e.g. Wikipedia mentions that there were 20,277 priests in France in 2007, for a total population of about 64 million people). In other terms, we have $P(T) \ll 1$: the belief of the addressee in T is very far from certainty.

To measure the effect of the argument C on the belief of T , i.e. to compare $P(T)$ to $P(T|C)$, we use Bayes' rule and get:

$$(2) P(T|C) = \frac{P(C|T) \cdot P(T)}{P(C)}$$

To compare $P(T|C)$ to $P(T)$, we can subtract them and we get:

$$(3) P(T|C) - P(T) = P(T) \left(\frac{P(C|T)}{P(C)} - 1 \right)$$

This means that $P(T|C)$ is higher than $P(T)$ iff $\frac{P(C|T)}{P(C)}$ is higher than 1. This is actually expected since world-knowledge tells us that $P(C|T)$ is rather high, i.e. the chances of Ted wearing a cassock when we know he is a priest are quite high, and certainly higher than the chances of Ted wearing a cassock without knowing anything else about him. So the argument C is helping to conclude T because the knowledge of C increases the belief in T .

This approach has successfully been used by H&O to show how so-called cases of fallacies can be persuasive, even though they should be considered to be logically invalid. In Sect. 2.2 I develop their account of the case of the argument from ignorance, but their framework has also been successfully applied to other cases of fallacies (e.g. the circular argument and the slippery slope argument) and also covers effects that can be attributed to the source of the argument rather than its form or content.

1.2 Argumentation in Linguistics

Initially, the argumentative perspective on discourse semantics was proposed and developed by Ducrot and Anscombe (see e.g. Anscombe and Ducrot 1983 for an

introduction). The main intuition behind this approach is that natural language interpretation is not reducible to matters of truth conditions, but should also include an argumentative component. A basic observation is the contrast in (4).

- (4) a. John should be reprimanded: he was **barely** on time for an important meeting.
 b. #John should be reprimanded: he was **almost** on time for an important meeting.

From a truth conditional point of view, John was on time for the meeting in (4-a) while he was not in (4-b). Nevertheless, (4-a) can be used as an argument for reprimanding John but (4-b) cannot. Anscombe and Ducrot postulate that this contrast is due to the respective argumentative effects of *barely* and *almost* which are independent of their truth conditional meaning postulates.

Recent work in that domain took a slightly different perspective by assuming that the argumentative effect of an item can be derived from its truth conditional component and the way it conveys information (Jayez and Tovena 2008; Merin 1999). These works use a Bayesian approach similar in spirit to the one used for Bayesian reasoning and explicitly relate the strength of an argument to its effect on the degree of belief in the conclusion after a conditional update.

Among the operators affecting the argumentative properties, one can define the class of invertors: these are the elements that switch the argumentative orientation of their host sentence. In other terms if p is used to argue for C and if Op is such an operator, then $Op(p)$ is an argument for the opposite conclusion, i.e. for $\neg C$. The most obvious element of this class is negation, as can be seen in the contrast in (5).

- (5) a. This ring is nice. \rightsquigarrow_{arg} Buy the ring.
 b. This ring is not nice. \rightsquigarrow_{arg} Do not buy the ring.

Other elements include adverbs like *barely*, *hardly* and *only* (Ducrot 1973). In this work, I will focus on the adverb *only*. The main feature that interests us is that, unlike negation, *only* apparently acts as an invertor while conveying the same informational content as its prejacent (i.e. its host). Combined with the approach of argumentation based on Bayesian reasoning, this leads to some conflicting predictions between the linguistic description of *only* and the predictions of the Bayesian reasoning approach.

In the next section, I explore these issues in more detail by showing how they are exemplified in the case of the *argument from ignorance*.

2 Exclusives and the Case of the Argument from Ignorance

2.1 The Semantics of *only*

2.1.1 The Basic Meaning of *only*

The adverb *only* has been the subject of numerous studies. Coppock and Beaver (2013) claim that one can number at least 27 distinct lexical entries attributed to this

element. It is beyond the scope of this article to do justice to every analysis proposed of *only*. I will thus begin by presenting rather non-controversial facts about *only* and then present a sample of the formal proposals proposed to capture these facts before proceeding to present the argumentative properties of *only*.

The central observation about a sentence using *only* is that it conveys two components: a positive one and a negative one. This is illustrated by the means of (6).

- (6) Robert only plays the bass.
 a. \rightsquigarrow Robert plays the bass.
 b. \rightsquigarrow Robert plays no other instrument than the bass.

The positive component of (6) is (6-a) and its contents seem to match the contents of the host sentence of *only* (also called prejacent). The negative component of (6) is (6-b) and it is this component that seem to convey the exclusive meaning giving its name to the class to which *only* belongs. To properly determine the contents of the main content it is necessary to identify the *associate* of *only*, i.e. the constituent which determines the set of alternative elements that will be excluded by the assertion. Information structure is usually assumed to play a key role in the identification of the associate (Krifka 1999; Rooth 1992).

The positive and negative components do not have the same assertoric status, and this has been the subject of many publications about *only*. The central observation is that negation does not affect these components in the same way: (7).

- (7) Robert does not only play the bass.
 a. \rightsquigarrow Robert plays the bass.
 b. $\not\rightsquigarrow$ Robert plays no other instrument than the bass.

While the positive component is still conveyed after embedding the sentence under a negation, this is not true for the negative one. This has been taken to indicate that the prejacent is presupposed by *only*, although this has been the matter of dispute (cf. below). The main point here is that, irrespective of their exact nature, the positive and negative components do not behave in the same way under negation, which is taken as an indication that they do not have the same informational status.

2.1.2 Formal Analyses of *only*

Accounts of the semantics of *only* vary on several points. The three main ones are the following:

- The informational status attributed to the positive and negative components.
- The scalar nature of *only*'s associate.
- The actual form of the positive and negative components.

I will have little to say on the debate about the informational status of the components. All analyses agree on the fact that the negative component corresponds to the asserted part of the utterance. This is sometimes referred as the main content or the at-issue content (see e.g. Simons et al. 2010; Jayez 2010 for considerations

on the notion of being *at-issue*). In this paper these terms have to be understood as being mostly equivalent although the terms *main content* will be preferred as a noncommittal one alongside with *at-issue*.

The status of the positive component however divides the community. It has been variously considered as a presupposition of various sorts (Rooth 1992; Krifka 1999; Geurts and van der Sandt 2004), a conversational implicature (van Rooij and Schulz 2004), a complex meaning arising from the interplay of a presupposition and a conversational implicature (Ippolito 2008) or a presupposition with a double weak/strong status (Zeevat 2009). Atlas (1993) and Horn (2002) are exceptions in considering that it can be a standard entailment of the utterance. For my purpose, I will assume that the prejacent is not a main content. This will be sufficient to test the hypothesis that nothing besides the main content of *only* is considered in the evaluation of an argument.

The question of the scalarity of *only*'s associate is related to a two-sided hypothesis. Firstly, some analyses assume that the associate of *only* belongs to a scale, i.e. an ordered set of alternatives. Secondly, the associate is supposed to be low on this scale, or at the very least not to occupy the top of the scale (Beyssade 2010; Coppock and Beaver 2013). This hypothesis helps to explain why sentences such as (8) appear degraded because even though *only*'s associate does belong to a scale, it is the top element of that scale (assuming that one knows that a royal flush is the highest possible hand in poker).

(8) # I only have a royal flush.

However, in cases such as (6), the scalar nature of the associate is less evident. One can of course assume that the alternatives to *playing the bass* are ordered, but it is not *necessary* to postulate the scalar nature of the associate to interpret the sentence. It is enough to state that all possible alternatives are false. Furthermore, one can also find examples where the associate of *only* occupies the top of a scale (Winterstein 2012). Nevertheless, it is not the point of the paper to discuss the scalarity hypothesis and I will eschew the question by only considering sentences that involve quantities and are therefore obviously scalar. This entails that the experimental results presented in Sect. 3 are only valid for the scalarity case. The extension to non-scalar cases is briefly addressed in Sect. 3.3, but a thorough generalization would require further work and experiments. Nevertheless, it should be noted that the cases that involve a non-controversially scalar associate of *only* are numerous enough to justify focusing on those in this paper.

The last point of contention is the exact form of the positive and negative component. The paraphrases given in (6) roughly correspond to what has first been proposed by Horn (1969) and adopted by many after him. Zeevat (2009)'s proposal is similar in spirit and postulates that the exclusive component conveys that there must exist an element alternative to the associate that could be subject to the same predication but is not. Coppock and Beaver (2013) follow Beaver and Clark (2008)'s proposal by assuming that *only* is scalar and that it makes an *at-issue* contribution that indicates that the prejacent is the strongest answer to the current question under discussion (CQ). Inversely, *only* presupposes that the prejacent is the weakest answer to the CQ.

As already said, I do not want to take an explicit stance on the question of the scalarity of *only*. However, since all the examples that I will use involve obviously scalar items (for reasons explained below), I will assume that in the cases at hand, the descriptions given by Coppock and Beaver are mostly correct and that *at most* or *no more than* are likely candidates to paraphrase the asserted content conveyed by *only*. A sufficient reason for this choice is that both appear natural when the associate of *only* is scalar, as in (9-a) and (9-b).¹

- (9) Lemmy drank only half the beer in the fridge.
 a. Lemmy drank at most half the beer in the fridge.
 b. Lemmy did not drink more than half the beer in the fridge.

The question of whether *at most* or *not more than* is the most adequate will be addressed in Sect. 3.

2.1.3 The Argumentative Effect of *only*

The use of *only* in a sentence has argumentative effects. For example even though having a master's degree can be used as an argument for being competent (10-a), modifying this information with *only* yields a counter argument (10-b).

- (10) a. Ritchie is quite competent, he has a master's degree.
 b. # Ritchie is quite competent, he only has a master's degree.

Based on the observation of examples like (10) (actually using French *seulement* which I assume to be roughly equivalent to *only*) Ducrot (1973) describes the adverb as having an argument reversal effect. If *p*, the prejacent of *only* argues in favor of *H* then "*only p*" will argue in favor $\neg H$. This places *only* in the class of invertors mentioned previously.

Unlike negation, this behavior of *only* is unexpected under the Bayesian perspective of argumentation detailed in Sect. 1.1. This is because when using *only* the meaning of the prejacent is still conveyed by the argument. It is even possible to construct utterances that do not differ in terms of informational content. For example, let's take the case of (11).

- (11) a. Half the studies showed that drug X has undesirable side-effects.
 b. Only half the studies showed that drug X has undesirable side-effects.

Intuitively, (11-a) should be heeded as a strong warning against using drug X, i.e. it argues against using drug X. The question is whether this is still true of (11-b). Ducrot's linguistic take on argumentation predicts that (11-b) should argue in the opposite direction of (11-a), meaning that it should be an argument in favor of using drug X.

¹This contrasts with situations where a scalar paraphrase might appear disputable, e.g. as in (21-a) below. In such cases the question of the scalarity of *only* becomes relevant, cf. (Winterstein 2012). As already stated we leave those cases outside the scope of this study.

On the other hand, the Bayesian reasoning approach defended by Hahn and Oaksford (2006, 2007) makes the opposite prediction. This is because both versions of (11) convey the same informational content, namely that half the studies showed undesirable side-effects and thus the other half showed no such effects.² Therefore the assertion of both versions of (11) should have the same effects since they will lead to the same conditional update of the probability function. This however seems to go against intuition. One way to reconcile the predictions of Ducrot's theory with the Bayesian account is to consider that when evaluating an argument like (11-b) not all the material conveyed by the speaker is taken into account. More precisely, the hypothesis is that only the at-issue content is considered. In the case of (11-b), this at-issue content is limited to the negative component conveyed by *only* and this can explain the different argumentative profiles. This hypothesis is experimentally tested in Sect. 3. However before moving on to the results of these experiments, I show that the Bayesian reasoning approach makes a further prediction on the orientation of both versions of (11).

2.2 The Argument from Ignorance

Besides predicting that both versions of (11) have similar effects, the Bayesian reasoning approach also predicts that the positive argument should win against the negative one. This entails that both versions of (11) should argue **against** using the drug X. This can be explained by observing that the arguments in (11) involve a form of argumentation known as the *argument of ignorance* which I introduce here.

Assuming the description of *only* that I adopted in Sect. 2.1, a sentence like (11-b) conveys two components, given in (12).

- (12) a. Half the studies showed that drug X has undesirable side-effects.
 b. Half the studies showed no undesirable side-effects associated with drug X.

In the argumentation tradition (12-b) is called an *argument from ignorance*. The *argumentum ad ignorantiam*, or *argument from ignorance*, is considered to be a

²Of course, it can be argued that this falls out only by adopting a strong reading of (11-a), i.e. one where *half* is interpreted in an exact manner. Some accounts claim that this is the result of a quantity implicature, which is notably defeasible and should not be integrated in the informational content of an utterance. However, the case of numerals and quantities like *half* is known to be peculiar in this regard. Typically, (i) could denote both a situation where the tank is more than half empty or less than half empty, meaning that *half* easily keeps its exact reading under negation, something which is harder for elements like *some*. I will therefore assume that the argument in (11-a) is done with an exact reading of *half*.

(i) The tank is not half empty.

It should however be noted that ultimately I argue for the fact that an argument is evaluated only in the light of its asserted content (see Sect. 3.3). Therefore, in the case of (11-a) the relevant reading for evaluating the argument will indeed be a lower-bounded one. Nevertheless, the utterance can still be considered as conveying an overall exact reading.

classical fallacy (Hamblin 1970). One of the common paraphrase of this argument is that “*absence of evidence means evidence of absence*”. More precisely, Walton (1996) gives the following characterization of this type of argument:

- (13) a. If *A* were true (false), it would be known (proved, presumed) to be true (false).
 b. *A* is not known (proved, presumed) to be true (false).
 c. Therefore, *A* is (presumably) false (true).

Arguments from ignorance are usually treated as fallacies because the inference to the falsity of *A* is logically valid only under the condition that the first premise (13-a) is true. This is often far from true as most tests used to demonstrate the truth of a conclusion are not perfect. However these arguments are routinely used in certain cases, most of all in law. As an example, French law will find a party guilty of damaging a plaintiff if and only if three elements can be established: the prejudice, the party’s responsibility and the causal link between the party’s behavior and the prejudice. Among the three, the causal link is usually the hardest to prove because it is difficult to prove that the prejudice would not have occurred had the accused party behaved differently (Pearl 2009). As an illustration, media covered several cases of people becoming sick after taking a faulty drug but getting no compensation because the link between the faulty drug and their illness could not be properly established. The defense argument was that there was no way to prove the patients would not have gotten sick anyway. Therefore, because of an absence of evidence for the causal link, the court ruled as if there were none and did not convict the drug company. This is a proper form of argument from ignorance since the absence of proof was considered as proof of absence.

In their work, Hahn and Oaksford (2006, 2007) propose an explanation for the fact that so-called fallacies sometimes appear unconvincing, while sometimes they appear sound enough to form the basis of legal argumentation. Their proposal is that the form of an argument itself is not indicative of its ability to convince, but that its content is what matters. They propose a Bayesian analysis of argumentation that captures this. Their account follows the three-fold classification of arguments from ignorance proposed by Walton (1996): *negative evidence*, *epistemic closure* and *shifting the burden of proof*. I will focus here on the case of negative evidence, illustrated in (14).

- (14) a. If drug *A* were toxic, it would produce toxic effects in legitimate tests.
 b. Drug *A* has not produced toxic effects in such tests.
 c. Therefore, *A* is not toxic.

The negative evidence refers to the absence of a test that proved toxic effects associated with drug *A*. Positive evidence would refer to the opposite situation, i.e. one where a positive test would be used as an argument for the toxicity of *A*. Recall that the fallacious character of the argument lies in the premise (14-a); this is because tests (even clinical ones) are not foolproof. H&O are interested in producing a model that embodies this characteristic as well as the following properties of negative evidence, which have all been experimentally tested and validated:

1. Negative arguments can be convincing in the right context: (15) sounds like a convincing argument.

(15) Drug *A* is not toxic: a recent study showed no undesirable side-effects associated with it.

2. Generally, negative arguments are less convincing than their positive counterparts: (16) is better than (15) (for the opposite conclusion).

(16) Drug *A* is toxic: a recent study showed undesirable side-effects.

The last property has an important consequence for the case of (11). It entails that positive evidence should prevail over negative evidence, and therefore that both (11-a) and (11-b) should argue in favor of the toxicity of drug *X*. The key point to arrive at this result is to assume that in both versions of (11) the total sum of the meanings conveyed is taken into account. If this is not the case, the two are no longer predicted to be equivalent. This point will be at the center of the experiments described in Sect. 3.

2.3 The Bayesian Treatment of the Argument from Ignorance

To show how a Bayesian model captures the last two facts, H&O model the force of positive and negative evidence in the following manner. First, let's define the following notation for the case of (15)–(16):

- e : observation of a toxic effect
- $\neg e$: no observation of a toxic effect
- T : *Drug A is toxic*
- $\neg T$: *Drug X is not toxic*

The effects of positive evidence and negative evidence are then respectively measured by observing $P(T|e)$ and $P(\neg T|\neg e)$, i.e. the probabilities that the drug is toxic (non toxic) knowing that there was (was not) an observation of a toxic effect in a legitimate test.

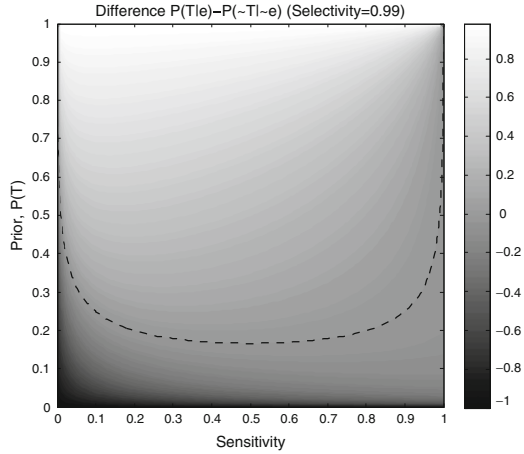
To access these two quantities, one can use Bayes's theorem and two quantities that routinely define clinical tests: *sensitivity* and *selectivity* (also called *specificity* by some):

Sensitivity (n) corresponds to the probability of observing a toxic effect knowing that the drug is toxic, i.e. $n = P(e|T)$. A low sensitivity increases the chances of having false negatives, i.e. negative tests in the presence of a toxic drug.

Selectivity (l) is the probability of observing a negative test when the drug is not toxic, i.e. $l = P(\neg e|\neg T)$. It captures the rate of false positives.

Although these quantities are usually defined in the context of clinical tests, they apply to any type of situation where an observation is used to infer some property.

Fig. 1 Difference between positive and negative evidence versus prior and sensitivity in the highly selective case



For example, owning a hammer is a clue for being a carpenter. It is highly selective: if one does not own a hammer, then the chances that he is a carpenter are very low. However it has a rather low sensitivity: many people besides carpenters own a hammer. Clinical tests are usually optimized in terms of sensitivity and selectivity in order to minimize the rates of false positives and false negatives (whilst keeping in mind that, usually, acting on one influences the other, e.g. lowering the false positives is likely to raise the false negatives).

Next, to access the quantities $P(T|e)$ and $P(\neg T|\neg e)$ we rely on Bayes’s theorem³:

$$(17) \text{ a. } P(T|e) = \frac{n \cdot P(T)}{n \cdot P(T) + (1-l) \cdot P(\neg T)}$$

$$\text{ b. } P(\neg T|\neg e) = \frac{l \cdot P(\neg T)}{l \cdot P(\neg T) + (1-n) \cdot P(T)}$$

H&O consider that positive arguments are better than negative ones iff $P(T|e) > P(\neg T|\neg e)$, i.e. if the probability of toxicity given a positive observation is higher than the probability of non-toxicity given a negative observation. Given (17), this inequality translates to $P(T)^2(n - n^2) > P(\neg T)^2(l - l^2)$. In the case agents are neutral regarding T , i.e. when $P(T) = P(\neg T)$, then the condition reduces to $l > n$, i.e. to selectivity being higher than sensitivity.

More generally, H&O claim that

... as long as [selectivity] is sufficiently high, $P(T|e)$ is greater than $P(\neg T|\neg e)$ over a large range of values of [sensitivity] even when [the prior] is low.

Figure 1 shows the difference between $P(T|e)$ and $P(\neg T|\neg e)$ in function of sensitivity and the prior h with a high fixed value of selectivity. As can be seen, most of the regions on the graph correspond to a positive difference between $P(T|e)$ and $P(\neg T|\neg e)$.

³Here recall that $n = P(e|T)$ and $l = P(\neg e|\neg T)$. In (17-a), the denominator $P(e)$ is expressed as the sum $P(e|\neg T)P(\neg T) + P(e|T) \cdot P(T)$ via the product rule. The same operation gives $P(\neg e)$ in (17-b).

H&O claim that having a very high selectivity is often met in practice (Oaksford and Hahn 2004), which then ensures that positive evidence prevails over negative evidence as desired.

2.4 Taking Stock

Throughout this section I presented various observations and hypotheses about the interpretation of *only* that are potentially contradictory:

- In sentences like (11), the use of *only* does not affect the truth-conditional content conveyed by the speaker: he conveys an equal amount of positive and negative evidence.
- Under a standard Bayesian reasoning approach, two arguments that share their truth-conditional content do not have different argumentative profiles.
- In the case at hand, the Bayesian reasoning makes the prediction that positive evidence prevails over negative evidence.
- The Bayesian reasoning approach makes predictions that are potentially in line with Ducrot's approach (i.e. by predicting that *only* is an inverter) if we assume that among the material conveyed by an utterance, only a subset of the content is factored in when evaluating an argument.

I put these observations in perspective with the main hypothesis we would like to defend, i.e. that an argument is evaluated by *only* considering the main content it conveys. The next section addresses those issues by presenting the results of several experiments.

3 Experiments

Two hypotheses come out of the discussion in the previous section. First, one wants to test the prediction of Ducrot regarding the inverter status of *only*. While intuitive enough, this hypothesis deserves testing. Second, assuming that *only* is an inverter, it is necessary to test whether its argumentative profile is indeed equivalent to its main or at-issue content alone (an hypothesis in line with previous assumptions (Jayez and Tovena 2008) and one that has been suggested in early works on argumentation but never properly investigated). Experiments were therefore run in order to test these two complementary hypotheses, summarized as follows:

Inversion Hypothesis : adding *only* to an argument inverts its argumentative properties, even when the content conveyed by the argument appears unchanged by the addition of *only*, and when Bayesian reasoning predicts one specific argumentative profile.

Main Content Hypothesis : the main content, or at-issue content, of an argument is what determines its argumentative profile.

The inversion hypothesis is related to the behaviour of *only*, the main content one is explanatory and is intended as a way to explain the inversion hypothesis should the experimental results confirm it. The details of the experimental protocol are presented in Sect. 3.1, results are given in Sect. 3.2 and discussed in Sect. 3.3.

3.1 Experimental Settings

3.1.1 Experimental Factors

The two hypotheses under scrutiny were tested by running four distinct experiments.

In order to test the inversion hypothesis, I compared arguments involving *only* to the same arguments without *only*. For testing the main content hypothesis, I compared arguments with *only* with arguments solely expressing the main content of *only*. Since multiple ways to quote the main content of *only* have been proposed I tested three possible paraphrases for it: *less than*, *at most* and *no more than*. The last two are notably used by Coppock and Beaver (2013), who mention on several occasions that *at most* is probably the most appropriate. The first one was deemed to be a suitable candidate as well for its proximity to the others in semantic terms and was also involved in the experiments. Each of these three options led to a different experiment.

Each experiment used one independent factor, called *Modification* which represents the way the argument is modified. *Modification* had three levels in each experiment. Each experiments involved the following two levels:

- *Bare*: no modification of the argument.
- *Only*: modification of the argument by adding *only*.

The third level of *Modification* differentiates the first three experiments. The values were respectively *Less than*, *At most* and *No more than* depending on the paraphrase of the main content of *only* that was tested. For each experiment I then compared:

- The responses for the *Only* and *Bare* levels to test the inversion hypothesis. If they are significantly different, it will support the hypothesis.
- The responses for the *Only* and *Less than/At most/No more than* to test the main content hypothesis and find the most appropriate paraphrase. A paraphrase will be judged appropriate if the way the argument is evaluated is not significantly different than the version with *only*.

Based on the outcome of the first three experiments, a fourth experiment was planned in order to strengthen the observations concerning the inversion hypothesis. It only used the *No more than* level for *Modification* and differed in the materials presented to the subjects (cf. below).

3.1.2 Experimental Protocol

All the experiments followed the same format, inspired by the one used by H&O in their own experiments.

Subjects had to answer an online questionnaire presenting paragraphs of texts followed by a question asking the subjects to rate at which degree one of the character presented in the paragraph was convinced by the argument presented in the paragraph. (18) is an example of one of the items used in one of the experiments.

- (18) a. Barbara is wondering whether she should take digesterol for her stomach troubles. Her friend Sophia tells her that half of the existing medical studies have found that digesterol has undesirable side effects.
 b. How strongly convinced do you think Barbara is that digesterol is dangerous given what Sophia tells her?

The answer to the question was given on a Likert scale (1 = *Not convinced at all*, 10 = *Entirely convinced*) and was the measured dependent factor in each experiment (called *Persuasiveness*).

For the first three experiments, the argumentative goal p that subjects had to evaluate was such that the *Bare* version was an argument for p and such that *Only* and *No more than* (and its alternatives) were arguments for $\neg p$. In the fourth experiment, this was inverted: the goal that was to be evaluated was supported by *Only* whereas *Bare* would argue against it. For example in the case of (18), the question was changed to (19).

- (19) How strongly convinced do you think Barbara is that digesterol is not dangerous given what Sophia tells her?

In total, each experiment presented 15 experimental items to the subjects along with 30 filler items not involving the tested factors. Target items all included a sentence of the form *half of the X (did) Y ...* where $X \text{ did } Y$ is taken to be a positive argument for the proposition used in the question (except in experiment 4), and taken to have the intuitively desired properties of sensitivity and selectivity (e.g. they referred to some scientific study). The only quantity used in the experiment was *half* so as to ensure there would be an equal amount of positive and negative evidence presented to the subjects (and thus ensure the predictions of the Bayesian reasoning approach).

The items were presented in pseudo-random order and using a latin square design to ensure a proper distribution of the target items and fillers. This means that each subject saw five instances of each level of *Modification* and that no single subject saw two versions of the same sentence instantiated with different levels of *Modification*. The latin square design ensured that all versions of each sentence were tested so that the scores associated with each level of the *Modification* variable represent an average of the scores for the 15 target sentences. The target sentences were identical in all three experiments except for the value instantiated by the third value of the *Modification* variable and the formulation of the question in the fourth experiment.

The final questionnaire was created using the *Ibex* tool and hosted on the *Ibex Farm* platform. Each questionnaire was answered by 33 participants recruited on the crowdsourcing *Crowdfunder* platform and who were paid 0.60\$ for their participation. All the participants were (self-declared) native speakers of English and were distinct (no participant answered more than one of the surveys).

3.2 Results

The figures in Fig. 2 summarize the results of the experiments. Each figure shows the results of one experiment by comparing the values of the dependent factor *Persuasiveness* (translated on a scale between 1 and 100) for each of the levels of the *Modification* factor.

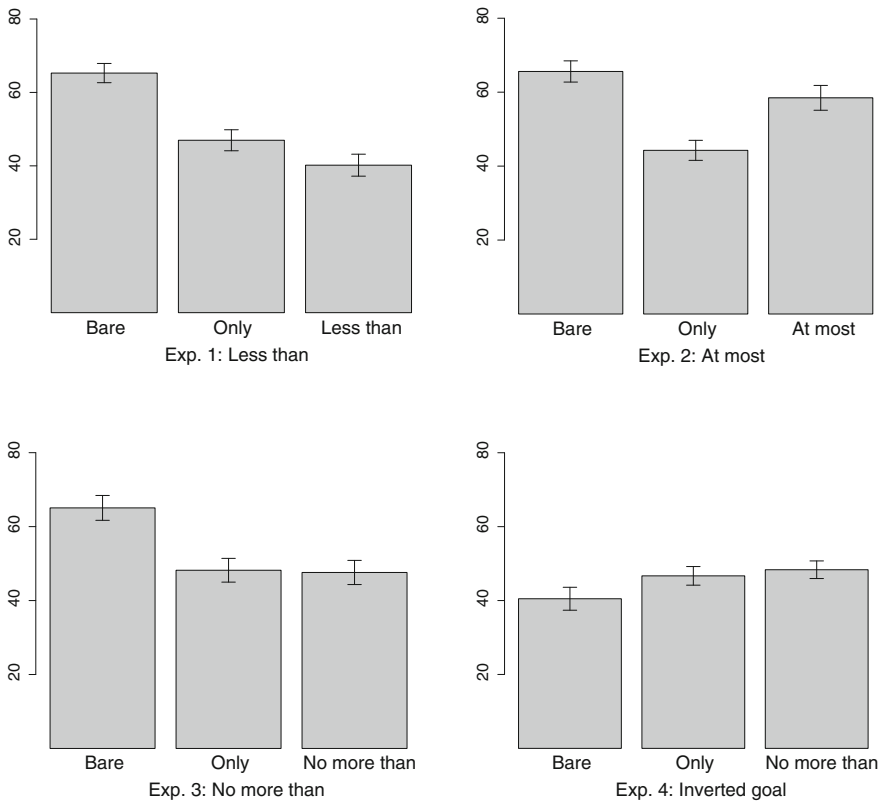


Fig. 2 Persuasiveness versus Modification, four experiments

3.2.1 Inversion Hypothesis

Each figure in Fig. 2 shows a difference between the `Bare` case on one hand and the `Only` case on the other hand. In the first three experiments the difference is very clear whereas it seems a bit attenuated in the fourth one. The significance of the differences have been tested by using paired Wilcoxon tests and all are highly significant, including the fourth experiment (*Less than* experiment: $V = 509, p = 4.751e - 06$; *At most* experiment: $V = 431, p = 4.062e - 06$; *No more than* experiment: $V = 408.5, p = 3.713e - 05$; *Inverted goal* experiment: $V = 125.5, p = 0.001126$).

The influence of the experiment was assessed using model reduction and likelihood ratio test in order to find whether the judgments for the `Only` and `Bare` cases were affected by the nature of the third level in the first three experiments. No effect of the experiment on the values of `Only` and `Bare` was found ($\chi^2 = 1.3282, p = 0.5147$), confirming the stability of the judgments over the experiments and further reinforcing the conclusion that the difference between the `Only` and `Bare` cases is significant.

Furthermore, if we add the scores of `Bare` and `Only` over all experiments (including the fourth one), the totals come close to 100, meaning that the values are complementary. The inversion hypothesis therefore appears confirmed: whenever *only* is present it affects the way the argument is evaluated and produces an argument whose strength depends on that of the bare version (cf. experiment 3 versus 4). This is discussed in the relevant section below.

3.2.2 Main Content Hypothesis

Table 1 summarizes the differences (and their significance) between the `Only` case and each of the three different alternatives proposed to approximate its main content.

As can be seen, the three proposed alternatives for *only* behave differently:

- *Less than* elicits judgments that are significantly lower than the ones obtained by using *only*.
- *At most* produces arguments that are significantly more persuasive than using *only* (but still not as persuasive as the `Bare` versions: the difference, although small, is still significant).

Table 1 Significance of the difference between `Only` and its proposed alternatives (paired Wilcoxon tests)

Alternative	Diff. w. <code>Only</code>	V	p
<i>Less than</i>	6.79	107	0.00336
<i>At most</i>	-14.2	425.5	7.435e-05
<i>No more than</i> (Exp. 3)	0.6	185	0.488
<i>No more than</i> (Exp. 4)	-0.71	317	0.5193

- *No more than* is not significantly different from *only* in experiment 3. This is what prompted the choice of this version for experiment 4 where no significant difference arises either. This supports the idea that *No more than* is an appropriate way to paraphrase the argumentative component of *only*.

Therefore, from the argumentative point of view and in the cases at hand, the best paraphrase for *only* is given by using the expression *no more than*.

3.3 Discussion

3.3.1 Inversion Hypothesis

The results of the experiments confirm the inversion hypothesis: the use of *only* drastically affects the argumentative orientation of its host argument in the cases at hand. I showed it to be true in the specific case where the quantity of positive and negative evidence that should be understood by the subjects is equivalent. Therefore the prediction that I attributed to the Bayesian reasoning approach on the grounds of its treatment of the argument of ignorance is falsified.

It is important to note here that the inversion hypothesis does not necessarily predict that the results of experiment 4 should mirror those of the first three experiments for the levels *Bare* and *Only*. The hypothesis merely predicts that the use of *only* should affect the evaluation of the argument in a direction that is opposite to the one defended by the *Bare* version. It does not however entail anything on the *strength* of the produced argument. In other words, the fact that the gap between these two levels is less important in experiment 4 than in the others does not argue against the inversion hypothesis.

A related issue is that if we compare the values for *Only* between the experiments 3 and 4, the difference is not significant (Wilcoxon non-paired test, $W = 15758.5$, $p = 0.6817$). Furthermore, both scores lie around the middle of the possible range of values (48.2 for the first experiment, 46.7 for the inverted experiment).

Seeing that the value for *only* does not seem affected by the shift in argumentative target and that the scores for the *Only* sentences lie around the middle of the range of possible values, one could hypothesize that subjects have a hard time interpreting examples involving *Only* and that instead of properly evaluating the arguments they always choose a middle range value as a kind of default. The choice for this default could be motivated on the grounds that the sentences involving *only* are more complex to treat and thus that subjects who are unsure might resort to a kind of non-committal “safe haven”.

I have no decisive argument to turn down this interpretation but I believe it is not right on track. My intuition is based on the fact that in the inverted goal experiment, evaluating the *Only* case should be more easy than evaluating the *Bare* case since the conclusion that is explicitly given is supported by *Only* while *Bare* goes against it. So even though the sentence is more complex, its argumentative evaluation

should be facilitated. This could also be a way to explain how the strength of the argument is affected when using a reversal operator like *only*, which could be a way to account for the fact that the gap in experiment 4 is less important than in the other three. More work should however be done on the cost of processing an utterance against an explicit conclusion for this last argument to be valid.

3.3.2 Main Content Hypothesis

Regarding the main content hypothesis, we saw that *no more than* is the formulation that matches the argumentative behavior of *only*. Therefore we can accept the main content hypothesis if we assume that *no more than* aptly represents the main content conveyed by *only*. Since this has been assumed before and appears relevant for the studied examples, the conclusion seems reasonable enough. However it should be kept in mind that this paraphrase is not always appropriate. First of all, it essentially applies when the associate of *only* is scalar. In a non-scalar case like (20) replacing *only* by *no more than* does not appear very natural, at least not as an answer to “*Who came to the party?*”

- (20) a. Only Lemmy came to the party.
 b. ? No more than Lemmy came to the party.

Furthermore, as pointed out in (Winterstein 2012), in some cases *only* does not yield an inversion effect but rather appears to strengthen an argument. This is intuitively the case in (21-a) which argues for (21-b) both with and without *only*. It is argued that this happens when the associate of *only* occupies the top of a scale.

- (21) a. Lemmy only drinks single malt scotch.
 b. $\overset{\sim}{\rightarrow}$ Lemmy is a connoisseur of whisky.
 arg

In those cases replacing *only* by *no more than* would probably not be a good argumentative paraphrase. The same could also be said for *at most* which is also purported to be a good approximation of the main content of *only*.

Such a case is interesting for the main content hypothesis. For (21-a), one could paraphrase the main content as in (22).

- (22) Lemmy does not drink low and middle range whiskies.

Intuitively, (22) argues in favor of (21-b), so the main content hypothesis can be maintained even in the case of (21-a). As claimed in (Winterstein 2012), the advantage of this position is that this entails that the argumentative profile of *only* can be derived from its semantics without having to postulate it as a primitive meaning (which is the traditional way of treating argumentation in the linguistic perspective).

3.3.3 Bayesian Reasoning

Assuming that the previous hypotheses are correct, we can draw a refined account of the way the Bayesian interpretation of natural language takes place. Essentially, I make more precise the nature of the content that is used for the conditional update of the speaker's beliefs by stating that only the at-issue content is used.

Assuming, with Zeevat (2013, 2014), that looking for the interpretation I of an utterance U means finding I such that $\text{argmax}_I p(I)p(U|I)$, then this last hypothesis reduces to say that in this process, U is restricted to its at-issue content in this operation.

The exact extension of the notion of *interpretation* is beyond the scope of the paper. However, one can assume that the speaker's goal, i.e. the proposition he or she wishes to support by making his assertion, does belong to I . Therefore, the process of abduction of the argumentative goal is also constrained by the main content hypothesis: when figuring out the potential goals the speaker might be defending, the reasoning is only done by considering the at-issue part of the argument.

We apply this hypothesis to the case of (11) (repeated here).

- (11) a. Half the studies showed that drug X has undesirable side-effects.
 b. Only half the studies showed that drug X has undesirable side-effects.

According to the main content hypothesis, the contents considered in the update are:

- (11-a) : *Half the study showed that drug X has undesirable side-effects.* = E
 (11-b) : *No more than half the study showed that drug X has undesirable side-effects.*
 \equiv *Half the study did not show that drug X has undesirable side-effects.* = $\neg E$

Let's define $\mathcal{G}_E = \{p|P(p|E) > P(p)\}$, i.e. the set of goals for which E is an argument and $\mathcal{G}_{\neg E} = \{p|P(p|\neg E) > P(p)\}$ the corresponding set for $\neg E$. As shown above, we have $T \in \mathcal{G}_E$ and $\neg T \in \mathcal{G}_{\neg E}$. In the case of (11-a), determining the interpretation of the utterance entails determining a subset of \mathcal{G}_E that will correspond to the goals attributed to the speaker. In the case at hand we can consider that the discourse situation induces a simple bipartition, i.e.: $\mathcal{G}_E \subset \{T, \neg T\}$, the speaker argues either in favor or against the toxicity. Since $\neg T$ is not in \mathcal{G}_E , we then attribute T as being the goal targeted by (11-a). The same reasoning applies for (11-b).

A more complex situation would not have involved a simple bipartition. In that case, determining the speaker's goal would have implied finding out the element H' of \mathcal{G}_E that satisfies some criterion. A plausible one would be that H' had to be the one that maximally affects the probability of asserting E , i.e. $H' = \text{argmax}_H p(H)p(H|E)$, which follows the intuition mentioned above regarding I .

On a final note, we can also wonder how to integrate the contribution of presuppositions in the argumentative reasoning. As suggested by a reviewer, presuppositions could be treated as constraints on the prior probability. In (11-b), this would mean that the prior $P(T)$ should be affected by the knowledge of E , i.e. set to $P(T') = P(T|E)$, before being updated to $P(T'|\neg E)$. The problem with this hypothesis is that it contradicts the main content hypothesis and predicts that both versions of (11) should

argue in the same way (cf. the discussion on the argument of ignorance in Sect. 2.2 where I show that positive arguments “win” over negative ones).

Therefore, if presuppositions enter the picture of argumentation reasoning, they have to be treated in a different manner. It is rather plausible that they might affect prior probabilities in a way. The protocols presented here cannot however shed light on the matter and I have to leave this to future work.

4 Conclusion

Using the specific case of the semantics of *only* I tried to bridge the gap between Bayesian theories of reasoning and theories that deal with the interpretation of natural language items. I focused on the specific case of the argumentative interpretation of natural language utterances and defended the hypothesis that natural language arguments are evaluated on the sole basis of their at-issue content. This was backed up by various experiments.

Notably, we have seen that applying this hypothesis to the case of *only* makes the right predictions. An immediate consequence is that it is not necessary to postulate a specific argumentative component inherent in *only*. Nor did I need using any constraint that stipulates that *only*'s associate has to be situated “low” on a scale. The only requirement we needed from *only* is that its main content excludes alternatives. In the case at hand the associate of *only* is scalar in nature which means that the only possible alternatives must be higher on the scale than the associate itself: lower values do not count as viable alternatives since they are entailed by the associate. In my account this is **not** due to *only* but a consequence of the nature of the associate and is thus compatible with all the accounts that have such a description of the main content of *only*.

An alternative view on the facts presented in this work is to consider the way an addressee would reason when faced with a sentence like (11-b). Since the informational content is the same as (11-a) and since the speaker chose to utter a more complex version involving *only*, the addressee is entitled to infer that the speaker has a non-standard attitude towards the contents he or she expresses and that this attitude is expressed via *only*. At this point we can either assume that the meaning postulate behind *only* encodes the necessary “negative” component, or we can assume that the addressee evaluates the speaker's specific attitude by focusing on the at-issue content of the utterance. This corresponds to the main content hypothesis I have defended so far, but puts it in a dialogical perspective where instead of mechanically updating the belief function of the agents, the process takes into account the agents attitudes towards the contents they express. Developing this account for *only* is left for future work. My main point here was to underline the specific profile of *only* and the fact that it could be derived by focusing on its at-issue content. A more general program

for research would be to evaluate in which measure one can further motivate various layers of meaning (presupposition, implicature etc.) as means of conveying information while keeping it hidden from some parts of the process of interpretation such as argumentation. This is left for future research.

References

- Anscombe, J. C., & Ducrot, O. (1983). *L'argumentation dans la langue*. Liège, Bruxelles: Pierre Mardaga.
- Atlas, J. D. (1993). The importance of being only. *Journal of Semantics*, 10, 301–318.
- Beaver, D. I., & Clark, B. Z. (2008). *Sense and sensitivity: How focus determines meaning*. Oxford: Wiley-Blackwell.
- Besnard, P., & Hunter, A. (2008). *Elements of argumentation*. Cambridge: MIT Press.
- Beysade, C. (2010). Seulement et ses usages scalaires. *Langue Française*, 165(1), 103–124.
- Coppock, E., & Beaver, D. (2013). Principles of the exclusive muddle. *Journal of Semantics*. doi: 10.1093/jos/fft007.
- Ducrot, O. (1973). *La preuve et le dire*. Paris: Mame.
- Geurts, B., & van der Sandt, R. (2004). Interpreting focus. *Theoretical Linguistics*, 30, 1–44.
- Hahn, U., & Oaksford, M. (2006). A Bayesian approach to informal argument fallacies. *Synthese*, 152, 207–236.
- Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: a Bayesian approach to reasoning fallacies. *Psychological Review*, 114(3), 704–732.
- Hamblin, C. L. (1970). *Fallacies*. London: Methuen.
- Horn, L. (1969). A presuppositional analysis of only and even. In: R.I. Binnick et al., (Eds.) *Papers from the Fifth Regional Meeting* (pp. 318–327). Chicago Linguistic Society.
- Horn, L. (2002). Assertoric Inertia and NPI Licensing. In *The Panels, CLS* (Vol. 38, pp. 55–82). Chicago Linguistics Society.
- Ippolito, M. (2008). On the meaning of *only*. *Journal of Semantics*, 25(1), 45–91.
- Jayez, J. (2010). Projective meaning and attachment. In M. Aloni, H. Bastiaanse, T. de Jager, & K. Schulz (Eds.), *Logic, language and meaning* (pp. 325–334). Berlin: Springer.
- Jayez, J., & Tovena, L. (2008) Presque and almost: How argumentation derives from comparative meaning. In O. Bonami, P. C. Hofherr (Eds.) *Empirical issues in syntax and semantics*, (Vol. 7, pp. 1–23). CNRS.
- Krifka, M.: Additive particles under stress, In *Proceedings of SALT 8* (pp. 111–128). CLC Publications, Cornell (1999).
- Merin, A. (1999). Information, relevance and social decision-making. In L. Moss, J. Ginzburg, & M. de Rijke (Eds.), *Logic, language, and computation* (Vol. 2, pp. 179–221). Stanford: CSLI Publications.
- Oaksford, M., & Chater, N. (2007). *Bayesian rationality—the probabilistic approach to human reasoning*. Oxford: Oxford University Press.
- Oaksford, M., & Hahn, U. (2004). A Bayesian approach to the argument from ignorance. *Canadian Journal of Experimental Psychology*, 58, 75–85.
- Pearl, J. (2009). *Causality: Models reasoning and inference* (2nd ed.). New York: Cambridge University Press.
- van Rooij, R., & Schulz, K. (2004). Exhaustive interpretation of complex sentences. *Journal of Logic, Language and Information*, 13, 491–519.
- Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics*, 1, 75–116.
- Simons, M., Tonhauser, J., Beaver, D., & Roberts, C. (2010). What projects and why. In *Proceedings of Semantics and Linguistic Theory (SALT) 20* (pp. 309–327). CLC Publications.
- Walton, D. N. (1996). *Arguments from ignorance*. Philadelphia: Pennsylvania State University Press.

- Winterstein, G. (2012). 'Only' without its scales. *Sprache und Datenverarbeitung*, 35–36, 29–47.
- Zeevat, H. (2009). Only as a mirative particle. *Sprache und Datenverarbeitung*, 33, 179–196.
- Zeevat, H. (2013). Accommodation in communication. *Manuscript*. ILLC: University of Amsterdam.
- Zeevat, H. (2014). *Language production and interpretation: Linguistics meets cognition*. Leiden: Brill.
- Zenker, F. (Ed.). (2013). *Bayesian argumentation: The practical side of probability*. Dordrecht: Springer.

Variations on a Bayesian Theme: Comparing Bayesian Models of Referential Reasoning

Ciyang Qing and Michael Franke

Abstract Recent developments in Bayesian experimental pragmatics have received much attention. The Rational Speech Act (RSA) model formalizes core concepts of traditional pragmatic theories quantitatively and makes predictions that fit empirical data nicely. In this paper, we analyze the RSA model and its relation to closely related game theoretic approaches, by spelling out its *belief*, *goal* and *action* components. We introduce some alternatives motivated from the game theoretic tradition and compare models incorporating these alternatives systematically to the original RSA model, using Bayesian model comparison, in terms of their ability to predict relevant empirical data. The result suggests that the RSA model could be adapted and extended to improve its predictive power, in particular by taking speaker preferences into account.

Keywords Referential expressions · Scalar implicature · Pragmatic reasoning · Referential game · Bayesian pragmatics · Rational speech-act theory · Game-theoretic pragmatics · Experimental pragmatics · Bayesian analysis · Model comparison

1 Introduction

Human language communication is efficient, in that people need not always say every detail in order to be understood. Rather, speakers often only say what is most relevant, and listeners can often seemingly effortlessly grasp the intended meaning beyond what is literally said. This ability to do pragmatic inference has been long

C. Qing (✉)
Linguistics, Stanford University,
Stanford, CA, USA
e-mail: qciyang@gmail.com

M. Franke
Linguistics, University of Tübingen,
Tübingen, Germany

studied and the *conversational implicature* theory by Grice (1975) is one of the most prominent theories in the field. However, since it is hard to formalize the *Cooperative Principle* and the *Conversational Maxims*, Grice's theory is not precise enough to give empirically testable predictions, especially quantitative predictions.

The Bayesian *Rational Speech Act* (RSA) model attempts to address this issue by using information-theoretic notions together with the Bayesian inference framework (Frank et al. 2009; Frank and Goodman 2012; Bergen et al. 2012; Goodman and Stuhlmüller 2013). It has been shown that the model could yield quantitative predictions that are highly correlated to actual human judgments (Frank and Goodman 2012; Goodman and Stuhlmüller 2013).

Despite the RSA model's theoretical promise and empirical success, if we analyze the design of the model, we will find several choices that are not *prima facie* obvious and might thus need further consideration and justification. These choices have their alternatives in closely related game-theoretic approaches to pragmatics (Rabin 1990; Stalnaker 2005; Benz and van Rooij 2007; Franke 2011; Jäger 2013), which have similar but slightly different conceptual motivations. Hence it is important to systematically compare the original RSA model with all these alternatives and test their predictions on empirical data, so as to gain a better understanding of the relation between these models. Doing so will help illuminate the theoretical commitments and empirical implications of each design choice, which will enhance our understanding of the nature of human pragmatic inference.

The outline of the paper is as follows. Section 2 introduce the referential communication games investigated by Frank and Goodman (2012) and the original RSA model proposed there. Section 3 analyzes a few design choices in this model and introduces their alternatives in game theoretic pragmatics, emphasizing the distinction between *belief*, *goal* and *action* and unifying various models as different interpretations of these three components. Section 4 reports on the results of an experiment similar to Frank and Goodman (2012). In Sect. 5, we compare the predictions of different models of our new data. Finally, we discuss the implications of the results of the model comparison in Sect. 6, concluding that the original RSA model can be improved on, in particular by taking speaker preferences into account, but we also point out that a more complex picture of probabilistic pragmatic reasoning about referential expressions awaits further exploration.

2 Referential Games and the RSA Model

Referential games are confined interactive reasoning tasks used to study pragmatic inference in a controlled environment (Stiller et al. 2011; Frank and Goodman 2012; Degen and Franke 2012). A referential game consists of a set of objects called a *context*. For example, a context might contain different shapes of different colors in various arrangements, such as shown in Fig. 1a. The speaker in the game is asked to refer to one of the objects (the *target*) by uttering an expression (typically a single word denoting a feature, i.e., color or shape, of the target) to the listener. The listener,

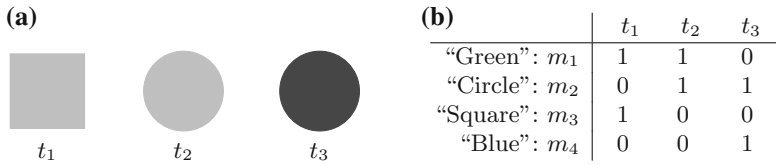


Fig. 1 A simple referential game, as used in Frank and Goodman (2012). **a** Example content: *green square* and *circle*, and *blue circle*. **b** Vocabulary and truth table

who does not know which object is the target, needs to recover the target based on the speaker’s choice of expression.

Let us first try to play this game ourselves, to get a better sense of what it is about and why it is relevant to human pragmatic reasoning. For simplicity let us assume that the utterances that the speaker is allowed to use are from a *vocabulary* which is commonly known between the speaker and the listener. The content of the vocabulary and the (truth-conditional/literal) meaning of each word are shown in Fig. 1b. Suppose we play as the speaker and the target is t_1 , the green square. We can either use “Square” or “Green” to truthfully refer to it, but it seems more prudent to use “Square”, as there is only one square in the context and thus the listener can easily identify it, whereas there are two green objects, which makes “Green” ambiguous. In terms of the Gricean *Maxim of Quantity*, we should use “Square” because it is more informative than “Green” in the given context; the surplus informativity moreover seems relevant. Similarly, we should use “Blue” to refer to t_3 , the blue circle. However, using “Blue” to refer to the blue circle might intuitively sound a little unnatural, as color terms are usually used as adjectives,¹ while we usually use nouns to refer to concrete objects. This inclination becomes more evident when we want to refer to t_2 , the green circle. While “Green” and “Circle” are equally ambiguous in this case, we might nevertheless prefer the latter.

Now let us turn to play as the listener. If we hear “Square” or “Blue”, we easily know the intended referent as there is no ambiguity, but what if we hear “Circle” (or “Green”)? There are two circles in the context that we need to choose from. On the one hand, the blue circle, having the unique color in the context, seems to be perceptually dominant and thus easily captures our attention. On the other hand, from the previous analysis we know that if the blue circle were the intended referent, the speaker could have chosen “Blue” which is not ambiguous and thus more informative. Hence the listener needs to balance two sources of information, i.e., the (presumably subconscious) perceptual salience of different objects and the rational expectation of the likelihood of the speaker making the utterance for each object.

To test these intuitions, Frank and Goodman (2012) recruited 745 US participants via Amazon’s Mechanical Turk to take part in a one-shot referential game, either as the speaker or the listener. The context of the game always consisted of three objects (e.g., Fig. 1a). Each object had three features: color (red, blue, or green),

¹They are used as nouns mostly to refer to the colors themselves.

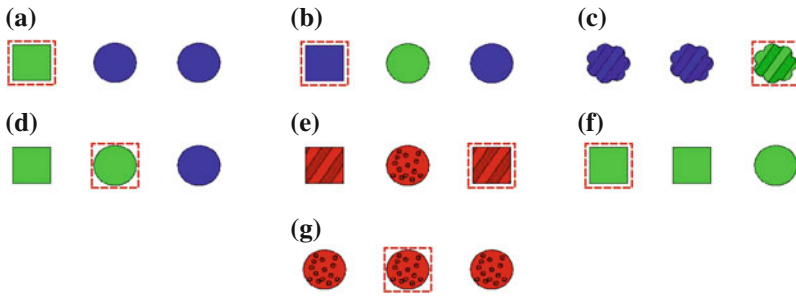


Fig. 2 Sample contexts for the 7 conditions used in Frank and Goodman (2012). The targets are indicated by *dotted lines*. Condition n/m means that the target has one feature shared by n objects and another feature shared by m objects in the context (including the target itself). For 2/2, either there is an object identical to the target, or the features are shared by different objects. **a** 1/1. **b** 1/2. **c** 1/3. **d** 2/2: identical (two *striped squares* and one *dotted circle*). **e** 2/2: different. **f** 2/3. **g** 3/3

shape (circle, square, or cloud), and texture (solid, or polka-dot, or striped). Every context had one feature fixed for all objects and the other two features systematically varied. The objects were randomly permuted and the target object was randomly assigned. Depending on how many objects shared the same value of a feature with the target, contexts were classified into 7 categories (see Fig. 2): 1/1, 1/2 (the target's first feature is unique, and there is another object sharing the same value of the second feature with the target), 1/3, 2/2 with the same object sharing both features with the target (which means there is another object identical to the target), 2/2 with different objects sharing the features, 2/3 and 3/3. The speakers were asked to choose a feature to refer to the target object.² In order to measure the perceptual salience, some of the listeners were told that the speaker was using a language that they did not understand and then they were asked to guess the intended referent. Other listeners were told a feature of the target and asked to choose the intended referent. Note that listeners were presented with pictures without dotted lines surrounding the target so they did not know the target. For each category in each of the speaker, salience and listener conditions, there were roughly 50 responses.³

The result of the experiment showed that speakers generally followed Grice's Maxim of Quantity, i.e., they tried to be informative by choosing the most unique feature of the target,⁴ and that listeners were influenced by both the perceptual salience

²More accurately, participants were asked to *bet* on features and referents. We will discuss this design choice when contrasting it with our replication in Sect. 4.

³As we can see, despite the fairly large number of participants recruited, the data in each condition are not that many. Also there might be a concern that some conditions might look less interesting than others. Again we will discuss these issues when we introduce our own replication in Sect. 4. Here we note that variety is important in order to test the general validity of a pragmatic theory. Thus the design used in Frank and Goodman (2012) served its purposes well, but may be adapted when the goal is different.

⁴Speakers' preference was not taken into account in Frank and Goodman (2012), so all features were randomly assigned and later lumped together in the analysis.

and the informativeness of the utterance that the speaker produced. However, the data, being quantitative in nature, cannot be fully explained by a classical qualitative Gricean account of scalar implicatures, so that a major challenge in terms of formal modeling becomes how to quantify notions such as informativeness and perceptual salience, and how these should be integrated to yield qualitative predictions.

The rational speech-act model (RSA) model addresses these issues by using information-theoretic concepts to measure informativeness and Bayesian inference to integrate different sources of information (Frank et al. 2009; Frank and Goodman 2012; Bergen et al. 2012; Goodman and Stuhlmüller 2013). In order to measure the informativeness of an utterance, the RSA model starts with the literal listener who upon receiving an utterance m does a conditioning on its literal meaning:

$$\rho_0(t | m) = \mathcal{U}(t | \llbracket m \rrbracket), \tag{1}$$

where \mathcal{U} is a uniform distribution over all the possible referents. The informativeness of utterance m for the intended referent t can be measured as the negative Kullback-Leibler divergence of the induced literal listener’s belief ρ_0 from the speaker’s own belief δ_t :

$$\text{Info}(m, t) = -\text{KL}(\delta_t \| \rho_0) = - \sum_{t'} \delta_t(t') \log \left(\frac{\delta_t(t')}{\rho_0(t')} \right), \tag{2}$$

where δ_t is a delta distribution with all probability mass on target object t , as the speaker knows her intended referent:

$$\delta_t(t') = \begin{cases} 1 & \text{if } t' = t \\ 0 & \text{otherwise} \end{cases} . \tag{3}$$

The speaker acts sub-optimally by choosing the utterance that soft-maximizes her expected utility, which is defined as the informativeness of the utterance subtracted by its cost:

$$\sigma(m | t) \propto \exp(\lambda_S \cdot \text{U}(m, t)) = \exp(\lambda_S \cdot (\text{Info}(m, t) - \text{Cost}(m))), \tag{4}$$

where λ_S is a parameter measuring the speaker’s degree of rationality, i.e., to what extent the speaker sticks to the strict optimum.⁵ The cost term is used to encode preference in different utterances, be it about the utterances’ lengths or syntactic categories.

⁵Mathematically, it can be shown that when $\lambda_S = 0$, $\sigma(m | t)$ is uniformly distributed over all messages, meaning that the speaker just randomly selects one of the messages without any optimality consideration. On the other hand, when $\lambda_S \rightarrow \infty$, $\sigma(m | t)$ is non-zero only when m yields the maximal utility, meaning that the speaker only selects the strictly optimal option(s). This rule is widely used in the AI literature (e.g., Sutton and Barto 1998) and is also known as *exponential choice rule*.

From (1)–(4) we obtain the speaker’s production rule⁶:

$$\sigma(m | t) \propto \exp(\lambda_S \cdot (\log \mathcal{U}(t | \llbracket m \rrbracket) - \text{Cost}(m))) . \quad (5)$$

The pragmatic listener in the RSA model, upon receiving the utterance m , performs a Bayesian update on his prior belief $\mathcal{S}(t)$ by using an estimate of the speaker’s behavior in (5):

$$\rho(t | m) \propto \mathcal{S}(t) \cdot \sigma(m | t) . \quad (6)$$

Bayes’ rule naturally integrates the perceptual salience of each object, which is treated as the prior $\mathcal{S}(t)$ and can be empirically measured, with listener’s expectation of the speaker being informative, which is incorporated as the likelihood, thus addressing the previously mentioned challenge of balancing different sources of information. Setting $\lambda = 1$ and $\text{Cost}(m) = 0$ for all m , Frank and Goodman (2012) obtained a highly significant correlation between the model prediction and their experiment data on actual speaker and listener behavior gathered from referential games of varying complexity.

3 Alternatives in Design Choices

The RSA model has a theoretical advantage over traditional formal pragmatic theories in that it provides quantitative predictions, and it has been shown to fit relevant empirical data well. However, if we want to learn more about pragmatic reasoning about referential expressions, it will be worthwhile to examine RSA carefully to pin down its major components, to spell out its main design choices and their underlying assumptions, and to test their contribution to the predictive power of the model through statistical comparison with natural alternatives. In this section, we will therefore compare the RSA model to closely related game theoretic approaches, that likewise assume that speakers and listeners form cascading beliefs about mutual behavior and seek to optimize their behavior based on these beliefs (Rabin 1990; Stalnaker 2005; Benz and van Rooij 2007; Franke 2011; Jäger 2013).⁷ We will describe a class of potential alternatives to RSA and then compare the predictive power of these alternatives on relevant empirical data in later sections.

⁶Note that we can further simplify (5) as follows: $\exp(\lambda_S \cdot (\log \mathcal{U}(t | \llbracket m \rrbracket) - \text{Cost}(m))) = \exp(\lambda_S \log \mathcal{U}(t | \llbracket m \rrbracket)) \cdot \exp(-\lambda_S \cdot \text{Cost}(m)) = (\mathcal{U}(t | \llbracket m \rrbracket))^{\lambda_S} \cdot \exp(-\lambda_S \cdot \text{Cost}(m))$. Incidentally, if we do not take costs into account, i.e., let $\text{Cost}(m) = 0$, then (5) reduces to $\sigma(m | t) \propto (\mathcal{U}(t | \llbracket m \rrbracket))^{\lambda_S}$, which is effectively a *power rule* that soft-maximizes $\mathcal{U}(t | \llbracket m \rrbracket)$. In this respect (5) can be seen as an action-oriented speaker model (the exact definition of which will be given in the next section) equipped with a different kind of soft-max rule. However, we note that this is a mere technical coincidence, and as a first step, throughout this paper we only consider the exponential rule as the soft-max rule. Worthy et al. (2008) provide a detailed discussion and empirical analyses on the differences between the exponential and power rules.

⁷See Franke and Jäger (2014) for overview and comparison of game theoretic and Bayesian models.

There are three closely related notions, i.e., *belief*, *goal* and *action*, that play a crucial role in RSA as well as game theoretic models. The relation and distinction between them can be best illustrated by looking at (5), repeated here:

$$\sigma(m | t) \propto \exp(\lambda_S \cdot (\log \mathcal{U}(t | \llbracket m \rrbracket) - \text{Cost}(m))) .$$

The term $\mathcal{U}(t | \llbracket m \rrbracket)$ is the speaker's *belief* about the listener's interpretation of the utterance. The expected utility $U(m, t) = \log \mathcal{U}(t | \llbracket m \rrbracket) - \text{Cost}(m)$ describes the speaker's *goal* of communication, i.e., inducing a belief that is as close as possible to her own with minimal effort. The production rule $\sigma(m | t)$ on the left-hand side specifies the speaker's *action*, as it defines the probability with which, according to RSA, speakers would choose a particular expression. The soft-max rule connects these notions, as a (sub-)rational agent's action depends on his belief and goal. While the RSA and game theoretic models all share this general architecture in their designs, they vary in the specific interpretations of the three ingredients, which reflect different views and emphasis on language and communication and give rise to a series of interesting and important empirical questions. Again, let us continue with the speaker model (5) to illustrate some points of divergence and formally define the corresponding alternatives.

First of all, although it is *prima facie* reasonable to hypothesize, as Frank and Goodman (2012) do, that the empirically measured perceptual salience $\mathcal{S}(\cdot)$ is common knowledge between speaker and listener, it does not actually affect RSA's production rule (5) at all. This makes it unclear in what sense the speaker can be said to know the listener's perceptual salience, since neither his beliefs nor actions depend on it. A natural variant, which is used in some game theoretic models, would be to replace the literal listener's uniform distribution $\mathcal{U}(t)$ in (1) with the salience prior $\mathcal{S}(t)$. This leads to the alternative production rule:

$$\sigma_S(m | t) \propto \exp(\lambda_S \cdot (\log \mathcal{S}(t | \llbracket m \rrbracket) - \text{Cost}(m))) . \quad (7)$$

Secondly, RSA measures the informativeness of an utterance m , which is a crucial part of the communicative goal, in terms of how close the induced belief of the literal listener is from the speaker's own belief. This means that the RSA model sees the goal of communication as conveying belief. While it is normally true that language does convey the belief of the speaker, it is questionable at least in this referential scenario whether letting the listener form a proper belief is the ultimate goal of the communication. After all, if the speaker wants to refer to something, it seems that in the end what matters is whether the listener actually picks out the intended referent successfully (e.g., when the speaker wants the listener to pass something). This view is inherent in game theoretic approaches where agents' beliefs are backed up by explicit formulations of their utilities. We might call this latter view *action-oriented*, in contrast to the *belief-oriented* view of communication which the RSA model adopts, as it interprets the goal of communication as invoking the

intended action rather than forming an accurate belief.⁸ Thus, according to this view, the informativeness of a message would be measured as the probability of the listener choosing the intended referent. Formally speaking, we have the action-oriented speaker model⁹:

$$\sigma_a(m | t) \propto \exp(\lambda_S \cdot (\rho_0(t | m) - \text{Cost}(m))) . \quad (8)$$

Hence we have four types of speaker models that differ in either the speaker's belief about the literal listener, or the speaker's goal of communication. We now introduce a uniform notation σ_{xy} , $x \in \{a, b\}$, $y \in \{\mathcal{U}, \mathcal{S}\}$ for them:

$$\sigma_{ay}(m | t) \propto \exp(\lambda_S \cdot (y(t | m) - \text{Cost}(m))), \quad (9)$$

$$\sigma_{by}(m | t) \propto \exp(\lambda_S \cdot (\log y(t | m) - \text{Cost}(m))), \quad (10)$$

where \mathcal{U} is the uniform prior and \mathcal{S} is the salience prior. For example, in the original RSA model, the speaker does not take listener's salience prior into account and he has a belief-oriented goal of communication. Thus it will be denoted as $\sigma_{b\mathcal{U}}$.

Finally, the original RSA speaker model (Frank and Goodman 2012) has $\text{Cost}(m) = 0$ for all m , which means that the potential speaker preference in different utterances is not taken into account (but not so in, e.g., Bergen et al. (2012)). As we previously pointed out, our intuition seems to suggest a preference for nouns over adjectives as referential expressions.¹⁰ Since it is an empirical question whether, to what direction or to what extent such a preference exists, we leave open all the possibilities. Technically speaking, since only the difference in costs matters, we define the cost function using a constant $c \in \mathbb{R}$

$$\text{Cost}(m) = c \text{ if } m \text{ is an adjective and } 0 \text{ otherwise} . \quad (11)$$

⁸Note that the two views are clearly intimately related, despite our emphasis on their differences here. An accurate belief is usually helpful for choosing a desired action. In fact, forming a belief can itself be seen as an action that the listener performs, which might well be all that the speaker cares about in many everyday situations, e.g., casual talks or responses to inquiries of information. Also, the relevant actions might vary greatly across situations which makes belief formation a practical alternative to serve as the goal. Nevertheless, they are still essentially different views in spite of the fact that they often coincide in practice. The referential communication scenario we investigate in this paper reflects the distinction.

⁹There might be concern about whether there should be another soft-max step for the literal listener's choice probability $\rho_0(t | m)$ in order for it to be action-oriented. This depends on the precise conceptual status of the literal listener in the model and to our knowledge there is no consensus yet. We tend to think of a literal listener as an automatic subconscious process that involves no optimality reasoning or decision-making. Thus we consider it appropriate to use $\rho_0(t | m)$ as the action of the literal listener. However, we do agree that in principle there can be a model whose literal listener does a soft-max over the salience prior, and such a model can be similarly included in the model comparison.

¹⁰The rich tradition on referential expression generation (Kramer and van Deemter 2012) has also identified the need to take speaker preferences into account, and Gatt et al. (2013) criticize RSA on these grounds as well.

If $c > 0$ it means there is a preference for nouns and if $c < 0$ then the preference is for adjectives. No preference exists if $c = 0$.

Now we turn to the listener model (6), where the boundary between belief, goal and action becomes less clear:

$$\rho(t | m) \propto \mathcal{S}(t) \cdot \sigma(m | t) .$$

Let us start from the relatively clear part. As pointed out previously, the likelihood term $\sigma(m | t)$ is the listener's belief about how the speaker behaves. One natural question is to what extent the listener's belief correlates with the actual production. We thus treat the speaker's production term $\sigma(m | t)$ in the listener's model as a parameter, making the association with a belief about production of the listener model explicit:

$$\rho(\sigma_{xy})(t | m) \propto \mathcal{S}(t) \cdot \sigma_{xy}(m | t) . \quad (12)$$

Note that the speaker's production rule has two parameters λ_S, c which are also included in the above specification of the listener model.

Next, even though our intuition suggests different objects have different perceptual salience and thus might affect our judgment, it is after all an empirical question whether it is relevant in the interpretation of referential expressions. It is in principle possible that the listener does not take into account the perceptual salience in his reasoning, which means he has a uniform prior over the referents:

$$\rho_{\mathcal{U}}(\sigma_{xy})(t | m) \propto \mathcal{U}(t) \cdot \sigma_{xy}(m | t) . \quad (13)$$

Finally, since the RSA model adopts a belief-oriented view on the goal of communication, its listener model only consists of the belief about the intended referent after hearing the utterance, obtained by a Bayes update. However, according to the action-oriented view, this is not quite the end of the story. The listener often needs to decide the exact referent of a referential expression. (Again, consider the case in which the listener is asked to pass something.) Hence in such cases the listener will choose an object by, essentially, (soft-)maximizing over his beliefs. Formally, the action-oriented listener model becomes:

$$\rho_{av}(\sigma_{xy})(t | m) \propto \exp(\lambda_L \cdot \rho_{bv}(\sigma_{xy})(t | m)), \quad (14)$$

where $v \in \{\mathcal{U}, \mathcal{S}\}$, λ_L is the parameter measuring the listener's degree of rationality, and ρ_{bv} is the belief-oriented model that does the Bayesian update:

$$\rho_{bv}(\sigma_{xy})(t | m) \propto v(t) \cdot \sigma_{xy}(m | t) . \quad (15)$$

For instance, the original RSA listener model is a belief-oriented one with the perceptual salience as prior, whose belief about the speaker is $\sigma_{b\mathcal{U}}$, hence it is denoted as $\rho_{b\mathcal{S}}(\sigma_{b\mathcal{U}})$.

Fig. 3 A sample context

4 The Experiment

In the previous section we analyzed the design of the original RSA model, spelled out the relation and distinction between belief, goal and action, and proposed a family of alternative models based on different interpretation of these notions. This gives rise to a series of interesting empirical questions regarding the underlying assumptions of the models. In order to gain insight into these questions by comparing the predictive power of different models, we conducted the following experiment to collect empirical data, which we will use to test the predictions of different models in the next section.

Participants

We recruited 1032 US participants via Amazon's Mechanical Turk. Each participant played a one-shot referential game either as a speaker or as a listener and received a payment of 5 cents. If a participant contributed multiple trials, only the first was included. All participants passed the attention check described below.

Materials and Procedure

Each participant saw a context consisting of three objects (Fig. 3 as an example) and reported on the number of objects in the context as an attention check. Participants in the listener conditions were told to imagine someone was talking to them who used some utterance (depending on the specific condition tested) to refer to exactly one of the objects in the context. Then they were asked to choose what they thought to be the intended referent. Participants in the speaker conditions were told to imagine they were talking to someone and they wanted to refer to one of the objects (indicated by a small red arrow¹¹). Then they were asked to choose between two words to refer to it.

Each context had a square and a circle sharing the same color, and another circle having a different color. Hence each object had two features: shape (square/circle) and color (green/blue) and each context had two ambiguous (shared by two objects) features. In the listener salience condition, participants were told that the person talking to them was using a foreign language they could not understand, and in the other two listener conditions, the utterances used were the ambiguous feature of shape and color, respectively. In the speaker conditions, the target could be any of the three objects: the one with the unique color, the one with the unique shape and the one with both features shared. The two words for the participants to choose between were the features of color and shape of the target object.

¹¹ Participants were told that the arrow was only for illustration and thus the person they were talking to could not see it.

In order to minimize the effect of the confounding factors, we counterbalanced the positions and colors of the objects, as well as the orders of the candidate words in the speaker conditions. Thus the unique color in a context could be either blue (as in Fig. 3) or green and the order of the objects could be any of the permutations.¹²

Design

There are four major design choices in our own experiment that differ from the original experiment in Frank and Goodman (2012). First of all, the original experiment used a *betting paradigm*, i.e., participants were asked to bet over a range of options and they were instructed that the amount of money that is bet on an option should reflect their confidence that the option is correct. Even though the betting paradigm has the merit of providing us with graded responses from each individual participant, the caveat is that it is unclear whether it measures beliefs or actions. This can lead to confusion when we are to fit the model predictions to the empirical data without knowing whether they are directly comparable. In addition, since the betting paradigm is more or less introspective in nature, it tends to be not very accurate. Thus we used a *forced choice design* instead, which clearly measures the action, does not rely on introspection and, technically speaking, provides for a straightforward likelihood function to be used in subsequent statistical model comparison. Secondly, since we decided to investigate the influence of the speaker's preference as well as the listener's perceptual salience, we focused on contexts equivalent to Fig. 1a, which require reasoning that is highly reminiscent of Gricean accounts of scalar implicature calculation, and examined different features separately. Thirdly, we only included the features of color and shape, with binary values (green/blue and square/circle) for two reasons. On the one hand, we wanted to ensure the vocabulary to be common knowledge, so we excluded the feature of texture. On the other hand, we excluded feature values red and cloud as they are significantly more salient than the others, which might override the pragmatic effect. Finally, we did not use dotted lines surrounding an object as the way to indicate the target to the speaker, as Frank and Goodman (2012) did, because that might unduly emphasize the feature of shape and thus be a confound in the experiment. Instead we used a small arrow pointing to the target.

Despite these differences, the remaining aspects in our experiment were almost identical to those of the original one, e.g., the phrasing of the instructions.

The results of the speaker and listener conditions of our experiment are shown as follows.

4.1 Speaker Conditions

There were 432 participants in the speaker conditions, 144 in each condition. The numbers of participants choosing each word in each condition are shown in Table 1.

¹²Due to the scale of our preliminary experiment we did not counterbalance the shape of the objects, i.e., a context always had one square and two circles.

Table 1 Speaker conditions

Target	“Green”	“Square”	“Blue”	“Circle”	Total
	9	135	–	–	144
	–	–	119	25	144
	63	–	–	81	144

Note that when the target was the object with the unique shape (as in the first row of the table), the feature of shape (“Square”) should be the optimal utterance because the listener could uniquely identify it. Similarly when the target was the object with the unique color (the second row), the optimal utterance would be the feature of color (“Blue” in this case). When the target was the object with both features shared, both features should be equally ambiguous because of the context’s symmetric nature.




From the data in Table 1, we can see that the speakers tend to choose the optimal feature more often when the target has the unique shape than when it has the unique color ($\chi^2 = 8.5$, $p < 0.01$). Even though they seem to prefer the feature of shape when both of the target’s features are shared, the difference is not statistically significant from uniform random choice ($\chi^2 = 2.25$, $p = 0.13$).

4.2 Listener Conditions

There were 600 participants in the listener conditions, 240 in the salience condition and 180 in each of the remaining two conditions. The numbers of participants choosing each object in each condition are shown in Table 2.

The result of the salience condition will be used as the empirical estimation of contextual salience. For the other two conditions, the object with both features shared (the green circle in the above table, but note that we counterbalanced colors) is what the Gricean pragmatic account predicts to be the target. The result in Table 2 shows that when the message is “Green”, listeners prefer the green circle which is predicted by the Gricean pragmatics, while they tend to stick to the blue circle which is more perceptually salient when the message is “Circle”. The behavioral patterns in both conditions are significantly different from uniform random choice, and they

Table 2 Listener conditions

				Total
Salience	71	30	139	240
“Green”	65	115	0	180
“Circle”	1	62	117	180

significantly differ from each other in whether they conform to the predictions by Gricean pragmatic theory.

5 Model Comparison

We use a Bayesian approach to model comparison (Jeffreys 1961; Kass and Raftery 1995; Jaynes 2003; Vandekerckhove et al. [in press](#)) to find out which alternative best explains the observed data. The models under investigation have unspecified parameters: the speaker's degree of rationality λ_S , the cost of adjectives c , and, for those listener models that have an action-oriented communication goal, the listener's degree of rationality λ_L . Since there is no principled theory to determine the value of the parameters, we will rely mostly on relatively uninformed hyperpriors (so-called to distinguish them from the salience priors). Based on a specification of hyperpriors, we calculate the models' *evidences* and compare them by their *Bayes factors*. The evidence of a model M is the weighted average of the likelihood of observing the data under all parameter values:

$$\text{Ev}(M) = \int \text{Pr}(\theta) \cdot \text{Pr}(D \mid M, \theta) d\theta, \quad (16)$$

where $\text{Pr}(\theta)$ is the hyperprior over parameter(-tuple) θ associated with model M and $\text{Pr}(D \mid M, \theta)$ is the likelihood of the observed data D given M and each concrete instantiation of the parameter(-tuple) θ . The Bayes factor $K_{M_2}^{M_1}$ is a comparative measure for the plausibility of model M_1 over M_2 , given their respective hyperpriors and the data in question:

$$K_{M_2}^{M_1} = \frac{\text{Ev}(M_1)}{\text{Ev}(M_2)}. \quad (17)$$

Model M_1 makes the data more likely whenever $K_{M_2}^{M_1} > 1$, but normally only a Bayes factor $K_{M_2}^{M_1} > 3$ (or sometimes $K_{M_2}^{M_1} > 5$) is considered substantial. Values $K_{M_2}^{M_1} > 10$ are considered strong evidence.

In a sense, comparison by Bayes factors is comparing models in a wider sense of the term: we actually compare pairs consisting of a model and its associated hyperprior. For clarity, we refer henceforth to a model-hyperprior pair as a Model.

Speaker data. First we look at the speaker models σ_{xy} , $x \in \{a, b\}$, $y \in \{U, S\}$. Each model has two parameters λ_S and c . We assume that they are independent of each other:

$$\text{Pr}(\lambda_S, c) = \text{Pr}(\lambda_S) \cdot \text{Pr}(c). \quad (18)$$

Table 3 Log-evidences of speaker models

Support of $P(c)$	$\sigma_{b\mathcal{U}}$	$\sigma_{a\mathcal{U}}$	$\sigma_{b\mathcal{S}}$	$\sigma_{a\mathcal{S}}$
$[0, 0]$	-67.48	-67.81	-104.14	-157.75
$(-0.4, 0.4)$	-32.92	-33.15	-81.44	-139.27

We are uncertain about the rationality of the speaker:

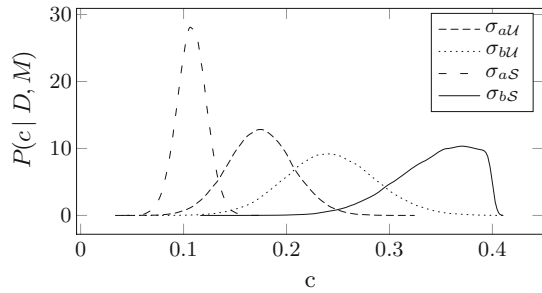
$$\Pr(\lambda_S) = \mathcal{U}_{(0,11)}(\lambda_S), \quad (19)$$

which is a uniform distribution over $(0, 11)$. Excluding λ -values ≥ 11 serves practical purposes only, but is innocuous since the regions of non-negligible posterior likelihood of λ lie safely in the chosen interval for all models. Next, to allow for the possibility of speaker preferences (nouns over adjectives or vice versa), we consider two types of hyperpriors for costs c . The first hyperprior has $\Pr(c) = \delta(c)$, the Dirac delta distribution, that assigns all probability mass to $c = 0$. This captures the assumption that there is no speaker preference. The second hyperprior is $\mathcal{U}_{(-0.4, 0.4)}$, which captures the notion that a preference exists, without commitment to either direction. We restrict our attention to the interval $(-0.4, 0.4)$, because we consider higher levels of cost implausible, given that utilities for successful communication live in $[0, 1]$ and that we believe that strive for communicative success should outrank preference satisfaction in a rational model of communication. Taken together, there are four speaker models, two hyperpriors for each, so that we compare eight Models with respect to their evidences.

Evidences of speaker Models were calculated by grid approximation. The results are shown as log-evidences in Table 3. (Notice that the Bayes factor between models is easily compared by taking the differences between log-evidences.) We can see that the data very strongly supports speaker models that do not take the listener's perceptual salience into account. Also, it seems that action-oriented models are slightly better than their belief-oriented counterparts, even though the relevant Bayes factors are not substantial by common standards. Finally, our data makes each speaker Model that does allow for a speaker preference strongly more plausible than its counterpart that does not. A look at the posterior likelihood of c for each speaker model informs us that our data supports the belief in a speaker preference for nouns, see Fig. 4.¹³ In sum, our data supports the view that the speaker does not take into account the perceptual salience of the listener, while having her own preference for shape terms over color terms.

¹³A technical aside: formally, action-based speaker models, as formulated in Sect. 3, assign a non-zero probability to the event that a speaker chooses a description which is not true of the target object (unlike belief-based models). But since subjects in our experiments could only select between true alternatives, the predictions of action-based speaker models were restricted to truthful choices in the reported model comparison.

Fig. 4 Posterior distributions over costs given the data for each speaker model



Listener data. Each listener model has a speaker model nested inside as a belief of the listener about the speaker’s behavior. Section 3 introduced a total of 16 potentially relevant listener models, but we will focus on a selection only. For one, we restrict our attention to those listener models that are either entirely belief-based or entirely action-based. In other words, we exclude as notionally inconsistent models like $\rho_a(\sigma_b)$ where the receiver part assumes an action-based goal structure and the speaker part a belief-based goal structure. For another, we assume that the listener’s model of the speaker is a reasonable one and therefore put to the side listener models that embed speaker models that are highly implausible, given the speaker data as discussed above. Effectively, this discards listener models that include speaker models that take the listener’s salience prior into account. These two principled restrictions leave us with four listener models to compare.¹⁴

Further variation comes from different relevant hyperpriors. Belief-based models of the listener have the same parameters as speaker models: the speaker’s rationality λ_S and the speaker’s preference costs c . Importantly though, hyperpriors over λ_S and c , although formally parallel to those for speaker models, have a different interpretation in listener models where they capture our prior beliefs about the listeners’ beliefs about the speakers’ likely rationality and preferences. This is true as well for action-based models, which also have an additional parameter λ_L . Hyperpriors for the latter encode our prior beliefs about the listener’s actual rationality.

We consider a variety of hyperpriors that differ in whether they take the speaker’s costs into account, whether the listener’s beliefs about the speaker’s rationality and preferences are uninformed (i.e., flat) or informed (i.e., given by the posterior likelihood of parameters given the actual speaker data) and, in action-based models, whether the listener’s level of rationality corresponds to his “informed” beliefs. The latter option effectively implements the assumption that there is a tight

¹⁴Similar to the remarks about speaker models in Footnote 13, we need to be mindful about the fact that action-based listener models assign non-zero probability to the choice of referents of which the given description is not true. Here, despite the fact that subjects in our experiments could in principle choose referents of which the given description is not true (this indeed happened once), we restricted the predictions of these models to only semantically compatible referent choices. We then discarded the one data point in which a subject choose a semantically non-compatible referent.

correlation between the speaker’s actual rationality, the listener’s actual rationality and the listener’s beliefs about the speaker’s rationality.

Concretely, we consider the following “flat” hyperpriors:

$$\begin{aligned}\Pr(\lambda_S, c) &= \mathcal{U}_{(0,11)}(\lambda_S) \cdot \Pr(c) \\ \Pr(\lambda_S, c, \lambda_L) &= \mathcal{U}_{(0,11)}(\lambda_S) \cdot \Pr(c) \cdot \mathcal{U}_{(0,11)}(\lambda_L),\end{aligned}$$

for belief-based and action-based models respectively, where $\Pr(c)$ is the hyperprior for the speaker’s preference parameter c . When $\Pr(c) = \delta(c)$ we have a hyperprior that does not take costs into account; otherwise we assume $\Pr(c) = \mathcal{U}_{(-0.4,0.4)}(c)$, as before.

Hyperpriors that capture the idea that the listener’s beliefs are good guesses of speaker behavior are modeled as if informed by the data from the speaker experiments:

$$\begin{aligned}\Pr(\lambda_S, c) &= \Pr(\lambda_S, c \mid D_S, M_S) \\ \Pr(\lambda_S, c, \lambda_L) &= \Pr(\lambda_S, c \mid D_S, M_S) \cdot \mathcal{U}_{(0,11)}(\lambda_L) .\end{aligned}$$

Here, D_S is the data from the speaker experiments, and M_S the relevant speaker model. For a given listener model, we consider only the embedded speaker model as relevant. We call hyperpriors of the above form “informed” or, in the case of action-based models, “informed uncorrelated”. We distinguish the latter from “informed correlated” hyperpriors of the form:

$$\Pr(\lambda_S, c, \lambda_L) = \Pr(\lambda_S, c \mid D_S, M_S) \cdot \Pr(\lambda_L \mid D_S, M_S),$$

where the listener’s rationality parameter is distributed according to the relevant posterior. All of the three types of informed hyperpriors were tested in two varieties: whether the listener takes the sender’s preferences into account or not. If he does not, the posterior $\Pr(\lambda_S, c \mid D_S, M_S)$ is derived from a speaker-hyperprior $\Pr(c) = \delta(c)$; otherwise from $\Pr(c) = \mathcal{U}_{(-0.4,0.4)}(c)$.

Taken together, we consider two belief-based models, paired with four hyperpriors, and two action-based models, paired with six hyperpriors (see Table 4). Comparing these ten Models is meant to address the following general questions:

1. Is the goal structure assumed by participants in our task belief-based or action-based?
2. Does the listener take the estimated salience prior into account or not?
3. Is the listener’s belief about the speaker’s level of rationality “correct”, i.e., in line with the observed speaker data?
4. Is the listener’s level of rationality related to the speaker’s actual level of rationality?

Answers to these questions can be found by comparing the evidences of the Models listed in Table 4.

Table 4 Log-evidences for listener models

prior type	Costs	$\rho_{bU}(\sigma_{bU})$	$\rho_{bS}(\sigma_{bU})$	$\rho_{aU}(\sigma_{aU})$	$\rho_{aS}(\sigma_{aU})$
Flat	No	-24.83	-11.66	-24.06	-10.62
Flat	Yes	-24.80	-9.84	-23.84	-10.81
Informed	No	-59.87	-10.44	-25.42	-8.48
Informed	Yes	-68.43	-13.35	-24.21	-11.85
Informed-correlated	No	NA	NA	-64.70	-6.80
Informed-correlated	Yes	NA	NA	-82.36	-29.01

The most striking contrast is that Models that do not take the salience prior into account fare much worse than those that do. The data makes a total rejection of Models that rely on uninformative salience priors strongly plausible.

Another eye-catching feature is that there is a clear winner in the list. The single best Model, which is substantially better than all the others, is action-oriented and takes perceptual salience into account. It has an informed-correlated hyperprior and assumes that the listener does not take the speaker’s preferences into account. This result is highly thought-provoking, but should be taken with a grain of salt. It might be that we merely got lucky by restricting the range of λ_S and λ_L just to a small region of reasonably high posterior likelihood for these parameters. Further experimentation should tell us whether informed-correlated hyperpriors are good predictors in general.

Most generally reliable are the results from flat hyperpriors. Here it is interesting to note that all of the models that take salience priors into account, with the exception of $\rho_{bU}(\sigma_{bS})$ are equally plausible, by common standards of Bayes factor comparisons. This is despite the fact that action-based models have an additional parameter.

Action-based models are also better at accommodating the idea that the listener’s estimate of the speaker behavior is roughly correct. The belief-based models are substantially or strongly less plausible in this case. In other words, action-based models, possibly due to the added flexibility of another parameter, are better at explaining production and comprehension data in unison.

This latter point is of interest with respect to the findings of Frank and Goodman (2012) that a single parameter choice $\lambda = 1$ provided a good fit for both production and comprehension data. The RSA model $\rho_{bU}(\sigma_{bS})$ with fixed parameter $\lambda = 1$ is a very poor predictor of our data. (We focus here on comprehension, but the same applies to production data.) Assuming $\lambda = 1$ as the null hypothesis of a nested-model comparison, we can use the Savage-Dickey method to compute a Bayes factor (Dickey and Lientz 1970; Wagenmakers et al. 2010). Let \mathcal{M} be the parameterized Model with model $\rho_{bU}(\sigma_{bS})$ and a flat hyperprior, not taking costs into account. Given our data we should adjust our beliefs in parameter value $\lambda = 1$ by the Bayes factor (computed approximately via MCMC sampling):

$$K = \frac{P(\lambda = 1 \mid D_L, \mathcal{M})}{P(\lambda = 1 \mid \mathcal{M})} \approx \frac{1.25e-9}{8.18e-2} = 1.52e-8$$

where D_L is the listener data. That means that our data provides very strong evidence that the null hypothesis $\lambda = 1$ is incorrect.

6 Conclusion

Our experiment and model comparison suggest the following. (1) On the conceptual level, it is helpful to clarify the distinction between belief, goal and action, and the Bayesian framework provides us with a natural way to structure these components into a formal model. By examining each component systematically, we can explicitly spell out the underlying assumptions and formulate alternative hypotheses to be tested. (2) On the empirical level, we tested the RSA model with a family of variants motivated from a game theoretic perspective on communication, and in particular, we investigated the speaker's preference as well as the listener's perceptual salience. Our data showed a more intricate picture of the roles that various factors play in pragmatic reasoning about referential expressions. We found evidence for a correlation between the speaker's and the listener's rationality. Either side appears to have his own biases, but appears to be negligent of the other's. Also, an action-based view might better reflect the goal of communication, at least in the forced-choice experimental setting that we used. Understanding the relation between the forced-choice and the betting paradigms would be an important next step to gain more insight into the differences between action- and belief-based goal structures.

The set of models we were able to compare explicitly here does clearly not exhaust the space of plausible candidates. For example, we excluded listener models in which the literal listener takes the salience prior into account. As an anonymous reviewer points out, this may seem paradoxical because the best listener models of our comparison do assume that the actual pragmatic listener takes salience into account, but does not believe that the speaker believes that he does. But actually, there is no friction here at all, because the listener model that is embedded in a speaker model is a different one from the full pragmatic listener model. If we think of the pragmatic listener as simulating his own hypothetical speaker behavior, there is nothing puzzling about factoring in salience only at a higher level of pragmatic inference.

A similar issue arises with respect to the rationality parameters in action-based listener models. We assumed that there are two rationality parameters, one for the listener's actual rationality and one for the listener's beliefs about the speaker's rationality. But we also saw that a model that correlates the set of joint values of these parameter pairs gave the best predictions of our data. This raises many interesting issues for further research, most of which we cannot answer here. The most obvious idea is, of course, to equate the speaker's and the listener's lambda and so that the listener would believe that the speaker is exactly as rational as the listener himself. Surprisingly, this model achieves a very poor fit on our data, despite its parsimony.

More research into correlations in perspective-taking and beliefs in rationality will hopefully shed more light on this fascinating issue.

In summary, our work shows that careful conceptual analysis of the design choices for quantitative models can lead to a better understanding of the phenomenon and further improvement in the formal model's predictive power. Of course, our restricted empirical data can only serve as a start and more data is needed to fuel further model comparison towards a more robust pragmatic theory of people's probabilistic reasoning about referential expressions.

Acknowledgments We are indebted to Judith Degen, Michael C. Frank, Noah D. Goodman and Daniel Lassiter, two anonymous reviewers, and the audience of the ESSLLI workshop "Bayesian Natural Language Semantics and Pragmatics" for stimulating feedback and discussion. Many thanks also to Henk Zeevat and Hans-Christian Schmitz for organizing mentioned workshop, and to Will Frager for help realizing our experiments. Michael Franke gratefully acknowledges financial support by NWO-VENI grant 275-80-004.

References

- Benz, A., & van Rooij, R. (2007). Optimal assertions, and what they implicate. A uniform game theoretic approach. *Topoi*, 26(1), 63–78. doi:[10.1007/s11245-006-9007-3](https://doi.org/10.1007/s11245-006-9007-3).
- Bergen, L., Goodman, N. D., & Levy, R. (2012). That's what she (could have) said: How alternative utterances affect language use. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 120–125). Austin: Cognitive Science Society.
- Degen, J., & Franke, M. (2012). Optimal reasoning about referential expressions. In S. Brown-Schmidt, J. Ginzburg, & S. Larsson (Eds.), *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue (SeineDial: SemDial 2012)* (pp. 2–11). Paris: France.
- Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, 41(1), 214–226.
- Franke, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084), 998. doi:[10.1126/science.1218633](https://doi.org/10.1126/science.1218633).
- Franke, M. C., Goodman, N. D., Lai, P., & Tenenbaum, J. B. (2009). Informative communication in word production and word learning. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1228–1233). Austin: Cognitive Science Society.
- Franke, M. (2011). Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics and Pragmatics*, 4(1), 1–82. doi:[10.3765/sp.4.1](https://doi.org/10.3765/sp.4.1).
- Franke, M., & Jäger, G. (2014). Pragmatic back-and-forth reasoning. In S. Pistoia Reda (Ed.), *Semantics, pragmatics and the case of scalar implicatures* (pp. 170–200). New York: Palgrave MacMillan. doi:[10.1057/9781137333285.0011](https://doi.org/10.1057/9781137333285.0011).
- Gatt, A., van Gompel, R. P. G., van Deemter, K., & Kramer, E. (2013). Are we Bayesian referring expression generators? In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 1228–1233). Austin: Cognitive Science Society.
- Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, 5(1), 173–184. doi:[10.1111/tops.12007](https://doi.org/10.1111/tops.12007).
- Grice, H. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41–58). New York: Academic Press.

- Jäger, G. (2013). Rationalizable signaling. *Erkenntnis*, 79(4), 673–706. doi:[10.1007/s10670-013-9462-3](https://doi.org/10.1007/s10670-013-9462-3).
- Jaynes, E. T. (2003). *Probability theory: the logic of science*. Cambridge: Cambridge University Press.
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford: Oxford University Press.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795. doi:[10.2307/2291091](https://doi.org/10.2307/2291091).
- Kramer, E., & van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1), 173–218. doi:[10.1162/coli_a_00088](https://doi.org/10.1162/coli_a_00088).
- Rabin, M. (1990). Communication between rational agents. *Journal of Economic Theory*, 51(1), 144–170. doi:[10.1016/0022-0531\(90\)90055-O](https://doi.org/10.1016/0022-0531(90)90055-O).
- Stalnaker, R. (2005). Saying and meaning, cheap talk and credibility. In A. Benz, G. Jäger, & R. van Rooij (Eds.), *Game theory and pragmatics* (pp. 83–100). New York: Palgrave MacMillan.
- Stiller, A., Goodman, N. D., & Frank, M. C. (2011). Ad-hoc scalar implicature in adults and children. In L. Carlson, C. Hölscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2134–2139). Austin: Cognitive Science Society.
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge: MIT Press.
- Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (in press). Model comparison and the principle of parsimony. In J. Busemeyer, J. Townsend, Z. J. Wang, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology*. Oxford: Oxford University Press.
- Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology*, 60(3), 158–189. doi:[10.1016/j.cogpsych.2009.12.001](https://doi.org/10.1016/j.cogpsych.2009.12.001).
- Worthy, D. A., Maddox, W. T., & Markman, A. B. (2008). Ratio and difference comparisons of expected reward in decision-making tasks. *Memory & Cognition*, 36(8), 1460–1469. doi:[10.3758/MC.36.8.1460](https://doi.org/10.3758/MC.36.8.1460).

Towards a Probabilistic Semantics for Vague Adjectives

Peter R. Sutton

Abstract A way of modelling the meanings of vague terms directly in terms of the uncertainties they give rise to is explored. These uncertainties are modelled with a probabilistic, Bayesian version of situation semantics on which meaning is captured as the correlations between uses of words and the types of situations those words are used to refer to. It is argued that doing so provides a framework from which vagueness arises naturally. It is also claimed that such a framework is in a position to better capture the boundarylessness of vague concepts.

Keywords Vagueness · Uncertainty · Probabilistic semantics · Adjectives · Bayesian · Probability · Situations · Situation theory

1 Introduction

A seemingly characteristic feature of vague terms is their *boundarylessness* (Sainsbury 1991). However, a challenge for truth-conditional accounts of the semantics for vague terms is that truth-conditions create sharp boundaries. One way to try to assuage this tension is to introduce uncertainty into semantics. For example, one could maintain a threshold for truth, but model vagueness as uncertainty about where the threshold, in some context, lies (Lassiter 2011; Frazee and Beaver 2010). However, arguably, uncertainty over thresholds nonetheless carves boundaries. In this paper, I will avoid thresholds and develop a way of introducing uncertainty directly into semantic representations. On this approach, sentences are taken to encode uncertainty over types of situations. This can be seen from two perspectives. Either as uncertainty about how the world is, given the way words are used, or as uncertainty about how to use words when faced with a type of situation to describe. The information that words convey is understood as a reflection of the correlations there

P.R. Sutton (✉)

Institute for Language and Information, Heinrich Heine University,
Universitätsstraße 1, 40225 Düsseldorf, Germany
e-mail: peter.r.sutton@icloud.com

are between features or properties in the world, and our uses of words to describe or refer to objects with those properties. By adopting this view, I will suggest that we can better capture the boundarylessness of terms, and that borderline cases, another characteristic feature of vagueness, emerge directly from the semantics.

In Sects. 2–3 the uncertainty associated with vague terms will be described. In Sects. 4–5, these informal ideas will be modelled using a probabilistic version of situation semantics. In Sect. 6, the approach will be linked to vagueness. In Sect. 7, the proposal will be compared with two related proposals in the recent literature.

2 Uncertainty

There are two ways that uses of a vague term can give rise to uncertainty. The first concerns the world and the second concerns the word. Say that we are told that John is tall, or that Mary's car is green. Given that we are competent in English, and, given that we have no reason to expect what we have been told to be incorrect, what we have been told makes it reasonable to believe certain things about the world. In the vagueness literature, when the semantics of vague terms are analysed, some less vague properties, concepts, or features are often cited in the analysis. For example, with respect to 'green', our judgements vary over the shades that things are; for 'tall', over the heights that things are; for 'bald', over how much (head) hair individuals have. On being told that John is tall, or that Mary's car is green, we are uncertain about what feature or property John has with respect to height, or that Mary's car has with respect to shade.

This uncertainty could be thought of as a series of graded beliefs. In the following, I will represent the content of these beliefs with reference to measures that may not have any psychological reality. For example, I will talk about someone's beliefs about heights measured in centimetres. However, it need not be assumed that our doxastic representations of heights refer to centimetres. If I talk of Mary's beliefs about John's height as distributing over heights in centimetres, this is merely a convenient notation. It does not require that Mary can say, in centimetres, how tall John is (or might be).

It is, however, possible that using such precise values introduces an artificial level of precision. It could be that the way we cognitively represent the world is also vague, in which case, there should be a mapping from vague words to a mapping from a vague representational level to properties in the world (such as being some height in centimetres or being some particular shade). However, by describing mappings between vague words and less or non-vague properties in the world directly, I am, at worst, oversimplifying the matter by skipping over a mental/cognitive level of representation. For at least some cases we ought to be able to describe how the information conveyed by words relates to the (more or less) precise properties to be found in situations in the world.

It would be unreasonable to believe that John is a specific height just on the basis of being told that he is tall, and it would be unreasonable to believe Mary's car to be

a specific shade just on the basis of being told it is green. Herein lies the first kind of uncertainty:

- (U1) Descriptions using vague expressions leave us uncertain about specific features/properties that objects in the world have.

U1 uncertainty is often, to some extent, eliminable in practice. We can, for example, go and look at John (maybe even measure him), or we can go and have a look at Mary's car. And yet, even if we are in command of these facts, uncertainty about how to use our language can still remain. If John is 180cm tall, or if Mary's car is an odd sort of turquoise shade, we might feel thoroughly uncertain whether or not 'tall' would be an effective word to use to describe John or 'green' an effective word to use to describe Mary's car.

By using the term 'effective', I mean to appeal to a notion of success. What is effective (for some purpose) is what succeeds in doing something/bringing about the desired result. What is effective for describing John, or Mary's car is going to be an interest relative matter, since what is effective for doing something in one situation may not be effective in another. A lot of the vagueness literature, insofar as it predominantly concerns solutions to the sorites paradox, takes the only important criterion for what is effective to be what description would be true. However, truth is not always what we need to establish, or are interested in. For example, if our aim is to communicate which individual John is in a crowd, irrespective of whatever story about the truth-conditions for 'tall' are, one might be able to use 'tall' to identify John because he is significantly taller than those around him. Here, what makes a description count as true will not be addressed in detail. Instead, I will continue to use the broader notion of *effective descriptions* which can be used to describe a second kind of uncertainty:

- (U2) Specific features/properties that objects in the world have can leave us uncertain about how to effectively describe them.

Relations hold between U1 and U2:

1. Were we equally certain of applying 'tall' to 185 cm John as to 190 cm Bill, then, all else being equal, we should be equally certain that John (Bill) is 185 cm in height when described as tall as 190 cm, and *vice versa*
2. Given 1, where U1 similarities give rise to U2 similarities and *vice versa*, we should expect U1 differences to give rise to U2 differences and *vice versa*.
3. Variations in U1 uncertainty can give rise to U2 uncertainty. Being told that John is tall may leave us uncertain about his height. We might be fairly certain that he is not around 170 cm, comparatively certain that he is around 190 cm, but highly uncertain about whether he is around 180 cm. Given 1 and 2, this creates U2 uncertainty in the use of 'tall'. If we are fairly certain that 'tall' effectively describes someone who is around 190 cm in height, but fairly certain that 'tall' wouldn't effectively describe someone around 170 cm in height, then we should be uncertain that 'tall' would effectively describe someone around 180 cm. Furthermore, we should also expect a range of U2 judgements to give rise to U1 uncertainty.

As emphasised by an anonymous reviewer, the presence of U1 uncertainty is not sufficient for vagueness, and the presence of U2 uncertainty is not necessary for vagueness. The relationship between vagueness and different kinds of uncertainty will be made clear in Sect. 6. In brief, however, there will be times when, given the way the uses of words correlate with types of situations, nothing in the meanings of the words themselves, or any outside information, will be able to resolve our uncertainty. For example, for some heights of people, we may have irresolvable uncertainty about whether to call them tall. This can arise because when someone is described as tall, we may be no more certain that they are that height than we would be were they to be described as not tall.

Two challenges arise for communicating with vague expressions. As a hearer, we must face uncertainty over what objects in the world are like, given the way they have been described. As speakers, we may, from time to time, face uncertainty over how to effectively describe things as being. However, arguably this uncertainty stems directly from the uses of such terms, and from how such uses correlate with the properties of situations those words describe or refer to. In the following, I will treat the strengths of these correlations *as* the information that those terms carry.¹ This information will be modelled using Bayesian conditional probabilities. It is in this sense that Bayesian notions will be at the heart of the semantics for vague terms.

3 Constituents

A semantics based on uncertainty will have to be able to attribute, to types of expressions, a role in the larger constructions in which they occur. Take our above examples: ‘John is tall’ and ‘Mary’s car is green’. Giving a semantics for the modifiers ‘tall’ and ‘green’ can then be taken to be a matter of accounting for what information they carry about some object. For example, ‘ x is green’ might be modelled as making it reasonable to believe that x is one of some rough range of shades.

However, here a disparity between ‘tall’ and ‘green’ arises. The x in ‘ x is tall’ is vital for getting any idea about what it is reasonable to believe. If we are only told that something, anything, is tall, be it a mountain, a molehill, a mouse, or a millipede, there is no height it would be more reasonable to believe this thing is than others.² This changes as soon as ‘tall’ is predicated of an NP or is applied to a CN in an NP: The ‘tall’ in ‘ x is a tall man’ seems to make it more reasonable to believe that x is some heights rather than others. But this suggests that part of the information which contributes to our expectations of heights is coming from the CN/predicate ‘man’. This can be seen by substitution of CNs/predicates.³ Compare how expectations of

¹This notion of semantic information is more rigorously developed in Sutton (2013).

²This may be too strong. There could be some priors we have, or reasoning we could engage in, that would make some heights more reasonable than others. My thanks go to Noah Goodman for helpful discussion.

³I do not mean to take, as assumed, a clear, semantically important division between CNs and adjectives. Below, I will treat words like ‘tall’ and ‘green’ as predicate modifiers and words like

heights differ for ‘*x* is a tall man’, ‘*x* is a tall molehill’ and ‘*x* is a tall mountain’. The situation for green seems to be similar. CNs make a difference to expectations.⁴ However, ‘*x* is green’ does seem to carry some expectations of its own: that *x* is roughly within some range of shades.

I take this to be a reason to see some differences in the semantics of various vague adjectives/modifiers. The modifier ‘green’, which seems to carry information on its own, will have a slightly different semantic shape to those modifiers (such as ‘tall’) that only seem to modify expectations based on information carried by nominal predicates. The details of this difference will be elaborated in Sect. 5.

What we have here is an argument for treating some modifiers as, in some sense, not really carrying information *per se*. Instead we can see them as encoding a modification on the information carried by a nominal predicate. Someone’s being described as a woman gives us a rough expectation as to her height. This may be a very broad range, but we have some expectations nonetheless. Something’s being described as a skyscraper gives us very different expectations as to its height.

One might worry that this puts a lot of weight on information carried by nominal predicates and/or background information and beliefs. However, all that is being assumed is that learning to categorise and classify objects with such predicates, in part, amounts to developing expectations as to basic visual cues such as size, shape and shade. An anonymous reviewer rightly points out that this approach begins to blur the boundaries between what counts as meaning and what counts as general knowledge about the world. A firmer distinction can be established again without overly affecting the account, however. Rather than seeing the meaning of, say, ‘man’ as being all the information one has learnt about men, we can view meanings as procedures (see Gargett et al. 2009) for accessing such knowledge and for determining the ways such knowledge should be combined.

The common modifiers ‘old’, ‘big’ and ‘long’, just like ‘tall’, only seem to give us reasonable expectations when applied to a nominal predicate (or when themselves predicated of an NP in cases such as ‘John is tall’). In turn, however, that suggests that nominal predicates such as ‘man’ carry information about what features men can reasonably be expected to have. Some tests (albeit ones based on linguistic intuitions) can be applied to get a grasp on what information this is. For example, at least for non-metaphorical uses, ‘*x* is a long man’ does not seem to make much sense/provide us with reasonable expectations about *x*’s length.⁵ This suggests that ‘man’ does not carry information pertaining to length. This is not to say that modifiers such as

(Footnote 3 continued)

‘car’ and ‘man’ as predicates. In many cases, the traditional classifications of words as CNs and adjectives will overlap with these semantic divisions. Nothing I say rests on whether they all do.

⁴This point is related to the commonly held idea that vague adjectives are context-sensitive or interpretable only relative to a comparison class (Cresswell 1974).

⁵In English at least, length is not simply a euphemism for height. Trains, passageways, halls, to name a few, can have a great length without having a great height.

‘green’ are not restricted either. For example, synaesthesia aside, ‘*x* has green ideas’ is hard to understand if ‘green’ relates to shade.⁶

Nominal predicates will be treated as encoding information based on correlations between things that those common nouns are used to classify and the features that the things they are used to classify have. For example, it is because there is a stronger correlation between men and some heights than there is with others, that ‘man’ carries information about heights. Adjectives will be modelled as modulating some specific aspect of the information that nominal predicates carry. For example, ‘tall’ will modify information relating to height carried by predicates like ‘man’.

A possible worry can be flagged here.⁷ In the literature on adjectives, it is common to appeal to a comparison class to capture their semantics (Cresswell 1974; Kamp 1975; Kennedy 1997; Graff Fara 2000). However, it has been pointed out that common nouns do not always determine comparison classes (Kennedy 2007). For example, BMWs are not the comparison class below, despite featuring as part of a modified NP:

1. Kyle’s car is an expensive BMW, though it’s not expensive for a BMW. In fact it’s the least expensive model they make. (Kennedy 2007)

The worry concerns whether the above analysis conflates modified nouns with comparison classes. The analysis can, however be modified slightly. Rather than demand of some adjectives that they simply modify a nominal predicate, we can instead require that there is some class of things to which it is being applied in any context (information about which can be modified by the adjective). This still leaves a difference with terms like ‘red’ which seem to carry some information (about, say, shades) independently of a lexically provided nominal predicate or a comparison class.⁸

4 Semantics: Preliminaries

4.1 *Correlations and Situations*

In the following, I will suggest one way to formalise U1 uncertainty (uncertainty over how the world is, given a description of it). I will borrow heavily from situation semantics (Barwise and Perry 1983; Cooper 1996; Devlin 2006), but will incorporate a probabilistic element into the standard categorical theory.

⁶An interesting possibility is that the application of domain specific modifiers to NPs that do not carry information about that domain will generate metaphorical interpretations. Metaphor, humour and other such subjects are outside of the scope of this paper. There may also be an interesting link to be explored between the ideas put forward here and work done on scalar adjectives in de Marnee et al. (2010).

⁷This was pointed out by an anonymous reviewer for the BNLSP 2013 Workshop.

⁸That is not to say that ‘red’ cannot relate to non-stereotypical shades (such as in ‘red onion’). Elsewhere (Sutton 2013), I suggest a way for this account to model the information carried by ‘pet fish’ where goldfish are not the most stereotypical fish, nor the most stereotypical pets.

In situation theory, meaning is captured via the notion of constraints. Constraints represent information channels, for example, types of situations in which there is rain in the city are informationally connected to types of situation in which the pavements are wet. The (context independent) linguistic meaning of an expression is held to be just a special case of a normal information channel: types of situation in which some expression is uttered are informationally related to types of situations in which some conditions obtain. For example, types of situations in which someone utters ‘It’s raining in London’ are informationally connected to types of situations in which there is rain in (at least part of) London (at some time).

It is via this type-type link or relation that agents are able to extract token-token information. For the case in hand, if one is in a token discourse situation of the type ‘*It’s raining in London*’ is uttered, one will be led to expect a token situation of the type *raining in London*. In standard situation semantics, ignoring a lot of details, the truth of what is said turns on whether the token described situation is of the right type.

The basic idea to be presented in this paper is that meaning can be treated as correlational. Instead of constraints holding between one utterance situation type and one described situation type, probability theory will be employed to capture different strengths of connections between an utterance situation type, and many described situation types. The meanings of expressions are therefore held to be correlations between discourse situations of a certain type, and described situations of a certain type. For example, the type of situation in which ‘John is tall’ is uttered will be correlated, to different extents, with types of situation in which John is one of many possible heights.

In Situation Theory, situations support *infons*. Infons are traditionally conceived as what contribute to forming informational items (Devlin 2006), or as types of situations (Cooper 1996).⁹ Another way to view them is as properties of situations (as opposed to properties of individuals). For example, (ignoring times) one property that a situation might have is where it’s raining somewhere¹⁰:

$$\langle\langle \text{rain}, \dot{l}, \text{yes} \rangle\rangle$$

The \dot{l} is a *location parameter*. Parameters can be seen as akin to variables. They can be bound via type abstraction (see below), or free. Ignoring quantifiers, free parameters are those not bound by abstraction in situation theoretic propositions. If a situation, s has some property (some infon), σ , then it is said that the situation

⁹This way of viewing infons (as types) propagates through into situation theoretic approaches with richer type systems. See, for example Cooper (2012).

¹⁰This notation for infons is essentially Devlin’s. However, for polarities, I adopt Barwise and Perry’s ‘yes’ and ‘no’ instead of Devlin’s ‘1’, ‘0’. This is to avoid potential confusion with the limit cases of probability values [0, 1].

supports the infon. That a situation supports an infon, in notation, is a situation theoretic proposition:

$$s \models \sigma$$

For the above infon, this would give:

$$s \models \langle\langle \text{rain}, \dot{l}, \text{yes} \rangle\rangle$$

Which says that in situation s , it is raining at some location. Situations are meant to be ways to classify objects and events in the world. They are not possible worlds. If a situation, in fact, has the property σ (is, in fact, of type σ /supports σ), then the proposition is true. However, we will not be seeking to associate linguistic expressions with propositions, nor will we be directly interested in particular situations and whether they support some infon. This is because constraints are defined between situation types.

Importantly, abstraction in situation theory allows one to talk about types. Abstracting over objects (understood loosely as individuals, locations, and times), creates *object types*. Taking our rain example, and abstracting over \dot{l} , this gives:

$$\lambda[\dot{l}](s \models \langle\langle \text{rain}, \dot{l}, \text{yes} \rangle\rangle)$$

Which is the type of locations in which it rains in situation s . When bound, parameters can be replaced with constants as with standard β -reduction. For example:

$$\begin{aligned} &\lambda[\dot{l}](s \models \langle\langle \text{rain}, \dot{l}, \text{yes} \rangle\rangle) . [\text{London}] \\ \Rightarrow &s \models \langle\langle \text{rain}, \text{London}, \text{yes} \rangle\rangle \end{aligned}$$

Situations can be abstracted over to give *situation types*¹¹ via *situation type abstraction*. Notation varies, I will adopt the following¹²:

$$\lambda[\dot{s}](\dot{s} \models \sigma)$$

This means the type of situation in which σ obtains. The \dot{s} is a parameter, as \dot{l} above, but \dot{s} is a parameter for situations. In this case, \dot{s} is bound. For example:

$$\lambda[\dot{s}](\dot{s} \models \langle\langle \text{rain}, \text{London}, \text{yes} \rangle\rangle)$$

Which is the type of situation in which it rains in London. Correlations (information channels) hold between types. For example the above type might correlate with the

¹¹More recent situation theoretic approaches take types to be objects. An example of this rich type-theoretic approach is Type Theory with Records (TTR) (Cooper 2012).

¹²This is close to Devlin's notation, but with Cooper's use of ' λ ' for abstractions instead of Devlin's ($\dot{s}|\dot{s} \models \sigma$). This variation is to avoid possible confusion resulting from the use of ' $|$ ' in probability theory.

type of situation in which the streets of London are wet. In the probabilistic version of Situation Theory to be developed, these correlations will be modelled as conditional probabilities (the probability of one type, given another). It is this that will ensure that if some concrete situation of a type arises, there will be a probability that some situation of another type will arise (or, arose, or will have arisen etc.).

The proposal to be made here is then that U1 uncertainty (where descriptions using vague expressions leave us uncertain about specific features/properties that objects in the world have), can be modelled as conditional probabilities between types of situation. I will say more on U2 probability in Sect. 7.

We will have situations of two kinds:

- (i) *Discourse situations*, \mathfrak{d} , are situations in which some specific discourse takes place. For example, we might have the discourse situation in which ‘red’ is used to classify some object, or, in which ‘tall’ is used to describe some person. Discourse situations will model the situations in which certain words are used (certain things are said).
- (ii) *Described situations*, \mathfrak{s} , are situations in which the world is some way. For example, we might have the described situation in which an object is some shade of colour, or, in which a person is 180cm in height. Described situations will model the kinds of ways the world can be.

Types can be formed by abstracting over discourse and described situations. For example:

$$\lambda[\dot{d}](\dot{d} \models \langle\langle \text{utters}, \dot{a}, \text{TALL}, j, \text{yes} \rangle\rangle)$$

Is the type of situation in which someone describes John as ‘tall’. Given such a type, there will be a probability of some described situation being of some type. For example, the type of described situation in which John is 180cm in height:

$$\lambda[\dot{s}](\dot{s} \models \langle\langle \text{height}=180\text{cm}, j, \text{yes} \rangle\rangle)$$

Varying probabilities will result from different height values. These probabilities will form a distribution. Abstracting away from the contribution made by the name ‘John’, this distribution will reflect the extent to which uses of the modifier ‘tall’ to describe, say, a human male, correlates with human males being certain heights. The heights that an agent will entertain will be restricted by her/his learning process (namely, an approximation over the kinds of humans, males etc. that they have experienced). It will be supposed that a rational agent can only have distributions that sum to 1.

The information carried by sentences will be captured as the conditional probability of there being a described situation of some type, given a discourse situation of some type. For example, the probability of there being a situation type in which John is 5ft, given a situation type in which John is described as tall, will be (comparatively) very low.¹³ Other conditional probabilities will be greater for greater height

¹³This can be so even given the assumption that speakers are not being deliberately deceptive. What is being tracked here is the extent to which, in general, properties such as heights of individuals correlate with types of utterances, such as describing them as tall.

properties John might have. A range of such probabilities (each with a different height property/one-place relation assigned to John in the described situation) will form a distribution (must sum to one).

Conditional probability values will represent something like credence. However, rather than just a degree of belief, we will be concerned with the degree to which it is reasonable to believe that some state of affairs obtains. This distinction is between what, given some actual description, a hearer believes (credence), and what, in virtue of the information carried by the utterance of the words, one is entitled to believe. Conditional probability values reflect correlations between uses of words and states of affairs. This incorporates all uses, both representations and misrepresentations of the world. The degree to which it is reasonable that some state of affairs obtains reflects this. How much reason or entitlement one has to believe the world to be some way will be a direct reflection of how words are used (how uses of words correlate with types of situations).

So, if being told that John is tall makes it reasonable to believe to a degree of n that he is 180 cm in height, this will be represented as assigning a probability value of n to the conditional probability of there being a situation of the type in which John is 180 cm in height, given a situation of the type in which he is described as tall. The value for n , against another larger number, represents a lower reasonable credence in a state of affairs than the higher number. Conditional probability distributions will be formed over a range of heights.

Types and Situation Types A more recent approach to situation theoretic semantics is systematically laid out in Cooper (2012) which adopts a rich proof theoretic type theory. Rather than treating types as sets of objects in a domain, this approach treats types as objects (in a domain of types). Situations/events then act as proofs of propositions. In what is to follow, I will adopt a fairly simple type theory with an, essentially, model theoretic semantics.¹⁴ This simplification will make it easier to focus on the probabilistic element in my account. A proof theoretic approach with a richer type theory could, in principle, adopt the core ideas in the proposal that I will make. Indeed, a probabilistic approach comparable to my own has just been proposed in Cooper et al. (2014).¹⁵

Properties It is important to note that it is not assumed that there is anything like the property *green* or the property *tall* that individuals have or lack. We will be assuming that individuals can have the property of being some height (this will just be the height that they are). We will also be assuming that certain objects have some hue or shade.¹⁶ These heights and shade properties will be mentioned in the formalism as properties of objects in infons.

¹⁴The hedging here is included because, although I assume a domain of situations (as a set), situations are best not viewed as a set.

¹⁵A difference between this, my own, and similar proposals will be explained in Sect. 7.

¹⁶I assume for simplicity that these properties are not vague. Nothing turns on this however. Even if properties were vague in the sense that someone may genuinely be of indeterminate (exact) height, one would still need to explain how words like ‘tall’ admit of wide ranges of those properties in a graded way.

4.2 Definitions

Types The type system that will be used will have three basic types.¹⁷ Type e will be the standard type for individuals (entities). Type s will be the type for situation. Rather than the traditional type t (for truth value), we will have type p (for probability). Basic types can be used to form functional types in the way normal to model theory.

Definition 1 Types

Basic types: $\{e, p, s\}$
 Functional types: If a, b are types, then $\langle a, b \rangle$ is a type

Expressions This formalism brings together notions from both probability theory and situation theory. As such, we will need to define how various parts of them interact. The vocabulary of the formalism is in Definition 2. Where necessary, types of expressions will be given as subscripts. For example, a variable, S , of type $\langle e, p \rangle$ will be written $S_{\langle e, p \rangle}$. The marking of numerical variables as type p in Definition 2 is an oversimplification. These will be interpreted as numerical normalising values. These values can have a value outside of $[0, 1]$, but are derivable from sums of values of interpretations of other type p expressions.

Definition 2 Vocabulary

Parameters (PAR):	Infinite set of parameters for all types $\dot{s}_1, \dots, \dot{d}_1, \dots, \dot{x}, \dots$
Constants (CON):	A possibly empty set of constants for every type
Supports:	\models
Conditional on:	$ $
Infons:	$\sigma, \tau, \langle \langle \text{height}=h, j, \text{yes} \rangle \rangle, \langle \langle \text{utters}, a, \text{TALL}, j, \text{no} \rangle \rangle$
Connectives:	\wedge, \neg, \vee
Lambda abstraction operator:	λ
Mathematical operators:	$\times, +, -$
Numerical variables:	\mathbf{C}, \mathbf{C}' of type $\langle p \rangle$
Brackets:	$(,), [,]$

A further few remarks on infons are needed at this juncture. Infons contain relations between objects and the polarities in infons ‘*yes*’ and ‘*no*’ indicate whether the relation holds between the objects. I assume a basic ontology of relations (such as being some height or being some shade). Propositions state that situations support infons. Positive polarities on infons indicate that any situation that supports the

¹⁷I will from now on suppress all reference to locations and times, which are also types in situation theory.

infon is of that type. A negative polarity would indicate that the situation is of the negative type. Infons can thus be understood as types of situations (Cooper 1996), or perhaps as properties of situations. The class of infons will be formed of all possible combinations of relations and objects in our ontology (plus a polarity).

The well-formed expressions of the language are defined in Definition 3 (where WE_a denotes well formed expressions of type a). Clauses (i), (ii), (iv) and (v) are standard. Clause (v) introduces the syntax for situation types. Clause (vi) gives what will eventually be interpreted as a conditional probability of one situation type given another. The connectives in clause (vii) will be discussed in greater detail below. Clause (viii) allows for expressions of type p to be linked by mathematical connectives.

Definition 3 Well-Formed Expressions:

- (i) If $\alpha \in PAR_a$, then $\alpha \in WE_a$,
- (ii) If $\alpha \in CON_a$, then $\alpha \in WE_a$,
- (iii) If $\alpha \in WE_{(a,b)}$ and $\beta \in WE_a$, then $\alpha(\beta) \in WE_b$,
- (iv) If $\alpha \in WE_b$, and $\beta \in WE_a$, then $\lambda[\beta](\alpha) \in WE_{(a,b)}$,
- (v) If $\alpha \in WE_s$, and if σ is an infon, then $\lambda[\alpha](\alpha \models \sigma) \in WE_p$,
- (vi) If $\phi, \psi \in WE_p$, then $\phi \mid \psi \in WE_p$,
- (vii) If $\phi, \psi \in WE_p$, then:
 - (a) $\neg\phi \in WE_p$,
 - (b) $\phi \wedge \psi \in WE_p$,
 - (c) $\phi \vee \psi \in WE_p$,
- (viii) If $\phi, \psi \in WE_p$, then
 - (a) $\phi \times \psi \in WE_p$,
 - (b) $\phi + \psi \in WE_p$,
 - (c) $\phi - \psi \in WE_p$,

As mentioned briefly earlier, there are two kinds of situation parameters that we will use: *described situations* (\dot{s}) and *discourse situations* (\dot{d}). For descriptive situation types, we will be interested in how the world is with respect to things like the shades things are (for colour terms), and the heights things are (for terms like ‘tall’). In discourse situation types, what will be described is the production of some utterance type.

The infons we will be concerned with are likewise of two kinds. Infons for descriptive situations will describe/be about, say, what height someone is. Immediately below I specify the infon, and below that, I give a described situation type including that infon:

$$\langle\langle \text{height}=180\text{cm}, j, \text{yes} \rangle\rangle$$

$$\lambda[\dot{s}](\dot{s} \models \langle\langle \text{height}=180\text{cm}, j, \text{yes} \rangle\rangle)$$

The latter is the type of situation in which the individual denoted by j (say, John) is 180cm in height. Infons for descriptive situation types will typically relate to

what has been said. For example, below, I specify a discourse infon, followed by a discourse situation type including that infon:

$$\langle\langle \text{utters}, \dot{a}, \text{TALL}, j, \text{yes} \rangle\rangle$$

$$\lambda[\dot{d}](\dot{d} \models \langle\langle \text{utters}, \dot{a}, \text{TALL}, j, \text{yes} \rangle\rangle)$$

The latter is the type of situation in which the individual denoted by j (say, John) has been described as tall. This is a simplification, however. In actual fact, there ought to be unbound parameters for locations and times etc. which are being suppressed for simplicity.

For either of these situation types (descriptive and discourse), many actual situations could be of those types (types where John is 180cm in height, or where he is described using ‘tall’).

Interpretations and Domains The interpretation function we shall use will be marked as: $p(\cdot)$. Hence, an expression ϕ of type p , will be interpreted as $p(\phi)$, or the probability of ϕ . Like interpretation functions in model theory, the interpretations of type p expressions will be built up of the interpretations of other expressions. I assume a domain for each basic type. If a is a basic type:

$$p(\alpha_{(a)}) \in \text{Dom}_{(a)}$$

Functionally typed expressions will be interpreted as functions from basic domains:

$$p(\alpha_{(a,b)}) \in \text{Dom}_{(b)}^{\text{Dom}_{(a)}}$$

The type e domain will be a set of individuals. The domain of type p will be the range $[0, 1]$. Situations are being taken as primitive. We can, for the sake of the formalism, assume a domain of situations. However, in the spirit of situation theory, situations should be taken to be our ways of conceptualising and carving up the world into parts. As well as the interpretation function, we will have a parameter anchoring function g , which assigns appropriate members of the domain to parameters. In situation theory, parameter anchoring functions have restrictions on their domains, I will pass over those details here.

The first four clauses of Definition 4 are fairly standard. Clause (v) simply states that maths operators in the object language are treated as standard when the interpretations of their operands are in the range $[0, 1]$. Clause (vi) simply states the standard definitions of probabilistic connectives (Kolmogorov 1950).

Spaces and Priors In order for these axioms and the probability function to be well defined, we must define a probability space. Probabilities will distribute over situation types. The probability value will be in the range $[0, 1]$ where this value indicates the probability of some situation being of the type specified by the infon. Conditional probabilities will be the probability of a situation being of some type, given that some situation is of another type. For simplicity, for all non-conditional

values, I assume that probabilities distribute evenly.¹⁸ For example, in the case where we have n discrete ranges of heights over which a distribution will be formed, the values for each unconditional situation type will be:

$$p(\lambda[\dot{s}] (\dot{s} \models \langle\langle \text{height} = h, \dot{x}, \text{yes} \rangle\rangle)^g) = \frac{1}{n}$$

Definition 4 Denotations

- (i) $p(\dot{x})^g = g(x)$ if $\dot{x} \in PAR$,
- (ii) $p(c)^g = p(c)$ if $c \in CON$,
- (iii) $p(\alpha(\beta))^g = p(\alpha)^g(p(\beta)^g)$,
- (iv) $p(\lambda[\dot{x}](\alpha))^g = f$ such that if $\dot{x} \in PAR_a$, and $c \in Dom_a$, $f(c) = p(\alpha)^g[\dot{x}:=c]$,
- (v) Maths Operators:

- (a) $p(\phi \times \psi)^g = p(\phi)^g \times p(\psi)^g$,
- (b) $p(\phi + \psi)^g = p(\phi)^g + p(\psi)^g$,
- (c) $p(\phi - \psi)^g = p(\phi)^g - p(\psi)^g$,

(vi) Probabilistic Connectives:

- (a) $p(\neg\phi)^g = 1 - p(\phi)^g$,
- (b) $p(\phi \wedge \psi)^g = p(\phi)^g \times p(\psi | \phi)^g$,
- (c) $p(\phi \vee \psi)^g = p(\phi)^g + p(\psi)^g - p(\phi \wedge \psi)^g$.

In practice, the range of situations and infons will be highly constrained. For described situations, constraints will come out of the semantic learning process, as well as the goals and purposes of the speakers. For example, certain ranges of shade/hue properties will constrain the range of described situations for colour terms in general, but further constraints may come from what objects are salient to speaker and hearer in the situation of utterance.

For discourse situations, I assume a space of two infons, formed of the same relation and object but with opposing polarities.

$$p(\lambda[\dot{d}] (\dot{d} \models \langle\langle R, a_1, \dots, a_n, \text{yes} \rangle\rangle)^g) = 0.5 = p(\lambda[\dot{d}] (\dot{d} \models \langle\langle R, a_1, \dots, a_n, \text{no} \rangle\rangle)^g)$$

However, when multiple terms are used, the space of possibilities may be bigger and so values may be lower.

Representationalism The formal semantics and linguistics debate over representationalism got going with the publication of work on Discourse Representation Theory (DRT) (for an overview, see Kamp and Reyle 1993). DRT questioned Montague's claim that the representational language of a formalism is dispensable. In DRT, it was claimed that to account for the treatment of some expressions (such as discourse anaphora), the representational (interpretational) level is not dispensable.

¹⁸In a more sophisticated model, one would wish the values of priors to be set in accordance with the learning experiences of agents. One way this has been implemented will be discussed in Sect. 7.2.

Non-representational, dynamic accounts were developed in response. The debate was never entirely settled, but I do not need to take up a position.

Linguistic Meaning In standard situation semantics, linguistic meaning is captured by way of *constraints*. For example, the linguistic meaning of ‘Mary is tall’ will be a link between an utterance situation type in which an utterance of ‘Mary is tall’ is made, and a described situation type in which Mary is tall. The introduction of probabilities into the situation theoretic framework, can be understood as the loosening of the strength of this link. Rather than taking an utterance of ‘Mary is tall’ to convey information in such a way as to give us the expectation that in some described situation Mary is tall, probabilistic constraints will hold between utterance situation types and a number of described situation types. In each of the described situation types, Mary will be some particular height, however, the strength with which we should expect Mary to be that height, given that she has been described as tall will vary. This uncertainty will be captured formally as a conditional probability distribution. For the example in hand this will be the following:

$$p(\lambda[s](s \models \langle\langle \text{height}=h, m, \text{yes} \rangle\rangle) \mid \lambda[d](d \models \langle\langle \text{Utters}, \dot{a}, \text{TALL}, m, \text{yes} \rangle\rangle))^g$$

The above formula will return probability values for different values of heights h . Put simply, given a type of situation in which Mary is described as tall, one should assign a greater probability to the described situation being one in which Mary is some heights rather than others. However, these values will be determined by more general correlations in which, say, people are described as tall and where people are some height or other (the details of this will emerge in the compositional semantics). The standard way we will represent and interpret declarative sentences will therefore be as a conditional probability formed with a described situation type and a discourse

Table 1 Interpretation of ‘Mary is tall’

h (cm)	$p(\lambda[s](s \models \langle\langle \text{height}=h, m, \text{yes} \rangle\rangle) \mid \lambda[d](d \models \langle\langle \text{Utters}, \dot{a}, \text{TALL}, m, \text{yes} \rangle\rangle))$
$h < 150$	0.01
$150 \leq h < 155$	0.01
$155 \leq h < 160$	0.01
$160 \leq h < 165$	0.01
$165 \leq h < 170$	0.01
$170 \leq h < 175$	0.02
$175 \leq h < 180$	0.06
$180 \leq h < 185$	0.11
$185 \leq h < 190$	0.22
$190 \leq h < 195$	0.32
$h > 195$	0.22

situation type where the resultant probability distribution describes the constraints that the words used in the utterance situation place on the described situation:

$$p(\lambda[\dot{s}](\dot{s} \models \tau) \mid \lambda[\dot{d}](\dot{d} \models \sigma))^g$$

For different infons substituted for τ , values will be in the range $[0, 1]$. These values will form a distribution. Returning to our example of ‘Mary is tall’, using some made-up values, the sentence could be interpreted as in Table 1. In Sect. 5, I will describe how such a sentence can be composed.

5 Semantics: Terms

5.1 Predicates

Since nominal predicates were argued to be important for the semantics of vague adjectives/modifiers (as the things bearing the information that they modify), we will now turn to them. As suggested in Sect. 3, nominal predicates may carry a lot of different information. Our basic semantic representation for such predicates will incorporate an argument the domain of which will be a selection of a type of information (such as a range of heights). This information type will be supplied either by context, or by a modifier, such as an adjective. In a departure from a simple view of predicates (that take just an object as an argument), nominal predicates will be modelled as a function from properties to a function from an individual to a probability, which is to say that they will be a function from properties to properties $((e, p), \langle e, p \rangle)$. The logical structure of nominal expressions will be such that two entities in the above will be the same. Put another way, singular nominal predicates are assumed to require updating in context with respect to the aspect of the information they carry (via an argument of type $\langle e, p \rangle$). Then, when provided with an individual (a type e argument), yield a value for the probability of that individual having that property, given the information provided. The schema for ‘man’ (and with appropriate substitution, other predicates) is:

$$\text{Man} : \lambda[\dot{S}](\lambda[\dot{x}](\dot{S} . [\dot{x}] \mid \lambda[\dot{d}](\dot{d} \models \langle\langle \text{Utters}, \dot{a}, \text{MAN}, \dot{x}, \text{yes} \rangle\rangle\rangle))_{(e,p), \langle e,p \rangle})$$

$\dot{x}, \dot{y} :=$ Parameters for individuals $((e))$

$\dot{S}, \dot{R} :=$ Parameters of functional type $\langle e, p \rangle$

e.g. $\lambda[\dot{y}](\lambda[\dot{s}](\dot{s} \models \langle\langle \text{height}=180\text{cm}, \dot{y}, \text{yes} \rangle\rangle\rangle))$

To get the information carried by ‘is a man’ with respect to the having of some property, we need to provide a property a man might have. Such a property (or range of properties) may be provided by the context. Below, however, we will see how

part of the semantic contribution of predicate modifiers is to provide such a property (such a range of properties).

Because nominal predicates are functions from properties of described situation types to a function from individuals to a probability, a worry over tractability and learnability arises. If an agent needs to learn all distributions that arise from interpreting ‘*x* is a man’ constructions, this would not be a tractable learning task. However, probabilistic learning affords a solution. We can assume that agents begin classifier learning tasks with flat prior distributions for all collections of properties. On being exposed to uses of a classifier, distributions will be adjusted for just those properties the objects are perceived to have. In this sense, if one has no reasonable expectations regarding some range of situations, given what has been said, simply because no one situation type is more plausible than another, then it is possible that the classifier used does not carry information about (the properties in) those situations.¹⁹

One property a man might have is of being a certain height. For example, he might be 180 cm in height:

$$\lambda[\dot{y}](\lambda[\dot{s}](\dot{s} \models \langle\langle \text{height}=180\text{cm}, \dot{y}, \text{yes} \rangle\rangle))$$

This kind of property can be applied to our representation of the predicate ‘man’:

$$\begin{aligned} & \lambda[\dot{S}](\lambda[\dot{x}](\dot{S} . [\dot{x}] | \lambda[\dot{d}](\dot{d} \models \langle\langle \text{utters}, \dot{a}, \text{MAN}, \dot{x}, \text{yes} \rangle\rangle)) . \\ & \quad [\lambda[\dot{y}](\lambda[\dot{s}](\dot{s} \models \langle\langle \text{height}=180\text{cm}, \dot{y}, \text{yes} \rangle\rangle)]) \\ \Rightarrow & \lambda[\dot{x}](\lambda[\dot{y}](\lambda[\dot{s}](\dot{s} \models \langle\langle \text{height}=180\text{cm}, \dot{y}, \text{yes} \rangle\rangle)) . [\dot{x}] | \\ & \quad \lambda[\dot{d}](\dot{d} \models \langle\langle \text{utters}, \dot{a}, \text{MAN}, \dot{x}, \text{yes} \rangle\rangle))) \\ \Rightarrow & \lambda[\dot{x}](\lambda[\dot{s}](\dot{s} \models \langle\langle \text{height}=180\text{cm}, \dot{x}, \text{yes} \rangle\rangle) | \lambda[\dot{d}](\dot{d} \models \langle\langle \text{utters}, \dot{a}, \text{MAN}, \dot{x}, \text{yes} \rangle\rangle)) \end{aligned}$$

This is now in the right shape to take an individual as an argument which will be described as a man in \dot{d} and assigned the height property in \dot{s} . Assuming (as a simplification) that names refer and carry no other information, this means that ‘John is a man’ can be represented, with respect to being 180 cm in height, as follows²⁰:

$$\lambda[\dot{s}](\dot{s} \models \langle\langle \text{height}=180\text{cm}, j, \text{yes} \rangle\rangle) | \lambda[\dot{d}](\dot{d} \models \langle\langle \text{utters}, \dot{a}, \text{MAN}, j, \text{yes} \rangle\rangle)$$

The interpretation of this will be a probability value that reflects, with respect to John having some height, the probability of John having that height, given that he has been described as a man.

¹⁹Of course, given particularly skewed learning data, anomalies in individuals semantic representations may occur. There will, nonetheless, be general patterns of use across whole language communities. What learners are assumed to be implicitly doing is approximating to the patterns of use in their learning communities as a whole.

²⁰I ignore the contribution of the indefinite article and treat this as a simple predication. I leave the modelling of quantification in this system for future research.

Of course, ‘John is a man’ will carry more information than this about John, but let us focus on information carried about his height. We can consider the range of heights John might be (given that he has been described as a man). There will be a rationality constraint on the values ‘John is a man’ receives with respect to John’s height. The values, must sum to 1. As a simplification, we may think in discrete values, although a more accurate representation would be a continuous function.

Put another way, providing different arguments of height properties for the formula above will generate a probability distribution. Table 2 shows one possible distribution for the height information carried by ‘John is a man’ simplified over ranges of heights.

Describing John as a man, in part, conveys information about his height: that it is fairly probable that he is around average height and highly probable that he is neither far below nor far above average height. This relates to the assumption that learning the classifier ‘man’, in part, involves learning, approximately, the ranges of heights men tend to come in. This needn’t be in centimetres. It need only be internalised in such a way as to aid, to some extent, the ability to identify and classify whether objects in one’s environment are men.

Tall As a predicate modifier, ‘tall’, when applied to a nominal predicate, increases our (reasonable) expectations of the entity (that predicate is applied to) having a greater height. ‘Tall’ will therefore be modelled as having two jobs to do when applied to a common noun like ‘man’: (i) it will pick out the information carried by ‘man’ with respect to height. (ii) it will be a function on the probability distribution that ‘man’ generates with respect to heights (it will make taller heights more probable and shorter heights less probable). In terms of a distribution curve, for a graph with probabilities on its y-axis and heights on its x-axis, it would shift the whole curve along the x-axis in the direction of greater heights. The representation of ‘tall’ is given below, along with a derivation for ‘tall man’.

Table 2 Interpretation of ‘John is a man’ with respect to height

h (cm)	$p(\lambda[s](s \models \langle\langle \text{height}=h, j, \text{yes} \rangle\rangle) \mid \lambda[d](d \models \langle\langle \text{utters}, a, \text{MAN}, j, \text{yes} \rangle\rangle))$
$h < 150$	0.01
$150 \leq h < 155$	0.04
$155 \leq h < 160$	0.08
$160 \leq h < 165$	0.12
$165 \leq h < 170$	0.15
$170 \leq h < 175$	0.20
$175 \leq h < 180$	0.15
$180 \leq h < 185$	0.12
$185 \leq h < 190$	0.08
$190 \leq h < 195$	0.04
$h > 195$	0.01

$$\begin{aligned}
\text{tall:} & \quad \lambda[\dot{P}](\lambda[\dot{y}](\mathbf{C} \times f_{\text{tall}}((\dot{P} \cdot [\dot{y}]) \cdot [(\lambda[\dot{z}](\dot{s} \models \langle\langle \text{height}=\text{h}, \dot{z}, \text{yes} \rangle\rangle))]))) \\
\text{man:} & \quad \lambda[\dot{S}](\lambda[\dot{x}](\langle\dot{S} \cdot [\dot{x}] \mid \dot{d} \models \langle\langle \text{utters}, \dot{a}, \text{MAN}, \dot{x}, \text{yes} \rangle\rangle))) \\
\text{tall man:} & \quad \lambda[\dot{P}](\lambda[\dot{y}](\mathbf{C} \times f_{\text{tall}}(\langle\dot{P} \cdot [\dot{y}]) \cdot [(\lambda[\dot{z}](\dot{s} \models \langle\langle \text{height}=\text{h}, \dot{z}, \text{yes} \rangle\rangle))]))) \cdot \\
& \quad \lambda[\dot{S}](\lambda[\dot{x}](\langle\dot{S} \cdot [\dot{x}] \mid \dot{d} \models \langle\langle \text{utters}, \dot{a}, \text{MAN}, \dot{x}, \text{yes} \rangle\rangle))) \\
& \quad \Rightarrow \lambda[\dot{y}](\mathbf{C} \times f_{\text{tall}}(\lambda[\dot{s}](\dot{s} \models \langle\langle \text{height}=\text{h}, \dot{y}, \text{yes} \rangle\rangle) \mid \\
& \quad \lambda[\dot{d}](\dot{d} \models \langle\langle \text{utters}, \dot{a}, \text{MAN}, \dot{y}, \text{yes} \rangle\rangle)))
\end{aligned}$$

\dot{P} := Parameter of type $\langle\langle e, p \rangle, \langle e, p \rangle\rangle$ (functional type from situation properties to a function from individuals to $[0, 1]$)

\mathbf{C} := Normalising value

f_{tall} := Expression of type $\langle p, p \rangle$

The interpretation of which is a function from individuals to a probability distribution for heights of that individual, given the individual's description as a tall man. Whatever function on distributions f_{tall} is, \mathbf{C} will be interpreted as the value that normalises the distribution.²¹ One possible interpretation for f_{tall} could be that it has the effect of shifting the distribution for 'man' up by 20 cm or so (and stretching it vertically).²² In Table 3, from the above toy distribution for 'John is a man', we can get a toy distribution for 'John is a tall man'.²³

Simply translated, being told that John is a man may carry all sorts of information. Some of this is that he is more likely than not to be within some margin of average

Table 3 Interpretation of 'man', and 'tall man' (unnormalised and normalised) with respect to height

h (cm)	$p(\Phi)$	$p(f_{\text{tall}}(\Phi))$	$p(\mathbf{C} \times f_{\text{tall}}(\Phi))$
$h < 150$	0.01	0.01	0.01
$150 \leq h < 155$	0.04	0.01	0.01
$155 \leq h < 160$	0.08	0.01	0.01
$160 \leq h < 165$	0.12	0.01	0.01
$165 \leq h < 170$	0.15	0.01	0.01
$170 \leq h < 175$	0.20	0.02	0.02
$175 \leq h < 180$	0.15	0.06	0.06
$180 \leq h < 185$	0.12	0.10	0.11
$185 \leq h < 190$	0.08	0.20	0.22
$190 \leq h < 195$	0.04	0.30	0.32
$h > 195$	0.01	0.20	0.22

²¹Which will be 1 over the sum of the modified distribution.

²²This could also be described by adjusting values of parameters on a Gaussian function.

²³I assume that 0.01 is the arbitrarily small value. Φ is an abbreviation for $\lambda[\dot{s}](\dot{s} \models \langle\langle \text{height}=\text{h}, \text{j}, \text{yes} \rangle\rangle \mid \dot{d} \models \langle\langle \text{utters}, \dot{a}, \text{MAN}, \text{j}, \text{yes} \rangle\rangle)$.

height. Being told that he is tall carries the information that it is highly probable that John is over average height. ‘tall’ selects a particular kind of information carried by its common noun (namely information about height), and it amplifies expectations for the upper bounds.

Importantly, after normalisation, we are simply left with a probability distribution (in the final column). Any of those values will be able to enter into formulas for further manipulation if need be. It will be convenient to mark, in the syntax of the formalism, that normalised distributions are mere distributions. This can amount to a dropping of the f and \mathbf{C} , but keeping a record of what the normalised distribution is over. Hence, the interpretation of formulas like:

$$\mathbf{C} \times f_{\text{tall}}(\lambda[\dot{s}](\dot{s} \models \langle\langle \text{height}=\text{h}, \text{j}, \text{yes} \rangle\rangle) \mid \lambda[\dot{d}](\dot{d} \models \langle\langle \text{utters}, \dot{a}, \text{MAN}, \text{j}, \text{yes} \rangle\rangle))$$

will, interpreted, be a probability distribution over heights, which can be rewritten with an updated discourse situation:

$$\lambda[\dot{s}](\dot{s} \models \langle\langle \text{height}=\text{h}, \text{j}, \text{yes} \rangle\rangle) \mid \lambda[\dot{d}](\dot{d} \models \langle\langle \text{utters}, \dot{a}, \text{TALL MAN}, \text{j}, \text{yes} \rangle\rangle)$$

However, in doing so we must be aware that we have increased the probability space of discourse situations. Now there are four possible discourse situation types since John might or might not be described as a man and might or might not be described as tall.

Green As noted in Sect. 3, modifiers like ‘green’ differ from ‘tall’: simply knowing that something (anything) is green *will* give reasonable expectations about what it is like. The way this will be modelled is that ‘green’ will also be a function on probability distributions, but whereas ‘tall’ was informally characterised as taking a distribution over heights and shifting it up (so that greater heights receive higher values and lower heights, lower values), ‘green’ will contribute something more stable to a distribution. It will select shades or hues as properties to distribute over, but it will have the effect of always weighting the distribution towards certain shades. For each shade that an object might be, ‘green’ provides a weighting over our expectations for what is being referred to as being one of those shades.

Formally, ‘green’ will look the same as ‘tall’ and the only difference will be in the functions f_{tall} and f_{green} . Immediately below is the derivation for ‘green car’. For some shade c , the interpretation of the above is a function from individuals to a probability value. For different values of c , this will form a distribution.

Unlike ‘tall’, which pulls height distributions up by some factor, ‘green’ will have a more constant character. Both functions will be of type $\langle p, p \rangle$, but whereas f_{tall} will have the effect of moving a distribution over heights upwards (increasing expectation of heights), f_{green} will always have the effect of flattening a distribution over shades where those shades are not usually described as green, and of elevating the distribution over shades that are usually described as green. The function for ‘green’ will pull any predicate distribution it is applied to towards the same points, namely, some range of shades. This will represent why it is more reasonable to expect something described as ‘green’ to be some shades. This also accounts for why ‘green’

seems to convey information on its own (about shades) in a way that ‘tall’ does not (about heights).

$$\begin{aligned}
 \text{green:} & \quad \lambda[\dot{P}](\lambda[\dot{y}](\mathbf{C} \times f_{\text{green}}((\dot{P} \cdot [y]) \cdot [\lambda[\dot{z}](\lambda[\dot{s}](\dot{s} \models \langle\langle \text{shade}=\text{c}, \dot{z}, \text{yes} \rangle\rangle)))))) \\
 \text{car:} & \quad \lambda[\dot{S}](\lambda[\dot{x}](\dot{S} \cdot [\dot{x}] \mid \lambda[\dot{d}](\dot{d} \models \langle\langle \text{utters}, \dot{a}, \text{CAR}, \dot{x}, \text{yes} \rangle\rangle))) \\
 \text{green car:} & \quad \lambda y. \mathbf{C} \times f_{\text{green}}(\lambda[\dot{s}](\dot{s} \models \langle\langle \text{shade}=\text{c}, \dot{y}, \text{yes} \rangle\rangle) \mid \\
 & \quad \lambda[\dot{d}](\dot{d} \models \langle\langle \text{utters}, \dot{a}, \text{CAR}, \dot{y}, \text{yes} \rangle\rangle))
 \end{aligned}$$

5.2 Connectives

For the interpretations of two declarative statements:

$$\begin{aligned}
 & p(\lambda[\dot{s}](\dot{s} \models \sigma) \mid \lambda[\dot{d}](\dot{d} \models \tau)) \\
 & p(\lambda[\dot{s}](\dot{s} \models \sigma) \mid \lambda[\dot{d}](\dot{d} \models \rho))
 \end{aligned}$$

The interpretation of their conjunction and disjunction will be:

$$\begin{aligned}
 & p(\lambda[\dot{s}](\dot{s} \models \sigma) \mid \lambda[\dot{d}](\dot{d} \models \tau) \wedge \lambda[\dot{d}](\dot{d} \models \rho)) \\
 & p(\lambda[\dot{s}](\dot{s} \models \sigma) \mid \lambda[\dot{d}](\dot{d} \models \tau) \vee \lambda[\dot{d}](\dot{d} \models \rho))
 \end{aligned}$$

Values for which can be obtained using Bayes’ Rule.²⁴ For example, in the conjunction case, this would yield:

$$\begin{aligned}
 & p(\lambda[\dot{s}](\dot{s} \models \sigma) \mid \lambda[\dot{d}](\dot{d} \models \tau) \wedge \lambda[\dot{d}](\dot{d} \models \rho)) = \\
 & \quad \frac{p(\lambda[\dot{s}](\dot{s} \models \sigma)) \wedge p(\lambda[\dot{d}](\dot{d} \models \tau) \wedge \lambda[\dot{d}](\dot{d} \models \rho))}{p(\lambda[\dot{d}](\dot{d} \models \tau) \wedge \lambda[\dot{d}](\dot{d} \models \rho))}
 \end{aligned}$$

6 Vagueness

On the semantics being developed, a phenomena (one that might well be called ‘vagueness’), emerges naturally from meaning representations. If terms leave us with U2 uncertainty, then sometimes there will be a clash between our intuitions of whether to apply a particular term in a particular case. U2 uncertainty (uncertainty over how to apply terms) is related to U1 uncertainty (uncertainty over what the world is like given a description). If U1 uncertainty is graded, then there will undoubtedly arise cases where U2 uncertainties clash: cases where we are no more certain about applying

²⁴Bayes’ Rule, stated in regular notation is: $P(C|A) = \frac{P(C \wedge A)}{P(A)}$.

a term than we are about applying its negation. Furthermore, some cases of clashes will be immune to further information. One could be in a position of full knowledge such that one knew exactly how words were used, all of the relevant properties of the objects being described, and all of the relevant contextual information. Nonetheless, clashes of equal uncertainty may remain unresolved. Such situations, I contend, are plausible explanations of so-called borderline cases.

For example, if the value for ‘John is tall’ with respect to height h is the same as the value for ‘John is not tall’ with respect to height h , then the meaning of ‘tall’ will provide us with no more reason to judge John to be tall than to judge him to be not tall. There is no more to know about the meaning of ‘tall’ which will resolve this uncertainty, and there may be no more to know about John or the context of utterance to do so either. If so, John would be a borderline case.

Probability values in the formalism can be seen to reflect reasons for making judgements. In viewing matters this way, we can begin to get a handle on boundarylessness. Even if our reasons for making a ‘tall’ judgement slightly outweigh our reasons for making a ‘not tall’ judgement, we might still not have *sufficient* reason to form a categorical judgement. The extent to which the values must differ for one or the other judgement to be true will not be a matter decided by the semantics for ‘tall’. In this sense, nothing in the meaning representation itself will determine a cut-off point for ‘tall’. It is due to this feature of the model provided that we can approximate boundarylessness for vague concepts.

7 Comparison with the Literature

7.1 Lassiter’s, and Frazee and Beaver’s Epistemic Models

Although Lassiter (2011) and Frazee and Beaver (2010) describe their positions from different perspectives, they amount to similar approaches. Lassiter’s article, though marginally later was developed independently and is fuller in detail. I will focus on it.

Similarly to my approach, Lassiter is interested in uncertainty in communication and worried about sharp boundaries. Rather than implementing probabilities directly into the meaning representations of words, Lassiter incorporates metalinguistic uncertainty over what is being expressed (in terms of a sharp proposition). He also includes uncertainty about the world, but we’ll focus on the former case.

Lassiter inherits, from other work in the scalar adjectives literature, the idea that their semantics should incorporate thresholds. In simple terms, he models uncertainty about where the threshold is when a term like ‘tall’ is used. We can, relative to a context, be pretty certain that thresholds are not too high or too low. Lassiter models this as a distribution over precise model theoretic objects which have sharp cut-offs. The effect is that if we learn someone’s height, we can get a value that reflects the probability that individual is tall.

However, Lassiter's emphasis is not on the sharp boundaries that such a position implies. For him (and for Frazee and Beaver), communication using 'tall' is about approximating, roughly, the standard in play for what would count as tall. It is these standards that we are uncertain about when someone uses the expression 'tall'. However, this general picture has the effect of implying that there is, nonetheless, a standard for 'tall' in every context. I have suggested that the approximation of standards is not what is encoded by our words. This amounts to the claim that uncertainty should enter at the level of describing the model theoretic object, not at the level of evaluating which classical model theoretic object is in play.

Lassiter's account is a big improvement on standard non-probabilistic approaches. However, the fact remains that the motivation for appealing to precise languages must still be given. If it turns out that we rarely, if ever, coordinate sufficiently to settle on exactly where a threshold for 'tall' should be, then why have such thresholds written so centrally into the semantics of such terms? We can drop precise languages and have a far more direct connection to what our terms mean by adopting the picture proposed.

7.2 Cooper et al.'s Probabilistic Type-Theoretic Approach

Semantic learning is at the forefront of Cooper et al. (2014), which is also the closest position to my own. I strongly suspect that the core of the two positions will be pretty inter-definable, although there are differences of emphasis.

U1 uncertainty is uncertainty about the world, given a description of it. U1 uncertainty is, effectively, what my account describes. Arguably, Cooper et al. focus on U2 uncertainty. U2 uncertainty is uncertainty about how to use words, given some known or perceived way the world is. I will say more about Cooper et al. formalism in a moment, but it is first worth remarking that, if modelled as I have presented, U1 and U2 uncertainty are inter-definable. The main weight of the meanings of utterances in my account rests on conditional probabilities of the form $p(\lambda[s](s \models \sigma) | \lambda[d](d \models \tau))$. In other words, the probability of the world being some way, given a description of it. If, indeed, Cooper et al.'s account describes U2 uncertainty, then it is possible that their results could be simulated via the use of the alternate conditional probability: $p(\lambda[d](d \models \tau) | \lambda[s](s \models \sigma))$ (or the probability that some description will be used, given that the world is some way). Importantly, given the priors $p(\lambda[s](s \models \sigma))$ and $p(\lambda[d](d \models \tau))$, these two conditional probabilities are simply ratios of each other.

Cooper et al.'s semantics uses a rich theory of types. The simple type theory I have used provided domains for basic types. Complex types were then constructed as functions of basic types. In rich type theories, types should not be thought of in these extensional terms, but instead as something like useful ways of classifying things (Cooper 2012, p. 275). However, propositions are also types. For example, the proposition *David runs* can be seen as a situation type (one in which David runs), and is true iff there is a situation in the world in which David runs. In Cooper et al.

(2014) agents are modelled as making probabilistic judgements about classifications. Propositions are still types, but judgements reflect the probabilities that there is a situation of the right type.

Of particular interest is how these type judgements are grounded. Central to Cooper et al.'s account is semantic learning, a wider discussion of which features in Cooper et al. (2013). On their learning model, a learner is exposed to multiple situations. In each one they make a judgement about whether the object they are shown is an apple or not an apple. After each judgement, the oracle that models the adult speaker gives a judgement. Initially, the learner cannot make a judgement, but after a few exposures to adult judgements, has enough data to begin to make judgements herself. In Cooper et al.'s model, 'apple' judgements are based on four properties (two colours and two shapes). Simplifying a little, when faced with a new object to classify, the value of the judgement is calculated as the conditional probability $p(\text{apple}|\text{observed properties})$.²⁵ To calculate that conditional probability, the learner uses priors and conditional probabilities such as $p(\text{property}|\text{apple})$, both of which are estimated directly from the adult judgements she has witnessed.

Translating more into my own vernacular, one could see probabilities of 'apple' judgements as approximating probabilities of discourse situations, and we could see probabilities of properties as approximating probabilities of described situations. The two stages of Cooper et al. learning account can then be described with the following procedure:

- (i) By directly witnessing adult speakers' linguistic behaviour, estimate probabilities of the form:

$$\begin{aligned} & p(\lambda[\dot{s}](\dot{s} \models \langle\langle \text{colour}_c, \dot{x}, \text{yes} \rangle\rangle)), p(\lambda[\dot{s}](\dot{s} \models \langle\langle \text{shape}_s, \dot{x}, \text{yes} \rangle\rangle)), \\ & p(\lambda[\dot{d}](\dot{d} \models \langle\langle \text{utters}, \dot{a}, \text{APPLE}, \dot{x}, \text{yes} \rangle\rangle)), \\ & p(\lambda[\dot{s}](\dot{s} \models \langle\langle \text{colour}_c, \dot{x}, \text{yes} \rangle\rangle) | \lambda[\dot{d}](\dot{d} \models \langle\langle \text{utters}, \dot{a}, \text{APPLE}, \dot{x}, \text{yes} \rangle\rangle)), \text{ and} \\ & p(\lambda[\dot{s}](\dot{s} \models \langle\langle \text{shape}_s, \dot{x}, \text{yes} \rangle\rangle) | \lambda[\dot{d}](\dot{d} \models \langle\langle \text{utters}, \dot{a}, \text{APPLE}, \dot{x}, \text{yes} \rangle\rangle)). \end{aligned}$$

- (ii) Use those values to calculate, of a novel object/context, the probability:

$$\begin{aligned} & p(\lambda[\dot{d}](\dot{d} \models \langle\langle \text{utters}, \dot{a}, \text{APPLE}, \dot{x}, \text{yes} \rangle\rangle) | \\ & \lambda[\dot{s}](\dot{s} \models \langle\langle \text{colour}_c, \dot{x}, \text{yes} \rangle\rangle) \wedge \lambda[\dot{s}](\dot{s} \models \langle\langle \text{shape}_s, \dot{x}, \text{yes} \rangle\rangle)). \end{aligned}$$

The process in (i) describes learning the kinds of meaning representations I have proposed. Cooper et al. show how the kind of information that can be learnt from adult speakers' linguistic behaviour can be turned into a classifier judgement. These judgements will not always be probability 1. The uncertainty that <1 judgements reflect, is, I would argue, just what I described as U2 uncertainty. A closer examination of the two accounts could be fruitful. It would certainly be useful if I were simply able to adopt Cooper et al. learning model.

²⁵Where the output judgement is decided by whether an 'apple' or an 'not-apple' judgement receives a higher value.

One distinct contribution the account in this paper makes is in the treatment of adjectives. Cooper et al.'s learning account shows how a nominal like 'apple' could be learnt from basic colour and shape observations. I have argued that vague modifiers should be viewed as functions on distributions given by nominal classifiers.

8 Conclusions

Situation theory captures sentence meaning as a link between discourse situation types and described situation types. Although the way that this idea has been developed here differs from the situation theoretic account in many ways, the basic spirit of the account remains. On the standard view, informational links hold between situation types. Individuals build connections with the world via learning (learning to decode the information carried by terms) and via communication. The reproduction of expressions for similar purposes entrenches these relations.

On the account presented here, the meanings of expressions are correlational informational links between discourse situation types and described situation types that arise from patterns in language use. People connect to these correlations via semantic learning, and they entrench correlations via reproduction of language to refer to similar situations.

Vagueness naturally arises from modelling such relations probabilistically. This can be viewed either as the boundarylessness of vague predicates achieved by separating, or at least distancing, truth conditions from semantic representations, or as borderline cases, where the meaning of a term such as 'tall' can provide competing reasons to make 'tall' and 'not tall' judgements.

However, the suggestions put forward here are rather programmatic and many issues have remained entirely unaddressed. Nonetheless, at least for vague terms, structured, probabilistic representations of meaning provide at least one viable route for attempting to capture the boundarylessness that seems characteristic of vagueness.

Acknowledgments I would like to thank the organisers and participants of the Bayesian Natural Language Semantics and Pragmatics workshop held at ESSLI 2013 and the anonymous reviewers for the workshop and this volume for the helpful comments and improvements that they have suggested. Thanks also to the participants of the King's College London Language and Cognition Seminar, with special thanks to Alex Clark, Ruth Kempson, Shalom Lappin, and Wilfried Meyer-Viol for their invaluable comments on earlier versions of this paper.

References

- Barwise, J., & Perry, J. (1983). *Situations and attitudes*. Stanford: CSLI.
- Cooper, R., Dobnik, S., Lappin, S., & Larsson, S. (2013). Probabilistic type theory and natural language semantics (manuscript) University of Gothenburg and King's College London.

- Cooper, R., Dobnik, S., Lappin, S., & Larsson, S. (2014). A probabilistic rich type theory for semantics interpretation. In *Proceedings of the EACL 2014 Workshop on Type Theory and Natural Language Semantics*, Gothenburg.
- Cooper, R. (1996). The role of situations in generalised quantifiers. In S. Lappin (Ed.), *The handbook of contemporary semantic theory* (pp. 65–86). Cambridge: Blackwell.
- Cooper, R. (2012). Type theory and semantics in flux. In R. Kempson, T. Fernando, & N. Asher (Eds.), *Philosophy of linguistics Handbook of the philosophy of science*, (vol. 14, pp. 271–323). Oxford: Elsevier.
- Cresswell, M. J. (1974). The semantics of degree. In B. Partee (Ed.), *Montague grammar* (pp. 261–292). New York: Academic Press.
- de Marnee, M., Manning, C.D., & Potts, C. (2010). Was it good? It was provocative. Learning the meaning of scalar adjectives. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 167–176). Stroudsburg, ACL.
- Devlin, K. (2006). Situation theory and situation semantics. In D. Gabbay & J. Woods (Eds.), *Handbook of the history of logic* (Vol. 7, pp. 601–664). Amsterdam: Elsevier.
- Fraeeze, J., & Beaver, D. (2010). Vagueness is rational under uncertainty. In M. Aloni, H. Bastiaanse, T. de Jager, & K. Schulz (Eds.), *Logic, language and meaning: 17th Amsterdam colloquium, Amsterdam, The Netherlands, December 16–18, 2009, Revised Selected Papers. LNAI* (Vol. 6042, pp. 153–162). Heidelberg: Springer.
- Gargett, A., Gregoromichelaki, E., Kempson, R., Purver, M., & Sato, Y. (2009). Grammar resources for modelling dialogue dynamically. *Cognitive Neurodynamics*, 3, 347–363.
- Graff Fara, D. (2000). Shifting sands: An interest relative theory of vagueness. *Philosophical Topics*, 28(1), 45–81.
- Kamp, H. (1975). Two theories about adjectives. In E. Keenan (Ed.), *Formal semantics of natural language* (pp. 123–155). Cambridge: Cambridge University Press.
- Kamp, H., & Reyle, U. (1993). *From discourse to logic*. Dordrecht: Kluwer.
- Kennedy, C. (1997). Projecting the adjective: The syntax and semantics of gradability and comparison. Ph.D. Thesis, University of California, Santa Cruz.
- Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, 30(1), 1–45.
- Kolmogorov, A. (1950). *Foundations of probability*. New York: Chelsea Publishing.
- Lassiter, D. (2011). Vagueness as probabilistic linguistic knowledge. In R. Nouwen, U. Sauerland, H. C. Schmitz, & R. van Rooij (Eds.), *Vagueness in communication. LNAI* (Vol. 6517, pp. 127–150). Heidelberg: Springer.
- Sainsbury, R. M. (1991). *Concepts without boundaries, Inaugural Lecture*. King's College London, London. Reprinted. In R. Keefe, P. Smith (Eds.), *Vagueness: A reader* (pp. 251–264). MIT Press, Cambridge (1996).
- Sutton, P. R. (2013). Vagueness, communication and semantic information. King's College London. Ph.D. Thesis.