On Bayesian pragmatics and categorical predictions

Anton Benz*

1 Introduction

Intuitions about pragmatic phenomena can sometimes be fuzzy. If a lecturer says that "All of my students solved some of the problems", does he mean that none solved all? And if he adds that "Some of my students solved all of the problems", what is the percentage of his students who did solve them? If there is a green square and two circles, a green and a blue one, and somebody is asked to bring the circle, which one would he choose? Experiments on pragmatic phenomena reveal a certain amount of variance among subjects, and rarely exhibit the clear distributions that would be predicted by formally oriented pragmatic theories. Michael Franke and Gerhard Jäger (this volume), henceforth F&J, outline an approach to pragmatic modelling that tries to account for this fuzziness. They aim at explanatory models that not only describe the uncertainty and variance found in pragmatic interpretations but that derive the observed behaviour from fundamental assumption about rational language use and the distribution of noise. Their approach replaces the categorical principles of neo-Gricean and grammatical approaches with probabilistic rules. F&J discuss two modelling techniques in some depth, the first employs discrete choice theory to explain the distribution of experimental outcomes in tests on referring expressions, and the second game theory to explain implicature about the speaker's type in bargaining situations.

Pragmatics as considered here is concerned with communicated meaning. Communicated meaning has many components. Well known is Grice (1989: Ch. 14) distinction between *what is said* and what is *implicated*, both of which are considered part of the intended speaker meaning. Besides this, there is a more fuzzy component consisting of expectations about what is likely the case. These expectations are not necessarily part of the intended speaker meaning. An example are the expectations raised by an utterance of "some A's are B" about the exact number of A's being B's. The exciting new aspect of discrete choice methods, as discussed by F&J, is that they bring such fuzzy phenomena into the reach of linguistic pragmatics. In this comment, I want to address the relation between probabilistic pragmatics and *old style* categorical predictions which are typical for Gricean intention based pragmatics (Grice 1989). I concentrate, in particular, on the Bayesian models of *soft* implicature (F&J Sec. 5) and fuzzy data

* **Corresponding Author: Anton Benz**, Zentrum für Allgemeine Sprachwissenschaft (ZAS), Berlin, E-Mail: benz@zas.gwz-berlin.de

in the reference game (F&J Sec. 4). In general, I am very sympathetic to the approach taken by the authors. However, I think that they tend to over-emphasis the probabilistic and fuzzy nature of pragmatics in general. Categorical models will continue to play an important role, in particularly, when taking language generation into account. By categorical I mean models that predict utterances either to have certain interpretations or not to have them, in contrast to models which assign only certain probabilities to interpretations. The following section elaborates on the different components of communicated meaning and raised expectations.

2 Modern Bayesianism and old Griceanism

A core problem of pragmatics is the observation that the amount of communicated information is much higher than the linguistically encoded information. Levinson (2000: 6, 28) observed that the verbal channel poses a *bottleneck* problem for information flow as its capacity is much too low to allow for effective communication. The solution that pragmaticists offer, whether they are neo-Griceans (Horn 1989, Levinson 1983, Levinson 2000) or post-Griceans (Sperber and Wilson 1995, Recanati 2004, Recanati 2011), is that some form of pragmatic enrichment must take place that augments structurally encoded information. Different schools pursue different strategies addressing this problem. Game theoretic approaches consider communication as a strategic problem, and a number of models have been proposed for explaining implicated meaning.[1] In recent years even advanced topics such as embedded implicatures of complex sentences have been successfully addressed (Franke 2009, Franke 2011, Benz 2012, Pavan 2013). F&J showed in one example[2] how a customer statement about the aesthetic quality of a product can indirectly communicate whether they are ready to pay a low, medium, or high price for it. These game theoretic models are categorical in the sense that implicature are inferred with certainty. In the standard theory (Levinson 1983: 131), implicated meaning is speaker intended meaning that becomes part of common ground. There is, however, a growing tendency to count as implicature any probabilistic expectation that may be raised by an utterance. Partly this is due to grammatical accounts (Fox 2007, Chierchia, Fox and Spector 2012, Chierchia 2013) which integrate scalar implicature into semantics such that sentences with scalar expressions become highly ambiguous and the different readings available in varying degrees (Chemla and Spector 2011). Partly this is also due to the rise of Bayesian accounts under which implicature become probabilistic inferences, but not communicated meaning.[3] In addition, there are experimental studies showing that *soft* scalar inferences exist which standard theories cannot account for. The work by Degen & Tannenhaus (2011, 2015) discussed by F&J belongs here. An interesting case are modified numerals. As the study by Cummins, Sauerland and Solt (2012) showed, numerals embedded under "more than" can give rise to expectation about upper bounds. For example, "more than 80" gives rise to the expectation *less than 100*, and "more than 70" to the expectation *less than 80*.

[1] Pioneering work include Parikh (1992), Parikh (2001), Merin (1999).

[2] Section 6, *Why be indirect?* Example (9).

[3] For example, Russell (2012), Potts, Lassiter, Levy and Frank (2015), and the applications of discrete choice theory discussed by F&J.

Cummins, Sauerland, and Solt considered these inferences as a kind of *weak* implicature. Krifka (1999) and Fox & Hackl (2006), in contrast, maintained that "more than" cancels the implicatures that may be generated by the numeral. Cummins (2013) explained them as inferences about the speaker's beliefs in an optimality theoretic model, where no principled difference between *weak* and *strong* implicature seem to exist. In (Benz 2015) I explained the difference between *soft* and *strong* implicature as the difference between induced expectations about general speaker beliefs, and inferences about speaker intended meaning. I would maintain that soft and strong implicature are two different kinds of beast, and that putting them into the same family leads to confusion. Let us consider an utterance of (S) "John has three children". Keeping the bottleneck problem in mind, how much pragmatic enrichment has to take place to arrive at the speaker's meaning? First "John" is a name that is carried by many people. The speaker would have to say something like "John, the person we are currently talking about" or "John, the person who is mutually known to us by this name" to express literally what is communicated by (S). Second, the possessive relation expressed by "has" is highly ambiguous. It may mean that John has three children in his team, visiting, in his care, etc. To literally express what (S) communicates, the speaker would have to say that "John is the biological father of three children". Third, there is the scalar inference from "three" to "not more than three". Fourth, there is an ambiguity in "children" which, in this utterance, may include adults. A core goal of pragmatic theories is to explain the mechanisms behind pragmatic enrichment, the conditions under which it occurs, or does not occur. With this example, we may contrast the example discussed by F&J taken from (Franke 2014): the implicature from "There are 10 circles. Some of the circles are white." to "(probably) N of the circles are white". Of course, if one asks "How many white circles are there?", then test subjects make guesses which show that certain numbers are more expected than others. Likewise, if one asked them after an utterance of (S) the question "How old are the children?", then certain numbers will be more frequent than others. I would not be surprised if subjects would provide different estimates when asked "How tall is the house?" after reading the sentence "John left the house and went down to the underground" than after reading the sentence "John left the house and went to his garage". The latter scenario is not as absurd as it may appear. Such diffuse expectation should play an important role in narrations, or, for example, in advertisement. In the case of modified numerals, they provide us with linguistically interesting insights about the relation between roundness and number interpretations. It is exciting that Bayesian probabilistic pragmatics offer us a framework with which to address these soft implicature. However, I think that soft implicature and implicature as part of speaker intended meaning should be kept conceptually apart, first because they give rise to different research questions, and second to forestall unnecessary polemics between *old-fashioned* Gricean approaches, and *modern* probabilistic accounts.

3 A re-analysis of the reference game

Probabilistic pragmatics is a part of linguistic pragmatics that extends into a number of other disciplines, among them artificial intelligence, cognitive science, economics and philosophy. It covers a group of theories that aim at simulating and explaining pragmatic behaviour. I want to begin by listing some properties that such pragmatic theories ideally should have: First, they should be specific. This means they should not only tell which interpretations or readings a sentence can have in general but also predict what a sentence means in a specific dialogue context. Second, the theories should not only explain utterance interpretation but also utterance generation. Third, the theories should be cognitively interpretable. Fourth, they should allow for the formulation of explicit algorithms that can be implemented in actual software. Fifth, they should be experimentally testable. Clearly, there is no theory that would satisfy all these requirements. What is more, these requirement create a certain tension that makes it difficult to satisfy them all at once. Cognitive adequacy and what F&J call *data orientedness* seems to call for probabilistic descriptions of behaviour. The assumptions of *rationality* and *optimality* lead to models describing how language *should* be used. This normative aspect is particularly important when it comes to computational applications, for example, natural language generation (see e.g. Stevens, Benz, Reuße and Klabunde 2015). Optimality and algorithmic simplicity both tend to favour non-probabilistic rules. In this section, I have a closer look at the reference game which illustrates this tension between probabilistic descriptive models and normative optimality based descriptions. I will keep reformulating F&J's model, and thereby derive a categorical core from which in turn a production algorithm can be derived. This reformulation exercise also illustrates the inherent indeterminacy of discrete choice models, i.e. the fact that for every model there exist, in general, infinitely many equivalent models with the same fit to the data.

In the reference game, a picture of three geometrical objects was shown to subjects, see Figure 1. There were two players, a speaker and a hearer. The speaker was given a referent r for which he had to produce a signal. After receiving the signal, the hearer had to guess which object was meant. The speaker's choice was restricted to one of the properties p of the referent. He could choose between a shape name, *square* or *circle*, and a colour adjective, *green* or *blue*.

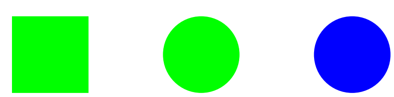


Figure 1: Experimental token of the reference game

F&J's model is based on *discrete choice theory*. They assume that the probability of the speaker choosing property p depends on the probability of a literally interpreting hearer choosing r given p and a (negative) cost parameter f for choosing an adjectives. The probability of a literally interpreting hearer choosing r are shown in Figure 2. Hence, the speaker's production probability is given by:

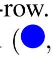
$$P(p|r;\lambda,f) = \frac{\exp(\lambda\,(P_{lit}(r|p)+f(p))}{\sum_{p'}\exp(\lambda\,(P_{lit}(r|p)+f(p))} \tag{1}$$

By numerical approximation, values for λ and f can be found that maximise the probability of the data given the speaker's production probability. As F&J report, the optimal values are $\lambda_{max} = 4.13$ and $f_{max} = -0.1$.[4]

	property			
referent	square	green	circle	blue
	1	$1/2$	0	0
	0	$1/2$	$1/2$	0
	0	0	$1/2$	1

Figure 2: Probabilities of a literally interpreting hearer choosing referent r given property p. $P_{lit}(r|p) = 1/|p|$ if $r \in p$, otherwise $P_{lit}(r|p) = 0$.

Eq. 1 tells us that the production probability depends on a payoff function $U(r,p) = P_{lit}(r|p) + f(p)$. In economics, payoff functions that differ only with respect to affine linear transformations are assumed to represent the same preferences. This means, if $U'(r,p) = a\,U(r,p) + t$ with $a > 0$, then U' and U are empirically equivalent. It is easy to see that the production probability does not change if a constant is added to $U(r,p)$. A multiplication with $a > 0$ changes the production probability, and so seems to have an effect. However, if the maximum likelihood is estimated, then the same maximum is reached at $\lambda'_{max} = \lambda_{max}/a$ and $f'_{max} = f_{max}$. What this means is that by rescaling the utility $U(r,p)$ we can obtain models that are equivalent with respect to descriptive value and fit to the data. I use such variations and show how an equivalent model with categorical rules can be derived.

Let us consider the probabilities shown in Figure 2, and concentrate on the cells which are different from 0. A speaker who has to choose a property for a referent r has only to consider the properties p with non-zero entries in the r-row. If we subtract $1/2$ everywhere, then all the cells are 0 except for (, square) and (, blue). This is equivalent to adding the constant 0.5 to $U(r,p)$, and therefore changes nothing. We then can rescale the remaining values by multiplying them with 2, such that there is a 1 in the upper left and in the lower right corner. The last rescaling halves λ_{max} and doubles f_{max}. In general, if we rescale the corner cells by multiplying them with ξ, then the new $\lambda'_{max} = \lambda_{max}/\xi$, and the new $f'_{max} = \xi\, f_{max}$. For example, for $\xi = 1/100$ and original $\lambda_{max} = 4.135$, $f_{max} = -0.0937$, the new values are $\lambda'_{max} = 413.5$, $f'_{max} = -0.000937$. The models resulting from rescaling are equivalent in the sense that the predicted production probabilities, which are shown in Figure 3, are the same. On the left side of Figure 3, also the observed production probabilities are listed. The Euclidean distance between the vectors indicated by the framed probabilities is $d = 0.042$.

We have achieved two things by this rescaling exercise: first, we have learned that λ_{max} is far less informative as one might think at first, and second, we have found a method for transforming an interactional, probabilistic model into an equivalent non-interactional, constraint based model. For the first point: one might think that λ provides a measure of the degree of rationality with which speakers choose alternative

[4]Values may slightly differ depending on software and algorithms used. Doing maximum likelihood estimation in Mathematica, I arrived at $\lambda = 4.135$, $f = -0.0937$. All following calculations are done with these values.

	property				property			
referent	square	green	circle	blue	square	green	circle	blue
■	$\boxed{.94}$	.06	0	0	$\boxed{.92}$	.08	0	0
●	0	$\boxed{.44}$	.56	0	0	$\boxed{.40}$	.60	0
●	0	0	$\boxed{.17}$	.83	0	0	$\boxed{.16}$	.84
	a) observed				b) predicted			

Figure 3: Production probabilities in the reference game ($\lambda = 4.135$, $f = -0.0937$).

utterances, with higher values of λ_{max} indicating a higher level of rationality. However, λ_{max} depends on the utility $U(r, p)$, which in turn depends on the values of f and the values in Fig 2. For example, that we find that λ'_{max} is a hundred times larger in the model rescaled with $\xi = {}^{1}/_{100}$ than in the original model does not imply that the speakers are 100 times more rational after rescaling. For the second point: after rescaling the values in Fig 2 such that only 1 remains in the upper left and lower right corner, we can interpret the result as the representation of a categorical constraint similar to an optimality theoretic constraint: *avoid ambiguity*. The utility $U(r, p)$ can then be interpreted as a representation of two constraints:

- (A) Avoid ambiguity
- (U) Use colour adjectives

This re-interprets the utilities, but does not yet explain the observed probabilities. However, this constraint based model can be turned into a probabilistic model by adding weights to the constraints: If α is the speaker's weight for constraint (A), and β the weight for (U), then the predicted probabilities can be calculated by normalising Figure 4:

referent	square	green	circle	blue	denominator
■	α	β	0	0	$\alpha + \beta$
●	0	β	1	0	$1 + \beta$
●	0	0	1	$\alpha \cdot \beta$	$1 + \alpha \cdot \beta$

Figure 4: Weighted speaker strategies

Optimising these weights such that the Euclidean distance between $(\alpha, \beta, 1)$ and the corresponding vector indicated by the framed observed values in Figure 3 leads to $\alpha = 7.59$ and $\beta = 0.73$ with a distance of $d = 0.037$.[5]

[5]The numerical value of d is slightly better than that for F&J's model. However, this may partly be due to imprecisions caused by software. Due to space limitations, we cannot explain here how the weights in Figure 4 can be derived from a logit model. The correspondence is given by $\alpha = \exp({}^{1}/_{2}\,\lambda)$ and $\beta = \exp(\lambda f)$. Further note that the values for λ_{max} and f_{max} were found by maximising the log-likelihood of the experimental results given λ and f, whereas α and β were found by minimising the Euklidean distance between the corresponding framed vectors shown in Figure 3.

The advantage of this transformation of F&J's logit model into a constraint based model is that we not only have a *descriptive* model with the same fit to observed probabilities but also a non-probabilistic *computational* model that generates referring expressions: choose the expression with highest weight! This rule can be translated into a linear production model in which one constraint is applied after the other: Follow constraint (A) (*avoid ambiguity*), and if this is not possible, violate (U) (i.e. violate *use colour terms*).[6] This tells us that the speaker can avoid the cognitively more costly reasoning about the hearer's expected reaction by simply following rules that only depend on linguistic properties of the referring expressions (U), and on objective facts about the utterance situation (A). Thereby, it provides an *explanatory* account based on a rudimentary cognitive *production* model. Is this rule also *optimal*? The reference game is a game of pure coordination. Hence, for every mixed strategy, there has to be a pure strategy with the same or even higher expected utility. If the observed speaker strategy corresponds at all to the observed hearer strategy, then the pure strategy that I proposed must be optimal. In fact, as the data provided by F&J show, the two strategies do not correspond. The optimal strategy against the hearer is to follow (A), and if this is not possible to follow (U) — and not to violate it. We observe here a conflict between observed strategies. However, from a *normative* perspective we can always provide a categorical production rule that is either optimal against the hearer strategy or is a pure representation of the speaker strategy.

The model presented by F&J is, of course, a toy model, and so is the one presented here. It may well be that F&J's model scales up better to situations in which referring expressions are ambiguous to varying degrees, or linguistic expressions come with a variety of different cost levels. However, the alternative model shows that it is sometimes possible to excavate from noisy probabilistic data the skeleton of a constraint based production model that is no less adequate.

4 Conclusion

The article by F&J demonstrated that probabilistic economic methods are a promising tool for analysing experimental data, and in particular, fuzzy expectations about general speaker beliefs, sometimes called *weak* or *soft* implicatures. I have argued, however, that these implicatures should not be mixed up with standard Gricean implicature understood as inferences towards the speaker's intended meaning. The repeated rescaling of parameters in the reference game demonstrated the indeterminacy of discrete choice models. The alternative models that I developed are not motivated by any flaws in the models provided by F&J, nor can they claim a better fit to the data. They incorporate, however, different constraints. They also illustrate the tension between a normative, optimality based perspective, and a purely descriptive perspective. The interpretation of discrete choice models is a very delicate matter. The question that arises is: what is the explanatory value of one specific model, even if it has a very good fit to the data, if there are infinitely many alternative models with the same fit?

[6]There is also an alternative model: violate (U) and choose a shape name, and then revise your choice if the colour adjective but not the shape name satisfies (A). In terms of fit to the data, there is no reason to prefer one model over the other.

Acknowledgements: This work was supported by Bundesministerium für Bildung und Forschung (BMBF, Grant Nr. 01UG1411), and Deutsche Forschungsgemeinschaft (DFG, Grant BE 4348/3-1).

References

Benz, Anton. 2012. Implicatures of complex sentences in error models. In Andrea Shalley (ed.), *Practical theories and empirical practice*, 273–306. Amsterdam: Benjamins.

Benz, Anton. 2015. Causal Bayesian networks, signalling games and implicature of 'more than n'. In Henk Zeevat & Hans-Christian Schmitz (eds.), *Bayesian natural language semantics and pragmatics*, 23–43. Heidelberg: Springer.

Chemla, Emmanuel & Benjamin Spector. 2011. Experimental evidence for embedded scalar implicatures. *Journal of Semantics* 28(3). 359–400.

Chierchia, Gennaro. 2013. *Logic in grammar: Polarity, free choice, and intervention*. Oxford: Oxford University Press.

Chierchia, Gennaro, Danny Fox & Benjamin Spector. 2012. Scalar implicature as a grammatical phenomenon. In Claudia Maienborn, Klaus von Heusinger & Paul Portner (eds.), *Semantics: An international handbook of natural language meaning*, vol. 3, 2297–2331. Berlin: De Gruyter Mouton.

Cummins, Chris. 2013. Modelling implicatures from modified numerals. *Lingua* 132. 103–114.

Cummins, Chris, Uli Sauerland & Stephanie Solt. 2012. Granularity and scalar implicature in numerical expressions. *Linguistics and Philosophy* 35(2). 135–169.

Degen, Judith & Michael K. Tanenhaus. 2011. Making inferences: The case of scalar implicature processing. *Proceedings of CogSci* 2011. 3299–3304.

Degen, Judith & Michael K. Tanenhaus. 2015. Processing scalar implicatures: A constraint-based approach. *Cognitive Science* 39(4). 667–710. doi: 10.1111/cogs.12171.

Fox, Danny. 2007. Free choice and the theory of scalar implicatures. In Uli Sauerland & Penka Stateva (eds.), *Presupposition and implicature in compositionsl semantics*, 71–120. Basingstoke: Palgrave Mcmillan.

Fox, Danny & Martin Hackl. 2006. The universal density of measurement. *Linguistics and Philosophy* 29(5). 537–586.

Franke, Michael. 2009. *Signal to act: Game theory in pragmatics*. Amsterdam: Universiteit van Amsterdam Dissertation.

Franke, Michael. 2011. Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics and Pragmatics* 4(1). 1–82.

Franke, Michael. 2014. Typical use of quantifiers: A probabilistic speaker model. *Proceedings of CogSci* 2014. 487–492.

Franke, Michael & Gerhard Jäger. This volume. Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft* 35(1). xx–xx.

Grice, Herbert Paul. 1989. *Studies in the way of words*. Cambridge, MA: Harvard University Press.

Horn, Laurence R. 1989. *A natural history of negation*. Chicago, IL: University of Chicago Press.

Krifka, Manfred. 1999. At least some determiners aren't determiners. In Ken Turner (ed.), *The semantics/pragmatics interface from different points of view*, 257–292. Oxford: Elsevier.

Levinson, Stephen C. 1983. *Pragmatics*. Cambridge: Cambridge University Press.

Levinson, Stephen C. 2000. *Presumptive meanings: The theory of generalized conversational implicatures*. Cambridge, MA: MIT Press.

Merin, Arthur. 1999. Information, relevance, and social decisionmaking: Some principles and results of decision-theoretic semantics. In Larry Moss, Jonathan Ginzburg & Martin de Rijke (eds.), *Logic, language, and information*, vol. 2, 179–221. Stanford, CA: CSLI Publications.

Parikh, Prashant. 1992. A game-theoretic account of implicature. In Yoram Moses (ed.), *Proceedings of the 4th conference on theoretical aspects of reasoning about knowledge*, Monterey, CA: Morgan Kaufmann.

Parikh, Prashant. 2001. *The use of language*. Stanford: CSLI Publications.

Pavan, Sascia. 2013. Quantity implicatures and iterated admissibility. *Linguistics and Philosophy* 36(4). 261–290.

Potts, Christopher, Daniel Lassiter, Roger Levy & Michael C. Frank. 2015. Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. Ms., Stanford University.

Recanati, Francois. 2004. *Literal meaning*. Cambridge: Cambridge University Press.

Recanati, Francois. 2011. *Truth-conditional pragmatics*. Oxford: Oxford University Press.

Russell, Benjamin. 2012. *Probabilistic reasoning and the computation of scalar implicatures*. Providence, RI: Department of Cognitive, Linguistic and Psychological Sciences at Brown University.

Sperber, Dan & Deirdre Wilson. 1995. *Relevance: Communication and cognition*. Oxford: Blackwell Publishing.

Stevens, Jon Scott, Anton Benz, Sebastian Reuße & Ralf Klabunde. 2015. A strategic reasoning model for generating alternative answers. In Chengqing Zong & Michael Strube (eds.), *Proceedings of the 53rd annual meeting of the ACL*, 534–542. Stroudsburg, PA: Association for Computational Linguistics.