Relevance and Bidirectional OT

Robert van Rooy*
ILLC/University of Amsterdam
vanrooy@hum.uva.nl

1 Introduction

According to optimality theoretic semantics (e.g. de Hoop & de Swart 2000) there exists a gap between the semantic representations of sentences and the thoughts actually communicated by utterances. How should this gap be filled? The obvious answer (Grice, 1957) seems to be that the hearer should recognize what the speaker thinks that the listener understands. Because this depends in turn, in a circular way, on what the listener thinks that the speaker has in mind, a game-theoretical framework seems natural to account for such situations. Intuitively, what goes on here is a game between a speaker and a hearer, where the former chooses a form to express the intended meaning, and the latter chooses a meaning corresponding to the form. Blutner's Bidirectional OT, based on the assumption that both speaker and hearer optimize their conversational actions seems perfectly suitable to implement this. But how can a hearer recognize the speaker's intentions? Gricean pragmatics (1975) suggests that she can do so by assuming that the speaker is cooperative and thus obeys the conversational maxims. Sperber & Wilson (1986) have suggested that these 4 conversational maxims can be reduced to the single principle of optimal relevance. In this paper I will discuss how far this can be done. I will argue that conversation involves resolving one of the participants' decision problems. After discussing bidirectional OT I will show how decision theory can be used to determine the utility of an interpretation in a mathematically precise way. Then I will discuss how this formal notion of utility, in combination with bidirectional OT, can account for a number of conversational implicatures and how it relates to (i) Sperber & Wilson's psychologically inspired notion of cognitive relevance; (ii) the Stalnakerian assertability conditions; (iii) the Gricean maxims of conversation, and (iv) the so-called Q and I principles of neo-Gricean pragmatics (Horn 1984; Levinson, 2000).

*The research of this work is supported by a fellowship from the Royal Netherlands Academy of Arts and Sciences (KNAW), which is gratefully acknowledged. I would like to thank David Beaver, Reinhard Blutner, Robyn Carstons, Gerhard Jäger, Marie Nilsenova, Katrin Schulz, Henk Zeevat, and the anonymous reviewers for their comments and suggestions. Some ideas developed in section 6.4 are closely related with some proposals made by Reinhard Blutner (p.c.) and Katrin Schulz (2001). Many thanks also to Raj Singh for commenting on an earlier version of this paper and for correcting my English.

2 Bidirectional OT

Optimality Theory (OT) assumes that a linguistic form should be interpreted in the *optimal* way. The crucial insight behind Blutner's (2000) bidirectional OT is that for the *hearer* to determine what the optimal interpretation of a given form is, he must also consider the *alternative expressions* the *speaker* could have used to express this meaning/interpretation. One way to implement this idea is to say that we not only require that *the hearer* finds the optimal meaning for a given form, but also that *the speaker* expresses the meaning he wants to communicate by using the optimal form. Thus, what is optimal is not just meanings with respect to forms, but rather form-meaning pairs. According to bidirectional OT the form-meaning pair $\langle f, m\rangle$ is **optimal** iff it satisfies both the S principle (i.e. is optimal for the *speaker*) and the H principle (i.e. is optimal for the hearer):[1,2]

$$\begin{array}{ll} (S) & \neg\exists f' :\ \langle f', m\rangle \in H\ \&\ \langle f, m\rangle < \langle f', m\rangle \\ (H) & \neg\exists m' :\ \langle f, m'\rangle \in S\ \&\ \langle f, m\rangle < \langle f, m'\rangle \end{array}$$

To turn the above definition of optimality into a predictive formalism, we have to know several things: (i) what are the alternative forms? (ii) what are the alternative meanings? and (iii) how should we interpret the ordering relation $<$?

In Blutner (1998, 2000) no restrictions are laid down on what alternative expressions/forms to take into account. Blutner (1998) proposes to let the alternative meanings be Carnapian *state descriptions*. The ordering relation is defined in terms of a *cost*-function, defined in turn on the *complexity* of the forms and the *conditional informativity* of the meanings. The cost of form-meaning pair $\langle f, m\rangle$, $c(\langle f, m\rangle)$, is then $compl(f) \times \inf(m/[[f]])$, where $compl(f)$ measures the complexity of form f; $[[f]]$ is the 'semantic' meaning of f; and $\inf(m/[[f]])$ measures the surprise that m holds when f is true.[3] I will sometimes call $\inf(m/[[f]])$ the *surprisal* that m holds if $[[f]]$ is true. If f is a sentence like 'John said hello to a secretary', we could assume that this gives rise to two interpretations: m, where the secretary is *female*, and m', where the secretary is *male*. Because secretaries are normally female, it holds that $P(m/[[f]]) > P(m'/[[f]])$, i.e., m is a more likely given $[[f]]$ than m' is, and thus $\inf(m/[[f]])$ is lower than $\inf(m'/[[f]])$, $\inf(m/[[f]]) < \inf(m'/[[f]])$. The ordering relation between form-meaning pairs is then defined as expected: $\langle f, m\rangle$ is preferred to $\langle f', m'\rangle$ iff the cost of the former is smaller than the cost of the latter, i.e., $\langle f, m\rangle > \langle f', m'\rangle$ iff $c(\langle f, m\rangle) < c(\langle f', m'\rangle)$. Thus, in particular $\langle f, m\rangle > \langle f, m'\rangle$ iff m is a more likely, or stereotypical, interpretation of f than m'' is. Blutner notes that by using this implementation he comes close to implementing Zipf's (1949) idea that interpretation can been seen as a balance of, on the one hand, the force to minimize the *speaker's effort* by preferring forms with a lower complexity, and, on the other, the force to minimize the *hearer's effort* by selecting the worlds that minimize the (conditional) surprise given the semantic meaning of the expression. What is the enriched, or preferred, meaning of sentence f? It is the union of meanings m such that $\langle f, m\rangle$ satisfies both the S and H principles. By using this mechanism, Blutner (1998) claims to be able to account for scalar and clausal implicatures classically accounted for in terms of Grice's maxim of Quantity, also known as Q-inferences, for the fact that sentences typically get interpreted in *stereotypical* ways (known as I inferences in neo-Gricean pragmatics), and for Horn's (1984) division of pragmatic labor.

2.1 Q inferences

Blutner's bidirectional OT accounts for classical quantity implicatures if we assume that the alternative meanings are worlds. Let's look at the *scalar* implicature derivable from $B \vee C$ that $B \wedge C$ is false and the *clausal* one that B and C are both possible. Let us assume we have four relevant worlds: w_0 where neither B nor C are true; w_1 where only B is true; w_2 where only C is true, and w_3 where both are true. Because $\inf(w/[[B \vee C]])$ has the same value for each world w in which '$B \vee C$' is true (or so let us assume), '$A \vee B$' could be interpreted as $\{w_1, w_2, w_3\}$ as far as the H-principle is concerned. However, w_3 is not optimal for the speaker because there is an alternative expression, '$B \wedge C$', such that the surprisal that w_3 holds after learning that this alternative expression is true is smaller than the surprisal that w_3 holds after learning that $B \vee C$ is true: $\inf(w_3/[[B \wedge C]]) < \inf(w_3/[[B \vee C]])$. As a result, '$B \vee C$' gets the exclusive interpretation: $\{w_1, w_2\}$. Notice that Blutner's bidirectional OT accounts both for the intuition that from the assertion '$B \vee C$' we conclude that '$B \wedge C$' is not true, i.e. the *scalar* implicature, and for the *clausal* implicature that $\Diamond B, \Diamond \neg B, \Diamond C$, and $\Diamond \neg C$ are all true. Notice that although the S principle *blocks* world w_3 from being 'part' of the meaning of $B \vee C$, this blocking is due to the conditional surprise that orders interpretations, and is *not* due to the fact that there is an alternative *cheaper form* that could express this interpretation/meaning. Blocking, in this case, is thus due to the ordering of meanings, which can depend on the expression being used. This analysis of blocking will be important in section 6 of this paper. In the next subsection, however, we will see that bidirectional OT also accounts for blocking due to the existence of more costly alternative expressions.

2.2 I inference and Horn's division of labor

Now we will see how due to the H principle sentences will be interpreted in stereotypical ways, and, due to the interaction of the S and H principles, marked expressions typically get a marked interpretation. Taken together this pattern is known as Horn's division of pragmatic labor. To illustrate, consider the following well-known example.

(1) a. John stopped the car.

b. John made the car stop.

Let us assume that both sentences are semantically true if John stopped the car either in a stereotypical way, m_{st}, or in an unusual way, m_u. In that case we typically interpret (1a) as meaning stereotypical stopping, while (1b) as non-stereotypical stopping. Blutner (1998) shows that this is predicted correctly from the interaction of the S and H principles: In case we learn that either (1a) or (1b) is true, the informativity, or surprisal, of m_{st} is smaller than the informativity of m_u, because the former's probability is higher. Because the complexity of (1b) is not smaller than the complexity of (1a), the sentence (1a) is interpreted as m_{st}. Thus, Blutner (1998) accounts for the intuition that sentences typically get the most plausible, or stereotypical, interpretation. To show that the marked form (1b) gets a marked meaning, notice that the interpretation m_{st} is blocked because there is alternative expression that could express m_{st} in a less complex way. Due

to the interaction of the S and H principles, the unmarked (1a) will get the stereotypical interpretation, while the marked (1b) will get the non-stereotypical interpretation.

2.3 OT and constraints

Although bidirectional OT has become rather popular recently to account for certain linguistic data (e.g. Blutner 2000; Zeevat 1999; Zeevat 2000; Aloni 2001; Krifka 2002), the specific way in which Blutner (1998) implemented the theory as I presented above has not been taken up. It is not assumed anymore that the form-meaning pairs are ordered in terms of an abstract cost-function. In particular, the idea is given up that the possible meanings of utterance B are ordered by the function $\inf(\cdot/[[B]])$ so as to minimize the hearer's effort to interpret. Instead, the analyses are based on Jäger's (2002) proposal to relate bidirectional OT more closely with standard OT approaches: derive the ordering relation between form-meaning pairs from a system of more specific ranked OT constraints, some of which are relevant only for ordering forms, others only for ordering meanings. A number of constraints for ordering meanings are very specific, other are more general and closely related with the assertability constraints of Stalnaker and conversational maxims of Grice.

This new way of doing bidirectional OT opens up many possibilities. But there is also a danger: if one can invent any OT constraint as long as it helps to describe the facts it is not clear to what extend OT is still explanatory. Remember that Blutner's formulation of bidirectional OT was motivated by the reduction of pragmatics to Zipf's general principle of minimizing speaker's and hearer's effort.[4] The main goal of this paper is to show how a number of specific OT constraints used in the literature to account for semantic/pragmatic phenomena can be motivated by, or reduced to, very general principles.

3 Bidirectional OT: Prospects and Problems

We saw that in Blutner's original statement of bidirectional OT the meanings are ordered in terms of one very simple general function: conditional informativity. In this section I want to show both the strength and limits of using this function. In section 3.2 I will argue that Blutner's use of the informativity function gives rise to a number of problems. These problems will motivate us to look for an alternative general function for ordering meanings. Before we come to that, however, I will argue for the strength of Blutner's informativity function: showing that a number of OT constraints proposed to account for some specific phenomenon can be reduced to this one function. The phenomenon to be discussed is *anaphora resolution* and the theory that was made to account for it is *centering theory*.

3.1 Centering in bidirectional OT

Centering theory is a theory designed to make predictions about anaphoric resolution and the interpretational coherence in discourses. The theory was originally stated by Grosz, Joshi & Weinstein (1983) in a procedural way and has

recently been given an attractive Optimality Theoretical *declarative* reformulation by Beaver (to appear).[5] The original procedural implementation makes use of two rules (called rule 1 and rule 2) which Beaver reduces to three violable OT constraints ordered in an hierarchical way.[6] I will show in this section how both the Beaverian constraints *and* the ordering between them follow from Blutner's (1998) original statement of bidirectional OT. This derivation crucially relies on a very similar derivation of the rules of original centering theory proposed by Hasida, Nagao and Miyata (1995). Although my derivation will just be a recoding of their's in Optimality Theoretical terms, the derivation is still worth going through, because it shows how specific constraints used in OT can be motivated independently by an ordering relation between form-meaning pairs that is based on a very abstract and general economically based function that orders meanings.

3.1.1 Centering Theory

The crucial notions of centering theory are the following:

- C_F^n = *forward looking centers*, the semantic entities referred to in the nth sentence in the discourse. They are ranked according to their salience, specified as *grammatical obliqueness*. Ranking is determined by the grammatical functions of the referring expressions in the utterance: (subject > direct object > indirect object > other complements > adjuncts)
- C_P^n = *preferred center of* n = highest ranked element of C_F^n.
- C_B^n = *backward looking center*: the highest ranked element in C_F^{n-1} that is referred to in the n-th sentence.

Centering theory is now based on two very simple ideas: First, that if a pronoun is used in an utterance, its preferred referent is the backward looking center of this utterance, called the *topic* of the previous utterance by Beaver (to appear). Beaver notes that this idea (known as *rule 1*) doesn't have to be stated conditionally once we adopt the OT framework, for now constraints are allowed to be violated. Beaver's constraint, PRO-TOP, to capture rule 1 of centering theory simply says: The topic must be pronominalized. The second idea of centering theory is that it is assumed that a discourse is more coherent when the topic remains constant, i.e. when for each utterance its backward looking center is the same as that of its previous utterance. This means that a discourse is maximally coherent (as far as anaphoric reference is concerned) if for each utterance n it holds that $C_B^n = C_B^{n-1}$ and $C_B^n = C_P^n$.

To illustrate, consider the following discourse:

(2) a. He_1 saw Jack_2 in the park_3.

b. He_4 stopped his car_5.

The three discourse entities/referents referred to in (2a) are DR_1 (He); DR_2 (Jack) and DR_3 (the park). DR_1 is the center (the C_B) of (2a) and also the preferred next center (the C_F) of (2a) and thus the backward center of (2b). Semantically speaking, the pronoun *he* in (2b) could refer back to both DR_1 and

DR_2. Giving the centering theoretical preference, however, it is predicted that it will refer to DR_1.

In the above example none of the centering constraints was violated. But what if one or more of these conditions is not satisfied? Which violation is less dramatic than others? According to *rule 2* of centering theory, transitions are preferred in the following ordering: CONTINUE > RETAIN > SMOOTH-SHIFT > ROUGH-SHIFT, where these names have the following denotations:

CONTINUE:	$C_B^n = C_B^{n-1}$ and $C_B^n = C_P^n$
RETAIN:	$C_B^n = C_B^{n-1}$ and $C_B^n \neq C_P^n$
SMOOTH-SHIFT:	$C_B^n \neq C_B^{n-1}$ and $C_B^n = C_P^n$
ROUGH-SHIFT:	$C_B^n \neq C_B^{n-1}$ and $C_B^n \neq C_P^n$

Beaver (to appear) notes that the ordering on these transitions can be captured straightforwardly when we assume that $C_B^n = C_B^{n-1}$ and $C_B^n = C_P^n$ are both separate optimality theoretical constraints, dubbed COHERE and ALIGN, respectively, and assume that COHERE is more important than ALIGN. By assuming in addition that the constraint PRO-TOP is more important than COHERE (and thus also than ALIGN), Beaver's OT reformulation (called COT) also captures the Centering theoretic claim that their rule 1 is more important than their rule 2.

3.1.2 Deriving the constraints

Although Centering Theory is normally seen as being purely *descriptive* in that it tries to predict pronoun resolution adequately, at least according to Beaver (to appear) its original motivation was economic in nature:

> One of the driving forces behind early Centering proposals of Joshi and associates was the idea that speakers choose forms which minimize processing costs to hearers. COT models the fact that it may be cheaper in the long-run to use a form which is in the short-term relatively expensive. For instance, a speaker may choose a form in which the topic is not in subject position because it will reduce the costs incurred by a *following* sentence in which a topic shift is needed. (Beaver, to appear, p.83)

By assuming that speaker's and hearer's try to minimize their effort, Blutner's bidirectional OT can be seen as a theory of rational language use. This suggests that we should be able to justify the rules of centering theory, or the Beaverian constraints and orderings between them, in terms of Blutner's general formalization of his theory. Following Hasida, Nagao and Miyata, I will suggest that this indeed can be done.

PRO-TOP: The constraint PRO-TOP only has an effect in Beaver's COT if the sentence contains pronouns. In that case it demands that one of them must refer to the backward looking center. To derive this constraint, let us assume that there is no lighter (anaphoric) expression than a pronoun. It follows from bidirectional OT that this pronoun must thus refer to the best possible meaning. Assume now that the semantic meaning of a pronoun is underspecified, and can be

interpreted as any of the elements of the set of forward looking centers of its previous utterance. Let the forward looking centers of utterance $n-1$, C_F^{n-1}, be the list $[a, b, c]$, with $C_B^n = a$. In that case we can assume that the semantic meaning of a pronoun in utterance n should be $\{a, b, c\}$. The elements of this list, however, are ordered by salience. In particular, the most probable referent of a pronoun in the nth utterance is its backward looking center, i.e. a. Thus, $\inf(a/\{a, b, c\})$ is smaller than both $\inf(b/\{a, b, c\})$ and $\inf(c/\{a, b, c\})$. Thus, the backward looking center, i.e. a, will be the best meaning, and, by bidirectional OT, will thus be the interpretation of the lightest anaphoric expression (a pronoun). So we see that PRO-TOP follows straightforwardly from Blutner's bidirectional OT.

COHERE: The constraint COHERE is satisfied iff $C_B^n = C_B^{n-1}$.[7] Notice that this constraint can only be violated when the the highest ranked element of C_F^{n-1} is not the same as C_B^{n-1}. In combination with PRO-TOP, this means that the highest ranked element of C_F^{n-1} could not be referred to in the $n-1$th utterance by a pronoun. Because the use of a pronoun is shorter, and requires less effort, than the use of a proper name or full description to refer to an object, Blutner's bidirectional OT predicts that it is better (for the $n-1$th utterance) to not violate the COHERE constraint.

ALIGN: The constraint ALIGN is satisfied iff $C_B^n = C_P^n$. The reason why bidirectional OT prefers this constraint to not be violated is very similar to the reason why it prefers COHERE to be satisfied, but now related to the nth utterance.

From the above derivations of the Beaverian constraints out of bidirectional OT we can also deduce that PRO-TOP is more important than the other constraints. To derive COHERE for instance, we referred to PRO-TOP, but not the other way around. Moreover, Beaver's ranking between COHERE and ALIGN can be understood also: a violation of COHERE is worse than a violation of ALIGN because the former violation leads to more effort in the $n-1$th utterance, while a violation of ALIGN can only have an effect in the nth utterance. In fact, a violation of COHERE *must* have an effort-like effect, while a violation of ALIGN need not have an effect, because it only puts constraints on the use of pronouns in *future* utterances.

3.2 Problems with original bidirectional OT

Although Blutner's bidirectional OT allows us to account for a number of conversational implicatures and can help to account for pronoun resolution in so far as it is able to explain the underlying principles of centering theory, there are serious problems with his analysis too. A major problem is that the analysis of *scalar implicatures* both *over* and *under*generates.[8]

3.2.1 Overgeneration

Blutner's (1998) original implementation of bidirectional OT overgenerates, because it predicts that whenever the semantic interpretation of B, $[[B]]$, entails the semantic interpretation of C, $[[C]]$, and the expressions B and C are equally complex, the assertion of C will have the scalar implicature that $[[B \wedge C]]$ is not true. The reason is that (on the assumption that worlds are reasonably equally distributed) for all $w \in [[C]]$: $\inf(w/[[B]]) > \inf(w/[[C]])$, which has the result

that for these worlds the form-meaning pairs $\langle$ 'B', $w\rangle$ are *blocked* by the S principle. But this is obviously false. Although we normally conclude from assertion (3a) that the stronger (3b) is not true, we typically don't infer that (3c) is false from the assertion of (3b):

(3) a. John *believes* that Susan is sick.

b. John *knows* that Susan is sick.

c. John *regrets* that Susan is sick.

Suppose $D \models C$ and $C \models B$, and suppose that w is only true in D, v also in C, and u also in B. If we then assume that the worlds are equally distributed, Blutner's formalization gives us the following table:

$\inf(\cdot/[[\cdot]])$	u	v	w
B	$\Rightarrow 1.4$	1.4	1.4
C	$*$	$\Rightarrow 1$	1
D	$*$	$*$	$\Rightarrow 0$

Notice that in this table the values of $\inf(v/[[B]]) = \inf(w/[[B]]) = 1$, for example, because the semantic meaning of C leaves open only two alternative interpretations, $[[C]] = \{v, w\}$, and learning that it should be interpreted as v (or as w) gives us one bit of information. The double arrow indicates how the expressions should be interpreted according to Blutner's formalization. Because D is only true in w, it will be interpreted in that way. C, in turn, is interpreted as v, because (i) w can better be expressed as D, because $\inf(w/[[D]]) < \inf(w/[[C]])$, and (ii) v can better be expressed by 'C' than by 'B' because $\inf(v/[[C]]) < \inf(v/[[B]])$. From this table we can conclude that for any C and D, it holds that if the former entails the latter, we can infer from the assertion that D is the case that C is false. We can get rid of this false prediction, of course, by stipulating that $\langle$know, believe$\rangle$ forms a scale, but $\langle$regret, know$\rangle$ does not, i.e. that C does not belong to the table. However, given the fact that the verbs 'believe', 'know' and 'regret' are lexicalized to the same degree, it is not at all easy to explain this asymmetry.

3.2.2 Pragmatic scales

Blutner's analysis of Q based implicatures, as any other analysis of scalars based on Grice's maxim of quantity, is also *not general enough*, because it cannot account for implicatures first discussed by Fauconnier (1975) and more extensively by Hirschberg (1984) that depend on *scales* where the meanings are logically independent and where the scalar behavior depends on the pragmatic context. For instance, if it is of great value to have an autograph of a famous movie star. However, it doesn't count anymore to have one of Woodward when you already have one of Newman. Thus, we can conclude from (4b) that (4c) is false, but not the other way around:

(4) a. Did you get Paul Newman's autograph?

b. I got Joanne Woodward's.

c. I got Paul Newman's.

An analysis of this scalar implicature in terms of informativity would have to say that (4b) is more informative than (4c). That, however, seems to be unnatural. So, we must agree with Levinson (2000) that there are limits to the use of Bar-Hillel & Carnap's (1953) informativity function to account for scalar implicatures:

> Clearly, there are limits to the utility of such a characterization of informativity (e.g. rather a lot depends on what properties we are actually interested in). But, it is useful as a first approximation. (Levinson, 2000, p. 31)

The point of this section is that the limits of this approximation are rather disturbing. The main goal of this article, however, is to make clear that the claim with which Levinson continues the above quote is simply wrong:

> — and besides, it is just about the only measure of semantic information available. (Levinson, 2000, p. 31)

In the next section I will introduce measures of information, utility, or relevance, that are much more appropriate to account for scalar implicatures than the 'first approximation' used by Levinson and also by Blutner.

4 Maximizing Utility

In this section I will first define a general decision theoretic notion of utility of propositions. I will then show that some specific measures that are found useful in accounting for linguistic phenomena turn out to be natural special cases of this general utility measure.

4.1 Decision theoretic utility

In Savage's (1954) decision theory, actions are taken to be primitives. If we assume that the utility of performing action a in world w is $U(a, w)$, we can define the *expected utility* of action a, $EU(a)$, with respect to probability function P as follows:

$$EU(a) \quad = \quad \sum_{w} P(w) \times U(a, w).$$

Let us now assume that our agent faces a *decision problem*, i.e. she wonders which of the alternative actions in A she should choose. A decision problem of an agent can be modeled as a triple, $\langle P, U, A\rangle$, containing (i) the agent's probability function, P, (ii) her utility function, U, and (iii) the alternative actions she considers, A. If she has to choose now, the agent simply should choose the action with the highest expected utility. But now suppose that she doesn't have to choose now, because she has the opportunity to first receive some useful information.

Before we can determine the utility of this new information, we first have to say how to determine the expected utility of an action conditional on learning

this information. For each action a_i, its conditional expected utility with respect to new proposition B, $EU(a_i, B)$ is

$$EU(a_i, B) \quad = \quad \sum_w P(w/B) \times U(a_i, w)$$

When our agent learns proposition B, she will of course choose that action in A which maximizes the above value: $max_i EU(a_i, B)$. In terms of this notion we can determine the value, or *relevance*, of the assertion B. Referring to a^* as the action that has the highest expected utility according to the original decision problem, $\langle P, U, A\rangle$, i.e. $max_i EU(a_i) = EU(a^*)$, we can determine the *utility value* of the *assertion* B, $UV(B)$, as follows:[9]

$$UV(B) \quad = \quad max_i EU(a_i, B) - EU(a^*)$$

It seems reasonable to claim that in a cooperative dialogue one assertion or interpretation, B, is 'better' than another, C, just in case the utility value of the former is higher than the utility value of the latter, $UV(B) > UV(C)$.

4.2 Special cases

4.2.1 Topic value

The above way to determine the utility value of assertions is very general and follows from general and standard decision theoretic considerations. Now we focus our attention to two special cases, cases where only special kinds of actions are considered and where the utility functions are special too.

If only truth is at stake, a decision problem can be modeled by a partition of the logical space.[10,11] In Shannon's (1948) Information Theory, the **entropy** of partition Q w.r.t. probability function P, $E(Q)$, is defined as $\sum_{q\in Q} P(q) \times \inf(q)$, where $\inf(q)$ denotes the *informativity* of q that Blutner used already to implement his H principle and is defined as $log_2 \frac{1}{P(q)}$. Thus, the entropy of Q is defined as follows:

$$E(Q) \quad = \quad \sum_{q\in Q} P(q) \times log_2 \frac{1}{P(q)}$$

This entropy $E(Q)$ measures the difficulty of the decision: the decision which element of Q is true is hardest when its elements are considered equally likely, and trivial in case one cell has probability 1. New information might *reduce* this *entropy*. Let us now denote the entropy of Q with respect to probability function P after B is learned by $E_B(Q)$:

$$E_B(Q) \quad = \quad \sum_{q\in Q} P(q/B) \times log_2 \frac{1}{P(q/B)}$$

Now we will equate the *reduction* of entropy, $E(Q) - E_B(Q)$, with the *Entropy value* of B with respect to decision problem Q and P, $EV_Q(B)$:

$$EV_Q(B) \quad = \quad E(Q) - E_B(Q)$$

Because learning B might flatten the distribution of the probabilities of the elements of Q, it should be clear that $EV_Q(B)$ might have a negative value. This won't happen when Q has a maximal entropy. The notion of entropy value gives rise to a linear order, $>$, on the usefulness of propositions, and we say that learning B is better than C in case $EV_Q(B) > EV_Q(C)$.

Suppose that partition Q has become relevant in a discourse either implicitly, or due to an explicit question asked by one of the participants, and that this question is very good in the sense that it has maximal entropy with respect to the relevant probability function. Now there are two reasons why B could reduce Q's entropy more than C, i.e., have a higher entropy value: either (i) because it *eliminates more cells* of the partition Q, or (ii) because it changes the probability distribution over the cells, i.e. it makes some cells of Q that have a positive probability more probable than others. Assume that we ignore the latter possibility, i.e., assume that when B is learned, each element of Q consistent with B has equal probability.[12] If we then quantify over probability functions, the above induced ordering relation comes down to the claim that B is better to learn than proposition C just in case B eliminates more cells of partition Q than C does:[13]

$$EV_Q(B) > EV_Q(C) \quad \text{iff} \quad \{q \in Q : \ B \cap q \neq \emptyset\} \subset \{q \in Q : \ C \cap q \neq \emptyset\}$$

It is worth remarking that in this way we have reduced the ordering of propositions in terms of entropy reduction to the ordering between answers that Groenendijk & Stokhof (1984) have proposed.

Can we also think of reduction of entropy itself, i.e. the entropy value of a proposition, $EV_Q(B)$, as a special case of the utility value of this proposition, $UV_Q(B)$ as discussed in the previous subsection? It turns out that we can (see van Rooy 2002 for proof) if we think of the alternative actions the decision maker considers in this case as probability distributions over the elements of Q.

4.2.2 Argumentative value

Ducrot (1973) argued that by making assertions we always want to argue for particular hypotheses, and analyzed linguistic expressions like *but* and *even* in terms of their argumentation orientation. More recently, Merin (1999) proposes to characterize the contexts in which such expressions can be used appropriately in terms of their *argumentative value*, and proposes to implement this argumentative view on language use by means of probability theory. Suppose that an agent wants to argue for hypothesis h and that the relevant information state, i.e. the common ground, is represented by probability function P. Notice that h is statistically dependent on proposition B iff learning B changes the probability of h, $P(h/B) \neq P(h)$. We might say that B is *positively relevant* with respect to h iff $P(h/B) > P(h)$. If $P(h/B) < P(h)$, B would be *negatively relevant*. Now we can define the *argumentative value* of proposition B with respect to hypothesis h, $AV_h(B)$, as follows:[14,15]

$$AV_h(B) \quad \stackrel{def}{=} \quad P(h/B) - P(h)$$

Assuming that an agent wants to argue for proposition h, we can order propositions linearly in terms of their argumentative value with respect to h. Thus,

we can say that B is a better argument for h than C is iff $AV_h(B) > AV_h(C)$. Notice that this ordering relation might behave quite differently from one based on informativity: if B is consistent with h and C is not, $AV_h(B) > AV_h(C)$ even if $C \models B$.

Can we also think of the argumentative value of a proposition as a special case of its utility value? To do so we should resolve two questions: (i) what are the alternative actions? and (ii) what is the natural utility function involved? Notice that just as in the previous case only probabilities are at stake. So, it seems reasonable to assume that the decision problem (for a third participant) is now a choice of a probability measure. For worlds, such a probability measure comes down to a truth-value function. Because the speaker wants to be in a world where h is true, it's a truth value function for h. The utility value can thus be defined as follows:

$$\begin{aligned} U(pr, w) &= 1, \text{ if } w \in h, \\ &= 0 \text{ otherwise} \end{aligned}$$

Now it is easy to see that the argumentative value of B with respect to 'goal' h is a special case of its utility value:

$$\begin{aligned} UV_h(B) &= max_i(a_i, B) - EU(a^*) \\ &= max_i \textstyle\sum_w P(w/B) \times U(pr_i, w) - \sum_w P(w) \times U(pr^*, w) \\ &= \textstyle\sum_{w \in h} P(w/B) - \sum_{w \in h} P(w) \\ &= P(h/B) - P(h) \\ &= AV_h(B) \end{aligned}$$

5 Sperber & Wilson's Relevance as Utility

One of the central maxims of Gricean pragmatics is *Be Relevant*. Unfortunately, Grice stays rather vague about what he means with this maxim. Moreover, the constraint to be relevant seems to be just a qualitative condition, and not one that allows different interpretations to be compared with one another to see in how far they are relevant. Sperber & Wilson (1995) have argued that interpretation is guided by the principle of relevance, stating that sentences should be interpreted as relevantly as possible:

> *The Communicative principle of Relevance*:
>
> > Every utterance communicates a presumption of its own optimal relevance

For this principle to have some predictive force, we have to know what optimal relevance amounts to. According to Sperber & Wilson, the relevance of a proposition/interpretation depends on two factors: (i) the number of *contextual implications* that the interpretation gives rise to; and (ii) the *processing effort* needed to come to this interpretation (Sperber & Wilson, 1995, p. 125).

> *Extend condition 1*: An assumption is relevant in a context to the extend that its contextual effects in that context are large. *Extend condition 2*: An assumption is relevant in a context to the extend the effort required to process it in that context is small.

When does one interpretation, B, give rise to more contextual implications than another, C? At first it seems that this is the case whenever B is *more informative* than C, i.e., meaning that either B entails C, or that B rules out more *worlds* than C does.[16] The principle of relevance then seems to say that only in case B and C rule out equally many worlds, B is better than C if interpretation B is easier to 'grasp' than interpretation C. Although this seems to be Gazdar & Good's (1981), Merin's (1999) and Levinson's (2000) interpretation of Sperber & Wilson's notion of relevance, this can't be the reading they actually had in mind. For in that case it would be impossible to claim with Sperber & Wilson (1995) that in the context of (5a)-(5c), (6b) is not only more relevant than (6a), but also than (6c):

(5) a. People who are getting married should consult a doctor about possible hereditary risks to their children.

b. Two people both of whom have thalassemia should be warned against having children.

c. Susan has thalassemia.

(6) a. Susan, who has thalassemia, is getting married to Bill.

b. Susan is getting married to Bill, who has thalassemia.

c. Susan is getting married to Bill, who has thalassemia, and 1967 was a very good year for Bordeaux wine.

It is obvious that whether informativity is measured in terms of entailment, the number of worlds it eliminates, or the more abstract informativity function, 'inf', of Bar-Hillel & Carnap (1953), (6c) will come out as being more informative than (6b). However, when we think of increase of relevance as increase of utility value, in particular as increase of entropy value $UV(\cdot)$ as defined in the previous section, our analysis arguably makes better predictions. On the assumption that speakers are fully rational, and thus try to maximize their utility, we can assume that the speaker meant that interpretation of the sentence which has the highest utility value for the hearer. Thus, if sentence B with an underspecified meaning gives rise to a number of interpretations $B_i, ..., B_n$, the assumption gives rise to the hypothesis that the speaker meant that the interpretation with the highest utility will be chosen:

$$M(B) \quad = \quad max_i UV(B_i)$$

In case the speaker tries to maximize the entropy value, we have to assume that another agent faces a question that the speaker tries helping to solve. This seems a natural way to account for Sperber & Wilson's claim that (6b) is preferred to (6a). The reason is that in the above discourse two decision problems seem to be important that could be represented by the following two issues/questions (where the *wh*-phrases range over Susan and Bill):

(7) a. Who should consult a doctor?

b. Who should be warned against having children?

If we now assume that the number of contextual implications correlates positively with the number of eliminated cells of the partitions induced by the above questions, we predict that the number of contextual implications due to (6b) and (6c) is higher than that number due to (6a), and that (6c) doesn't give rise to more implications than (6b) does. Utterance (6a) resolves the first issue for Susan and Bill, while utterances (6b) and (6c) resolve also the second issue for both of these individuals. So, it seems not unreasonable to claim that one aspect of Sperber & Wilson's notion of relevance can be captured by our notion of utility.

However, Sperber & Wilson (1995) also claim that (6b) is more relevant than (6c), because the latter gives some extra *irrelevant* information which only costs *extra* interpretation *effort*. Fortunately, there is an easy way to capture this aspect of relevance too. Just say that in case the utility of B equals the utility of C, e.g. eliminates equally many cells of the salient partition, B is still more relevant than C in case the latter gives more information that is useless to solve the decision problem than the former (formally this means that relevance gives rise to a lexicographical ordering):

$$R(B) > R(C) \quad \text{iff} \quad \begin{array}{ll} \text{(i)} & UV(B) > UV(C), \text{ or} \\ \text{(ii)} & UV(B) = UV(C) \text{ and } \inf(B) < \inf(C) \end{array}$$

In case the utility value of proposition B is measured by the number of cells of the relevant partition that is eliminated, the ordering relation induced by relevance is almost the same as the ordering relation discussed by Groenendijk & Stokhof (1984) meant to capture the intuition when one answer is better than another. They claim that when B and C eliminate the same cells of a partition, B is still better than C in case C gives more information that is irrelevant to the question at hand, i.e. when $C \subset B$.

I certainly don't want to suggest that Sperber & Wilson's notion of relevance is fully captured in the way described above. However, by making use of decision theory, a general theory of rationality that also applies to non-cooperative behavior, more aspects of their notion can be captured than just 'being an answer to a question'.

> Achieving optimal relevance, then, is less demanding than obeying the Gricean maxims. In particular, it is possible to be optimally relevant without being 'as informative as required' by the current purposes of the exchange (Grice's first maxim of quantity): for instance by keeping secret something that it would be relevant to the audience to know. It seems to us to be a matter of common experience that the degree of co-operation described by Grice is not automatically expected of communicators. (S & W, 1995, p. 162)

Indeed, when the goal is to make certain kinds of worlds true, or to argue for a particular hypothesis, maximal utility doesn't come down to being 'as informative as required', i.e. to eliminate as many cells of the relevant partition as possible. In these cases the utility of a proposition is its argumentative value, and it might well be that to maximize this value one should not give as much

information as possible: the probability of proposition h might be greater after learning just B, $P(h/B)$, than after learning the more informative proposition $B \wedge C$, $P(h/B \wedge C)$. In that case it is certainly more useful, though perhaps not very co-operative, to say only B.

So, I think that some aspect of Sperber & Wilson's notion of relevance can be captured by our very general decision theoretic notion of utility. In particular their notion of 'number of contextual implications' can be seen as correlating with being a 'good answer to a question'. The other side of their notion of relevance, the notion of 'processing effort' is obviously more difficult to formalize. However, at least some of the intuitions of Sperber & Wilson can be captured by assuming that in case two propositions, or two interpretations of a certain utterance, are equally useful, one is more relevant than another when the former gives less extra information than the latter.

Notice that this lexicographical analysis allows us to account for some examples that typically involve stereotypical interpretations. A sentence like (8a) is typically interpreted as (8b) because it is the most probable meaning:

(8) a. John said 'Hello' to the secretary.

b. John said 'Hello' to the *female* secretary.

We can account for an example like this, as for other so-called I inferences discussed in section 3, if we assume that its stereotypical interpretation and its alternative(s) are equally useful. In that case we predict that the most probable meaning is the most relevant one, giving rise to the stereotypical interpretation.

I don't believe, however, that by taking utility and effort into account in a lexicographical way as suggested above I can analyze successfully all the kinds of examples Sperber & Wilson's notion of relevance is meant to take care of: I predict that a more stereotypical interpretation of an utterance is preferred only if none of the other interpretations is more useful, whereas they seem to suggest that a stereotypical interpretation can be the most relevant one although there might be other interpretations that, after all the processing is done, turn out to have (in my terms) a higher utility value.[17]

> [...] the order in which hypotheses are tested affects their relevance. As a result, the principle of relevance does not generally warrant the selection of more than one interpretation for a single ostensive stimulus.
>
> [...] Consider the following utterance, for instance:
>
> (65) George has a big cat.
>
> In an ordinary situation, the first interpretation of (65) to occur to the hearer will be that George has a big *domestic* cat. [...] the first interpretation consistent with the principle of relevance was the best hypothesis. All other interpretations would manifestly falsify [...] the presumption of relevance. (Sperber & Wilson, 1986, pp. 167-168)

Although my lexicographical analysis of relevance doesn't seem to be fully adequate/sufficient to capture the effects of effort, we will see that by thinking of my notion of 'maximizing relevance' as only one of the two guiding principles of bidirectional OT, some other effects of 'minimizing effort' can be captured in this more general framework.

6 Stalnakerian constraints and Gricean maxims

In section 3 of this paper I have shown how Beaver's (2000) OT constraints used to capture centering theory could be motivated by reducing them to Blutner's general informativity function. In this section I want to do something very similar with respect to other constraints used in OT to account for semantic/pragmatic phenomena. In particular, I want to discuss to what extend Stalnaker's assertability conditions and Grice's conversational maxims can be motivated by the general presumption of optimal relevance/utility in combination with Blutner's bidirectional OT. Grice's maxim of quantity, and the implicatures it is usually said to account for, will be our main concern. Because both Stalnaker and Grice assume that participants of a conversation behave cooperatively, this section will deal almost exclusively with utility value instantiated as entropy reduction.

6.1 Stalnaker's Assertion conditions

In his very influential article 'Assertion', Stalnaker (1978) states 3 principles that have come to be known as Stalnaker's assertion conditions that he claims 'can be defended as essential conditions of rational communication'. Let's see to what extend these three principles can be based upon our decision theoretic approach. I will discuss them in reverse order.

6.1.1 Avoid ambiguity

Stalnaker's third principle basically says that speakers should *avoid ambiguity*. Can this principle be motivated from our decision theoretic point of view? I think we can. First note that according to our analysis, a sentence can be truly ambiguous only if there are at least two interpretations of this sentence that are optimally relevant. Now suppose our hearer faces a decision problem and hears a truly ambiguous sentence. In that case it might be that according to one interpretation the agent is advised to do one action, e.g. a, while according to the other interpretation he is advised to do action b. This has the result that the hearer doesn't know what to do, and, worse, might choose the wrong action. This is certainly something we don't want a cooperative speaker to be responsible for, and thus we shouldn't allow her to use a truly ambiguous sentence.[18]

6.1.2 Presupposition

Stalnaker's second condition advises the speaker to use only sentences that express a proposition in each world of the context, which means that (certain kinds of) its linguistic presuppositions have to already be common ground. It appears to make little sense to make this principle a hard constraint: although the verb

know is normally assumed to trigger a factive presupposition, it is not really problematic to use a sentence like *John knows that Mary is coming* even though Mary's coming is not yet common knowledge. Although such examples seem to violate the principle, it is standardly assumed with Lewis (1979) that the constraint can be rescued by assuming that in these cases we first *accommodate* the context such that the principle holds after all. Be that as it may, it still seem bad conversational practice to change contexts by means of presupposition accommodation. Moreover, some presuppositions seems to be accommodated more easily than others. In fact, in their use of OT to account for semantic/pragmatic phenomena, Zeevat (1999, 2000) and Aloni (2001) propose a violable constraint to ban presupposition accommodation. Can we give an explanation for why this constraint makes sense?

The explanation cannot be straightforward by using our analysis of relevance: presupposition accommodation enriches the context with new information and we saw that new (consistent) information can never have a negative utility. To explain why it is better conversational practice to enrich the context by asserting it than by presupposing it, we have to distinguish the ways in which presupposition and assertion are allowed to change the context. In a rich and very stimulating article, Merin (1999) proposes that (argumentative) relevance helps here: he claims that presupposition B is allowed to be informative with respect to the context, but that this new information should not have a positive relevance. I find this proposal very intuitive, but I don't think it can be a hard constraint: though perhaps not very polite, I find it sometimes a useful strategy to influence people *indirectly* by means of presupposition. Moreover, it is unclear to me how the presumption of optimal relevance can explain Merin's proposal.

Although I am not able (yet?) to explain the ban on accommodation by a presumption of optimal relevance,[19] a closely related principle proposed by Van der Sandt (1992) that prefers binding to accommodation seems to have a natural relevance-theoretical explanation. The principle says that if new information is accommodated to the context, it is better to *bind* this new information to already existing *discourse referents* of the context than to introduce new such referents. The 'motivation' for this principle is based on the fact to be discussed in section 6.2 that in special cases maximizing utility comes down to maximizing informativity. If the context already contains the information that a certain (underspecified) individual has property P, and it is presupposed (by a presupposition trigger like *too*) that somebody has property Q, it is more informative to assume that it is the same individual having property P and Q than to assume that the properties are distributed over (possibly) different individuals ($\inf(\exists x[Px \wedge Qx]) > \inf(\exists x Px \wedge \exists y Qy)$). In fact, this explanation is the natural analogue to Levinson's (2000, p. 273) explanation of why co-reference is preferred to disjoint reference. Unfortunately, however, I am not at all convinced of this explanation of the preference for co-reference. I find explanations in terms of maximizing *coherence* between clauses proposed by proponents of centering theory, and by authors like Hobbs (1979) and Asher & Lascarides (1998) much more natural. Remarkably enough, as we saw in section 3.1, the centering theoretical explanation for the preference for coreference can be motivated by the opposite assumption that expressions should be interpreted in the *least surprising* way: the interpretation selected is the one for which the (conditional) informativity is *lowest*. As we saw

in section 5, this follows from the presumption of optimal relevance (from the hearer's point of view) only if we make the counterintuive assumption that the utility of the resulting interpretation of the sentence in which the pronoun occurs is independent of the choice of reference of the pronoun.

6.1.3 Be Consistent!

Stalnaker's first assertion conditions demands two things: (i) to be consistent, and (ii) to be informative. To motivate (i), we have to see why inconsistency is bad.

Suppose B is inconsistent with $W(P) = \{w \in W : P(w) > 0\}$. Now there are two possible explanations. According to the standard way we say that $P(C/B)$ is undefined in case $B \cap W(P) = \emptyset$. It seems reasonable to stipulate that in that case $UV(B)$ is undefined as well, 'explaining' why learning information inconsistent with the context is bad. But we are obviously able to learn new information that is blatantly inconsistent with what we believed before. Can we give a decision theoretic motivation for why speakers should be consistent with what is commonly assumed even if we take this fact seriously? Suppose we allow $P(\cdot/B)$ to be defined even though B is inconsistent with $W(P)$, but that the result will be that P is *revised* by new information B, resulting in probability function $P^*_B(\cdot)$,[20] with the effect that $W(P) \cap W(P^*_B) = \emptyset$. The problem of revision with inconsistent information, however, is that it is normally not clear what the best way to do so is: there are typically more alternative ways to revise ones belief state that are equally optimal. In our case this means that there are typically several P^i_Bs that count as optimal revisions of P by B. Because the agent can't choose between them, he doesn't. He either feels 'ambiguous' about which belief state he is in, and the motivation given in the previous subsection applies here as well. Alternatively, (but less naturally) we might represent his belief state as a linear combination of the optimal probability functions after revision. According to this latter possibility, many more worlds will be consistent with the new probability function than with the old one. This has the result that there might be many more actions than the ones considered before that could be optimal in (at least) one of the worlds consistent with what is believed, which means that the *risk* that our agent will choose the wrong action has *increased*.

6.1.4 Be informative!

The second part of Stalnaker's first assertion condition demands that new information has to be *informative* with respect to what is commonly assumed, i.e. the context represented by our probability function P. Suppose now that our utterance has B as a relevant interpretation and thus has a strictly positive utility value: $UV(B) > 0$. Then it is easy to see that this interpretation must also be informative, i.e. incompatible with at least some worlds in $W(P)$.

If $UV(B) > 0$, it has to be the case that $max_a \sum_w P(w/B) \times U(a, w) > \sum_w P(w) \times U(a^*, w)$. This, however, can be the case only if either learning B has the result that an action different from a^* has the highest expected utility afterwards and thus will be chosen, or the preferred action remains the same, but the expected utility of this action is higher after learning B than before. But either one of those can happen only in case B at least eliminates some worlds

in $W(P)$ and thus is informative. Because the entropy value and argumentative value are both special cases of utility value, we have shown that old 'news' can never be useful. The other way around, however, doesn't follow: A proposition can be informative with respect to probability function P without being relevant.

6.2 Maximal informativity: the I principle

In the previous subsection we saw that a necessary condition for a proposition to be relevant, or useful, is to be informative. In case an utterance allows for more than one interpretation, our analysis predicts that the preferred one should at least be informative. According to Atlas & Levinson's (1981) and Levinson's (2000) I principle and Horn's (1984) R principle something more is demanded: the preferred interpretation is the one which is *maximally informative*.[21] Although, as we saw in section 3, there are good reasons not to assume this principle in its full force, to account for a certain range of phenomena it seems to predict correctly. In this section I show that in certain special circumstances usefulness reduces to informativity.

6.2.1 Entropy value

First, note that it is obvious that in case B *eliminates* more *cells* of the relevant partition than C does and cells are taken to be as fine-grained as worlds, eliminating more cells means eliminating more worlds. On the extra assumption that the cells of the partition are equally likely, it also means that B has in that case a higher 'inf' value.

Second, this result generalizes quite straightforwardly when relevance is measured in terms of *reduction of entropy*. If W is the set of all worlds, the entropy value of proposition B, $EV_W(B)$, is then $E(W) - E_B(W)$. It is obviously the case that this value gets higher when the entropy of W conditional on B gets lower. Thus, if we can show that $E_B(W) < E_C(W)$ iff $\inf(B) > \inf(C)$, we show that in these special cases maximizing entropy reduction comes down to maximizing informativity. As shown in van Rooy (2002), this can indeed be done in case the probabilities are equally distributed over the worlds. To illustrate, notice first that for every world w it holds that $w \in B$ or $w \notin B$, so that we can equate $E_B(W)$ with $\sum_{w \in A} P(w/B) \times -log_2 P(w/B)$. Suppose now that we have 8 worlds, and that $P(B) = 1/4$. Then B is true in 2 of the 8 worlds, and thus $E_B(W) = 2 \times (\frac{1/8}{1/4} \times -log_2 \frac{1/8}{1/4}) = 2 \times (1/2 \times -log_2 \frac{1}{2}) = 2 \times 1/2 = 1$. Now suppose that $P(C) = 1/2$, and thus that C is true in 4 of the 8 eight worlds. In that case it holds that $E_C(W) = 4 \times (\frac{1/8}{1/2} \times -log_2 \frac{1/8}{1/2}) = 4 \times (1/4 \times 2) = 2$. Because $E_B(W) < E_C(W)$ it also is the case that $EV_W(B) > EV_W(C)$. We can conclude that in these special circumstances the relevance of proposition B is higher in case its probability is lower, which holds exactly when its informativity value, $\inf(B)$, is higher. Thus, in these circumstances reduction of entropy is monotone increasing with respect to informativity, and maximization of the one comes down to maximization of the other.

6.2.2 Argumentative value

Finally, we can show that in special cases the *argumentative value* of a proposition is also monotone increasing with respect to its 'inf' value. In section 4.2.2 we argued that proposition B has a positive argumentative value with respect to h, i.e. $AV_h(B) > 0$, just in case $P(h/B) > P(h)$. Notice that $P(h/B) > P(h)$ iff $P(h/B)/P(h) > 1$ iff $P(B/h)/P(B) > 1$. In fact, the measure $P(\cdot/h)/P(\cdot)$ is continuously monotone increasing with respect to our $AV_h(\cdot)$, meaning that if the one gets higher (lower), the other gets higher (lower) too. Notice that when $h \models B$, $P(B/h)/P(B) = \frac{1}{P(B)}$. The function $\frac{1}{P(\cdot)}$, in turn, is continuously monotone increasing with respect to Bar-Hillel & Carnap's (1953) informativity function, because $\inf(\cdot) = log\frac{1}{P(\cdot)}$. Thus, if h entails the arguments given, the measure $P(\cdot/h)/P(\cdot)$ is continuously monotone increasing with respect to $\inf(\cdot)$. But this means that in these cases also our $AV_h(\cdot)$ is continuously monotone increasing with respect to $\inf(\cdot)$. We can conclude that in special circumstances the requirement to select the maximally relevant interpretation of a sentence comes down to selecting its most informative interpretation.

6.2.3 Sufficiently informative

In this section I have interpreted the I principle as the demand to interpret the sentence in the most informative way in the sense of Bar-Hillel & Carnap's (1953) informativity function 'inf'. Although Horn and especially Levinson make use of the I (or R) principle under this interpretation, their explicit statement of the principle actually demands only that the most informative interpretation 'sufficient to achieve your communicational ends' (Levinson, 2000, p. 114) be taken. And indeed, under this interpretation the I principle is close to what Grice's (1989) second maxim of Quantity asks for. Notice that in case relevance is measured in terms of entropy value, we might say that informativity is measured with respect to the goals/topics the discourse participants are interested in. Before we discuss such an interpretation of Grice's maxim, however, it is useful to first discuss his maxim of manner, and see to what extend our analysis can capture it.

6.3 Manner

Grice's maxim of manner asks the speaker to *be perspicuous*, which by itself gives rise to the following four (sub)maxims:

1. Avoid obscurity of expression.
2. Avoid ambiguity
3. Be brief (avoid unnecessary prolixity).
4. Be orderly.

As Grice notes himself, the maxim of manner is rather different from the others because it relates 'not (like the previous categories) to what is said but, rather, to HOW what is said is to be said'. Still, the first two submaxims can, I believe,

be motivated by our general decision theoretical approach in the same way as I motivated Stalnaker's third assertion condition. The other two submaxims seem to be very close to Zipf's principle of *minimizing effort*, a principle that was already captured adequately, or so we argued, by Blutner's interpretation of the S principle in his bidirectional OT.

This, then, suggests a way of combining the presumption of optimal relevance/utility with bidirectional OT: Blutner's S principle stays as it is, capturing Grice's last two submaxims of manner and part of Zipf's minimization of effort, but his ordering on interpretations used in the H principle should be induced (at least in a number of cases) by the above discussed notion of relevance.[22] If we do that, we are ready to see to what extend we can account for the effects of Grice's maxim of Quantity.

6.4 Quantity and Q implicatures

6.4.1 The maxims and their interpretations

Grice's maxim of quantity talks about the quantity of information to be provided, and thus seems most closely related with our quantitative analysis of relevance. Quantity comes with the following two maxims:

1. Make your contribution as informative as is required (for the current purposes of the exchange).

2. Do not make your contribution more informative than is required.

As we have seen in section 1, the first maxim of quantity was interpreted by Horn, Gazdar and others as meaning something like 'say as much as you can', and this maxim was taken to be responsible for many so-called *generalized* conversational implicatures: the *scalar* and the *clausal* ones. However, as noted by Gazdar (1979), it is not straightforward to interpret and/or formalize the maxim in its full generality:

> To formalize this maxim as it stands, that is in its full generality, we would have to (a) be able to quantify over informativeness, and (b) have some function which when applied to a conversation and a point within it would yield as its value the level of informativeness required. (Gazdar, 1979, p. 49)

It seems that our analysis of topic-dependent relevance, i.e. entropy value, provides exactly what Gazdar asked for. According to our treatment, B can only have a higher entropy value than C in case it is more informative. Moreover, our topic-dependent analysis of relevance also says in what sense a sentence can be more informative than required: In case proposition B resolves the decision problem, any stronger proposition C will resolve the decision problem too. In that case, however, C will give extra, irrelevant, information and, according to our analysis of relevance, interpretation B is then preferred to interpretation C.

So, how does the presumption of optimal relevance relates to Grice's Quantity maxims? First, it predicts that an agent might give information that is maximally relevant without being 'as informative as required'. In case you want to argue for

hypothesis h, or make true a world where h holds, it might be more useful to say less than is required. In those cases, of course Grice's cooperative principle is not at work, so the deviation should not come as a big surprise

However, when we limit ourselves to utility as entropy reduction, the proposal to ask for the hearer to interpret the utterance as maximally relevant seems to have a straightforward connection with Grice's Quantity maxims. But first we have to make clear how we understand these maxims.

According to the standard reading of the first maxim of quantity, as interpreted by Horn, Gazdar, Levinson and others, it says that the speaker could not make an alternative claim relevant to the conversation with a *stronger/more specific* conventional/semantic meaning. In this reading this maxim is responsible for the standard treatment of scalar and clausal implicatures.

The second maxim of quantity is normally (e.g. Horn, Levinson) taken to mean the opposite: it allows the speaker to use a sentence with a very *weak/general* conventional/semantic meaning if she can rely on the hearer to interpret the sentence in the intended stronger/more specific way because that is what the purpose of the exchange requires.

The proposal to ask the hearer to interpret the utterance as relevantly as possible seems in accordance with Grice's second submaxim of quantity, but in contradiction with the first one.

However, there might be another way to interpret Grice.[23] Suppose that Grice, in stating his maxims, already took the hearer's perspective into account. In that case, Grice's first maxim of quantity says something very close to our demand to choose that interpretation of a sentence which has the highest entropy value, while the second maxim can then be interpreted as saying that in case two interpretations of a sentence have an equally high entropy value (for instance, if both completely resolve the issue), the less informative one is preferred. On this reading of Grice, Quantity reduces to our lexicographical definition of relevance repeated below:

$$R(A) > R(B) \quad \text{iff} \quad \begin{array}{ll} \text{(i)} & UV(A) > UV(B), \text{ or} \\ \text{(ii)} & UV(A) = UV(B) \text{ and } \inf(A) < \inf(B) \end{array}$$

Notice that it is the first submaxim that is standardly used to derive scalar and clausal implicatures. However, as we saw in section 3, this maxim is also responsible for the overgeneration: for instance, it is not clear how we can rule out a scale like $\langle Regret, Know \rangle$ other than by stipulation. As observed by Groenendijk & Stokhof (1984), and earlier by Atlas & Levinson (1981), it is also responsible for the false prediction that answers to *Who*-questions are typically *not* interpreted as being exhaustive, for otherwise this would have been done explicitly. Perhaps, despite its overwhelming use in (neo)-Gricean pragmatics, we can, and thus should, do without Grice's first maxim of Quantity once we have our principle of optimal relevance together with bidirectional Optimality Theory. In the remainder of this paper we will see how far we can pursue this line of thought.

6.4.2 'Exactly' interpretation of numerals

A first example. We have to account for the fact that in most contexts number terms get an 'exactly' interpretation. At the same time (cf. Kempson (1986),

Kadmon (1987), Zeevat (1994), van Kuppevelt (1996)), the analysis should also explain why the sentence

(9) John has 3 children

does not get this 'exactly' interpretation when given as an answer to the question:

(10) Does John have 3 children?

Let us assume with neo-Griceans that number-terms semantically get an 'at least' meaning. In that case the second maxim of Quantity, and our maxim of optimal relevance, seem to do the trick: When the question is

(11) How many children does John have?

the question gives rise to the partition $\{\lambda v[\text{John has exactly } n \text{ children}] : n \in N\}$, where each cell contains only worlds where it is true that John has *exactly* n children. Thus, the exact number of children that John has is relevant, and we should look for the most informative reading of (9). But what is this most informative reading? Assuming that one reading is more informative than another if it eliminates more cells of the partition, it should be a reading of the form $\lambda v[\text{John has exactly } n \text{ children}]$ that is compatible with the semantic meaning of (9): $\lambda v[\text{John has at least } n \text{ children}]$. Intuitively, this most informative reading should be the one saying that John has exactly 3 children. Unfortunately, informativity by itself cannot enforce this reading: The reading 'exactly 3' is not the only one compatible with (9) when numerals have an 'at least' meaning; a reading like 'exactly 4' is so too. Why should (9) not be interpreted as an exhaustive answer incompatible with John's having exactly 3 children? The question seems silly, but this is only so because we take the answer to be so obvious: because for the other cells we use other numbers. Thus, alternative expressions should come into the picture after all. And with the alternative expressions, also Blutner's bidirectional OT.

If we then assume that the probabilities are equally distributed over the worlds, that it is already assumed that John has children, but not more than 4, a bidirectional formalization in terms of relevance gives rise to the following table, with the desired outcome:

$EV_{(11)}([[\cdot]])$	1	2	3	4
'1'	$\Rightarrow 0$	0	0	0
'2'	$*$	$\Rightarrow 0.4$	0.4	0.4
'3'	$*$	$*$	$\Rightarrow 1$	1
'4'	$*$	$*$	$*$	$\Rightarrow 2$

Notice that '3' doesn't mean 4 because this meaning is *blocked*: there is another expression for which 4 is a better meaning, i.e. a meaning with a higher relevance.[24]

6.4.3 Cancellation

Notice that this much could have been done already by Blutner's (1998) ordering relation between meanings in terms of (conditional) informativity. But we also saw that that analysis was both (i) too general, and (ii) not general enough. Let us first discuss the cases that Blutner's ordering in terms of informativity could not account for.

First, when the question answered by (9) is not (11) but (10) instead, the answer intuitively does not rule out that John has 4 children. When the meanings/worlds are ordered by conditional informativity $\inf(m/[[\cdot]])$, however, this is what is predicted: informativity alone doesn't care about relevance. To make correct predictions, Blutner (1998), following standard analyses of conversational implicatures, would have to allow for implicatures that can be *canceled* for reasons of relevance. When the ordering depends on relevance, on the other hand, things are different. In that case both answer '3' , i.e. (9), and answer '4' would have a relevance of 1 in worlds where John has 3 or 4 children. Because '3' and '4' are equally complex, bidirectional OT predicts that (9) now does not give rise to the inference that John doesn't have more than 3 children. Thus, no cancellation is needed, just like Kadmon (1987), Zeevat (1994), and van Kuppevelt propose.

6.4.4 Exhaustivity

Above we have seen that bidirectional OT predicts that answers involving numerical terms are interpreted *exhaustively*. Groenendijk & Stokhof (1984) make use of an explicit exhaustivity operator to account for this. But their operator accounts not only for standard *scalar* inferences, but also for the intuition that when (12b) is given as an answer to (12a), the answer is interpreted as meaning that *only* John went to the party:

(12) a. Who went to the party?

b. John went to the party.

As Groenendijk & Stokhof (1984) note themselves, this is certainly not an inference following from Grice's first maxim of Quantity. That maxim would rather predict that the answer should *not* be interpreted exhaustively. However, the inference *does* follow in bidirectional OT from the assumption that answers should be interpreted maximally relevant. Suppose that only a and b are the relevant persons for question (12a). In that case the bidirectional table looks as follows:

$EV_Q([[\cdot]])$	$\emptyset$	a	b	ab
'Nobody'	$\Rightarrow 2$	$*$	$*$	$*$
'a'	$*$	$\Rightarrow 1$	$*$	1
'b'	$*$	$*$	$\Rightarrow 1$	1
'ab'	$*$	$*$	$*$	$\Rightarrow 2$
'not a'	1	$*$	1	$*$
'not b'	1	1	$*$	$*$
'not a and not b'	2	$*$	$*$	$*$

Notice that in this table complexity plays a crucial rule. Meaning b, for example, is expressed by 'b' and not by 'not a' because the former is less complex than the latter. From this table we can conclude that in this example (12b) is predicted to mean that John was the *only* one who went to the party. This seems perfect. Still, as we will see in section 7, the analysis of exhaustivity can't be so straightforward anymore once we look at examples just a little bit more complicated than the one discussed here. But before we come to that, let us first discuss some cases that can't be handled straightforwardly by making use of Groenendijk & Stokhof's explicit exhaustivity operator, but that are unproblematic on our account.

6.4.5 Mention some

To account for the intuition that (12b) is treated as an *exhaustive* answer to question (12a), we have assumed that the decision problem is which answer to question (12a) is true, and that the question itself gave rise to a partition. But we might give up both of these assumptions. First, we might assume that the question is not represented as a partition, but treated as a mention-some question where its answers might overlap. If the worlds are the same as in the above example, and if it is assumed that at least one of $\{a, b\}$ went to the party, the question can be represented as $\{\{a, ab\}, \{b, ab\}\}$. Second, we might propose that the question is still represented as a partition, but that the decision problem is such that one action is best in world a, the other in world b, but both are equally good in world ab. Whether we now determine the relevance of answers with respect to the non-partitional question in the first case, or with respect to the decision problem in the second, the entropy value will be the same. In both cases the possible answers give rise to the following table:

$EV_Q([[\cdot]])$	a	b	ab
'a'	$\Rightarrow 1$	$*$	$\Rightarrow 1$
'b'	$*$	$\Rightarrow 1$	$\Rightarrow 1$
'ab'	$*$	$*$	1
'not a'	$*$	1	$*$
'not b'	1	$*$	$*$

Notice that in this case (i) the answers 'a' and 'b' are not interpreted exhaustively, and (ii) it is predicted that answer 'ab' will not be given, because there is no need to specify this world separately by the use of a more costly expression. It seems to me that both predictions are born out by the facts.

6.4.6 Pragmatic scales

We have seen in section 2 that informativity (alone) cannot account for the fact that in the context of question (4a), repeated here as (13a), we conclude from (13b) that (13c) is true, but we don't infer (13e) from (13d):

(13) a. Did you get Paul Newman's autograph?

b. I got Joanne Woodward's.

c. I didn't get Paul Newman's

d. Yes/I got Paul Newman's.

e. I didn't get Joanne Woodward's

An analysis in terms of relevance can do much better, but now we have to use Merin's (1999) notion. This seems reasonable in this case: the answerer wants to convince the questioner to accept that we are in a world where she has an autograph of somebody with a high prestige, and, if possible, an autograph with a higher prestige than the questioner himself. Let us assume that the questioner does not yet know that the answerer got an autograph of a famous movie star in the first place, that having an autograph of such a person is of great value, but that it doesn't count anymore to have one of Woodward when you already (or also) have one of Newman. In that case we get something like the following table (where the numbers might be different, but the ordinal relations between the numbers remain the same):

$AV_h([[\cdot]])$	$\neg N \wedge \neg W$	$\neg N \wedge W$	$N \wedge \neg W$	$N \wedge W$
'No'	$\Rightarrow 0$	0	0	0
'Woodward'	$*$	$\Rightarrow 0.7$	$*$	0.7
'Woodward and not Newman'	$*$	0.7	$*$	$*$
'Yes'	$*$	$*$	$\Rightarrow 1$	$\Rightarrow 1$
'not Woodward and Newman'	$*$	$*$	1	$*$

Notice that in this case the answers where both persons are mentioned are ruled out for reasons of speaker effort, and that relevance does the rest.

6.4.7 Limiting overgeneration

In this section we have seen that by replacing the informativity ordering relation on meanings by one of relevance, we can account for more scalar implicatures then before. But this analysis overgenerates neither as much as the ordering relation that Blutner proposed, nor as much as the standard neo-Gricean (e.g. Horn, Levinson) treatment of scalar implicatures in terms of Grice's first maxim of quantity. In section 3 we saw that ordering by informativity wrongly predicts that if B follows from C, the assertion 'B' always gives rise to the implicature that C is false. This prediction doesn't follow anymore once we order meanings in terms of relevance. The reason is that although B might follow from C, this doesn't necessarily mean that in the $B \wedge \neg C$-worlds assertion 'C' has a higher relevance with respect to the question under discussion than assertion 'B'. In fact, if the extra information that C asserts on top of B is *irrelevant* to the topic of the conversation, it is predicted that the relevance of C in those worlds is *lower* than the relevance of B. For instance, in case the question is how sure John is that Susan is sick, it is predicted that in every world where John knows that Susan is sick, (14a) has a higher relevance than (14b):

(14) a. John *knows* that Susan is sick.

b. John *regrets* that Susan is sick.

This gives rise to the correct prediction that in the context of such a question (14a) does not give rise to the inference that (14b) is false. I conclude that in combination with bidirectional OT, the assumption of optimal relevance predicts better with respect to scalar implicatures than Grice's first maxim of quantity under its standard reading.

Green (1995) has argued that the wrong prediction of neo-Griceans is due to a wrong reading of Grice's first maxim of Quantity. Neo-Griceans have standardly assumed that Quantity 1 means that the speaker is making the strongest statement she is able to make on the matter at hand (i.e. saying as much as she can). Green argued that Grice only requires, however, that the speaker makes a contribution which is (at least) as informative as is required, i.e. *informationally sufficient*. But if that is so, and if we also assume that Quantity 2 means that the speaker should not say something stronger than is required, it seems that Grice himself already correctly predicts that in the context of the question described above (14a) doesn't give rise to the implicature that (14b) is false. I think Green gives a new, interesting, and empirically more adequate, interpretation of Grice's maxim. Be that as it may, to formalize this reading of Grice, we have to say what it means to be as informative *as required*. To account for that, however, it seems we still need a notion of relevance. The purpose of this subsection, however, was to argue that once we have a notion of (optimal) relevance, in combination with bidirectional OT, we do not need the Gricean maxim of quantity anymore.

7 Maximization of relevance as Exhaustification

In the previous section we have seen how our use of relevance in bidirectional OT explains why an answer like *John went to the party* to the question *Who went to the party?* is typically interpreted exhaustively when the interrogative sentence should be interpreted as a mention-all question. But I noted already that things are not as straightforward as they seem. There are (at least) two reasons for this: (i) we limited ourselves to the simple case where only a few individuals were taken to be relevant; (ii) we considered only how to encode the *cells* of a partition and have not taken *partial answers* into account. With respect to the second problem, we have not discussed yet the perhaps most obvious problem for the standard analysis of scalar implicatures: the fact that from the answer 'a or b' to the question *Who is coming?* it is wrongly predicted that neither a nor b will come. The reason for this false prediction is that both the answer a and the answer b would entail the answer actually given, and thus, by the standard reading of quantity 1, are ruled out.[25] Our analysis does not generate this problem, but gives rise to another one: how should we interpret 'a or b' in the first place, and how can we explain that such a disjunctive answer normally gives rise to an *exclusive* reading? One might try to extend the bidirectional analysis by taking more alternative expressions into account, and also more meanings than just the cells of the partition. As it turns out, this is not a trivial enterprice. Instead of getting involved into this enterprise, let me discuss another problem of our approach which suggests a somewhat different line of attack.

In sections 2 and 3, I have shown the potential of bidirectional OT when meanings are ordered in terms of Blutner's conditional informativity function. After that I have argued that with this way of ordering meanings we encounter

difficulties in accounting for certain examples and have shown that bidirectional OT makes better predictions if we assume with Sperber & Wilson that sentences are interpreted as relevantly as possible. To account for that we assumed that meanings are ordered in terms of our decision theoretic notion of utility. Although we saw in the previous section that by making use of relevance/utility in bidirectional OT we can account for many Q-implicatures, it should be clear that such an analysis is not really suited to account for I-implicatures. To account for these latter kind of implicatures we had to assume that the information given is *irrelevant*. Our discussion of why stereotypical interpretations, and in particular co-referential interpretations of pronouns, are preferred suggested, however, that this assumption is implausible and that our lexicographical analysis of relevance isn't quite satisfactory. Thus, it seems that if we want to account for implicatures in terms of a single general function, we either have to use something like the conditional *informativity* function as used by Blutner, or the assumption that we interpret things as *relevantly* as possible and account for that in decision theoretical terms. If we choose for the first option, we can account for I-implicatures to stereotypical interpretations, but we can't account for Q implicatures. If we go for the second option, however, it are rather the I-implicatures that we cannot account for adequately anymore. So it seems that our search for a single general principle in terms of which all kinds of implicatures can be handled ended unsuccessfully. In this final main section, however, I want to suggest that prospects are not that dim.

The new idea is to shift once again to another reading of Grice's maxims. First, we followed Blutner (1998) in taking his interpretation principle that is based on the conditional informativity function as an implementation of Grice's *first* submaxim of quantity as understood by Horn, Levinson and others: Say as much as you can! Afterwards, we have used utility in accordance with Sperber & Wilson's principle to *interpret* sentences as maximally relevant, which can be based on Grice's *second* submaxim of quantity: Don't say more than you must! But perhaps we should make use of utility not from the hearer's, but rather from the *speaker*'s point of view. In that case it seems natural to use utility to interpret Grice's first maxim of quantity, so that it reads: Speak as relevantly as you can! From our earlier discussion it seemed that if we want to account for Q-implicatures in terms of Grice's first maxim of quantity, we have to make crucial use of *alternative expressions*. This use of alternative expressions, however, was seen to be dangerous: without limitations the analysis would overgenerate enormously. In this final section I would like to suggest that by adopting an *exhaustivity operator*, – in fact by changing Groenendijk & Stokhof's (1984) context-independent exhaustivity operator into one that is based on a relevance-ordering –, we can actually account for both I-implicatures and Q-implicatures with just one operation.

Groenendijk & Stokhof (1984) propose to account for the intuition that the answer *Peter comes* to question *Who comes?* should normally be read exhaustively by introducing an explicit exhaustivity operator that is applied to answers and the abstracts (predicates) underlying the questions to derive the exhaustive interpretation. Although their exhaustivity operator is very appealing and predicts correctly when assertions are given as answers to so-called *mention-all* questions, it also faces some crucial problems. First, it gives the wrong result if

applied to answers given to *mention-some* questions. Second, it cannot account for Hirschberg's examples of *scalar* readings. To solve both of these problems, the following exhaustivity operator can be defined (see van Rooy & Schulz, 2003) which is dependent on a relevance-ordering '>':

$$[[exh]] \quad = \quad \lambda T \lambda P.\{w \in W| \quad P(w) \in T(w) \wedge \neg \exists t \in T(w) : \lambda v[P(w) \subseteq P(v)] > \lambda v[t \subseteq P(v)]\}$$

This operator takes a term-answer T and a question-predicate P and turns it into a proposition. Described informally, it does the following: in each world, T denotes a generalized quantifier, i.e., gives a set of possible extensions for P. exh takes all these possibilities $t \in T(w)$ and compares the utility value of the propositions $\lambda v[t \subseteq P(v)]$. P can only be one of these possibilities that are minimal values in this order. This exhaustivity operator can be thought of as a generalization of Groenendijk & Stokhof's exhaustivity operator. The two operators give rise to (almost) identical results in case the relevance ordering '>' reduces to entailment, or the subset relation '$\subseteq$'. As a consequence, our operator accounts for many of the implicatures traditionally accounted for in terms of Grice's maxim of quantity. Just like our above OT tables, it accounts for the fact that when *Who came?* is answered by *John* we conclude that *only* John came. However, it also accounts for exhaustive interpretations of explicit partial answers, like disjunctive answers like *John or Bill* or an indefinite answer like *A man*. From the latter answer we can conclude by means of exhaustive interpretation that not all men came, an implicature standardly triggered by the ⟨all, some⟩ scale. The analysis also accounts for the exclusive reading of disjunctive sentences: if (15a) is answered by (15b), the latter is interpreted as (15c) after application of our exhaustivity operator:

(15) a. Did John walk?

b. John walked or Mary walked.

c. John walked or Mary walked, but not both.

Because the relevance relation '>' need not come down to entailment, our exhaustivity operator can account for phenomena Groenendijk & Stokhof cannot account for. First, it has no problems with answers given to mention-some readings of *wh*-questions as discussed in section 6.4.5. In those cases we predict that exhaustification has no effect. Second, the ordering relation on which we base our analysis of exhaustivity might come down to, for instance, autographic prestige, which means that also the examples in (12) can be handled correctly.

Notice that our exhaustification analysis not only predicts intuitions standardly accounted for in terms of the Q principle; also some I-implicatures are accounted for. Just like for Groenendijk & Stokhof's operator, we predict that if the question is *Who quacks?* the answer *Every duck quacks* is predicted to imply that every quacker is a duck. Horn (2000) calls this inference *conversion* and explicitly proposes to account for it in terms of the I-principle. Something similar holds for the inference from *if* to *if and only if*.

Studying Horn (1984) and Levinson (2000) carefully, one sees that two very different kinds of inferences are supposed to be accounted for in terms of the

I-principle. On the one hand, we have the *strengthening* inferences as discussed directly above, from *if* to *if and only if*, for example. More typical I-implicatures, however, are inferences from a sentence to its *stereotypical*, or most probable, interpretation. As we will see, we can capture these I-implicatures by means of an operator that is very close to our exhaustivity operator.

The exhaustivity operator given above is defined in terms of an ordering based on utility. As we saw in section 6.2.1, however, in special cases this utility ordering reduces to one based on informativity. In that case the exhaustivity operator looks as follows:

$$[[exh]] = \lambda T \lambda P.\{w \in W| \; P(w) \in T(w) \wedge \neg \exists t \in T(w) : \inf(\lambda v[P(w) \subseteq P(v)]) > \inf(\lambda v[t \subseteq P(v)])\}$$

Let us now assume that a sentence S gives rise to a set of possible interpretations in any world, that $S(w)$ denotes this set $\{m_1, ..., m_n\}$, and that $[[m]]$ denotes the proposition in which m is true. In that case, exhaustivity comes down to the following:

$$[[exh]] = \lambda S \lambda P.\{w \in W| \; P(w) \in S(w) \wedge \neg \exists m \in S(w) : w \in [[m]] \wedge \inf(\lambda v[P(w) \subseteq P(v)]) > \inf(\lambda v[m \subseteq P(v)])\}$$

But what does this formula mean? In particular, how should we interpret question-predicate P in this case? Well, notice that for standard *wh*-questions we assume that P just denotes a property from worlds to a set of individuals: the extension is the set of all individuals that have property P in that world. For sentences, we can assume something similar. Suppose S is a sentence like *John killed the sheriff*. We might then assume, for instance, that P is a function from worlds to ways in which John killed the sheriff in those worlds. Let's assume that for any world w, $P(w)$ denotes a set. Suppose that in w, John killed the sheriff in a stereotypical way, i.e. by knife or pistol. In that case $P(w)$ denotes the singleton set consisting of the state description saying that John killed the sheriff in this stereotypical way, and $\lambda v[P(w) \subseteq P(v)]$ denotes the proposition corresponding with this state description.

Because $\inf(A) > \inf(B)$ if and only if $P(A) < P(B)$, we see that for these special cases our exhaustivity operator picks out the most likely, or stereotypical, interpretation of S. Compare this last formula with Blutner's (1998) formalization of the I-principle in terms of conditional informativity (assuming that $[[S]]$ denotes the set of worlds in which S is true under any interpretation):

$$I\text{-principle} = \lambda S.\{w \in [[m]]| \; m \in S(w) \wedge \neg \exists m' \in S(w) : \; w \in [[m']] \wedge \inf([[m]]/[[S]]) > \inf([[m']]/[[S]])\}$$

One can see that they differ at two points: (i) whereas our interpretation rule considers only alternative interpretations of predicate P, Blutner allows the alternative interpretations of a sentence to vary in much more unconstrained ways; (ii) whereas Blutner considers *conditional* informativity of the state descriptions after the semantic meaning of S is learned, we consider the informativity of the state descriptions themselves. If we assume that also Blutner allows only for variations with respect to a particular predicate, and if the probability ratios between the elements of S do not change after you learn that S is the case, i.e., if we make the

following assumption: $\forall m, m' \in S : P(m/S) > P(m'/S)$ iff $P(m) > P(m')$, our exhaustivity principle and Blutner's formalization of the I-principle come down to the same. But this suggests that we have come to the remarkable conclusion that both Q and I implicatures can, in principle, be accounted for by the same principle of exhaustive interpretation!

8 Bidirectional OT and Horn's division of labor

In the previous section we have reduced both the Q and the I principle to the principle that we interpret sentences exhaustively. We saw that this assumes that *speakers* are relevance optimizers. However, doesn't that mean that as a result we have to give up on Blutner's bidirectional OT? In particular, how could we now account for Horn's (1984) division of pragmatic labour, so elegantly explained in terms of Blutner's OT, and so important to explain why marked expressions typically get non-stereotypical interpretations?

The solution to this problem readily suggests itself: we can still make use of bidirectional OT, but we base the theory not on the Q (or S) and I (or H) principles, but rather on the principles of *relevance maximization* (the R principle) and *effort* minimalization (the E principle). We have seen that many Q (and some I) implicatures can be captured by our assumption of relevance maximization. The inference to stereotypical interpretation can be accounted for by the I principle, which should, I believe, be part of the principle to minimize effort. The I principle does not mention alternative expressions. To account for markedness phenomena, however, or Horn's division of pragmatic labor, the E principle should take alternative expressions into account as well.

Notice that when we explain interpretation as a balancing act between relevance and effort, our analysis seems very close to Sperber & Wilson's (1986) analysis of natural language in terms of their Theory of Relevance. However, there is an important distinction: whereas Sperber & Wilson seek to maximize relevance from the *hearer*'s point of view, we crucially assume that it is the *speaker* who wants to maximize her relevance. This conclusion, I take it, is very much in accordance with Zeevat's (2000) criticism of Blutner's original formulation of bidirectional OT. Blutner crucially assumed that the hearer wants to minimize his effort to understand what the speaker meant. Zeevat argues forcefully that this gives too much responsibility to the hearer: he just has to find out what the speaker meant. So it seems that just like Sperber & Wilson, also Blutner overrated the responsibilty of the hearer in the interpretation process: both maximization of relevance *and* minimization of effort are primarily important from the speaker's point of view. But if we minimize the role of the hearer in this way, it seems that the understanding of bidirectional OT as I appealed to in the introduction of this paper – as an interpretation game between speaker and hearer – is not as straightforward as it seemed. Indeed, I believe that we should think of bidirectional OT primarily as a theory that explains why certain linguistic conventions – in particular Horn's division of pragmatic labor and some principles of centering theory – typically emerge, and that these general conventions, in turn, explain why participants of a particular conversation say and interpret sentences in the way they do.[26] However, although bidirectional OT should be thought of primarily as a theory of language organization, these organizational principles

can only be explained in terms of economical language use.

References

Aloni, M. (2001). Pragmatics for propositional attitudes. In R. van Rooy & M. Stokhof (Eds.), *Proceedings of the Thirteenth A'dam Colloquium*. Amsterdam: ILLC.

Asher, N. & Lascarides, A. (1998). Bridging. Journal of Semantics, *15*, 83–113.

Atlas, J. & Levinson, S. (1981). It-clefts, informativiteness and logical form. In P. Cole (Ed.), *Radical Pragmatics*. New York: Academic Publishes.

Bar-Hillel, Y. & Carnap, R. (1953). Semantic information. In *Proceedings of the Symposium on Applications of Communication Theory*. London: Butterworth Scientific Publications.

Beaver, D. (to appear). Centering and the optimization of discourse. Linguistics and Philosophy.

Blutner, R. (1998). Lexical pragmatics. Journal of Semantics, *15*, 115–162.

Blutner, R. (2000). Some aspects of optimality in natural language interpretation. Journal of Semantics, *17*, 189–216.

Clark, H. (1987). Relevance to what? Behavioral and Brain Sciences, *10*, 714–714.

Dalrymple, M., Kanazawa, M., Kim, Y., Mchombo, S., & Peters, S. (1998). Reciprocal expressions and the concept of reciprocity. Linguistics and Philosophy, *21*, 159–210.

de Hoop, H. & de Swart, H. (2000). Optimality theoretic semantics. Linguistics and Philosophy, *24*, 1–32.

Dekker, P. & van Rooy, R. (2000). Bi-directional optimality theory: An application of game theory. Journal of Semantics, *17*, 217–242.

Ducrot, O. (1973). *La peuvre et le dire*. Paris: Mame.

Fauconnier, G. (1975). Polarity and the scale principle. In *Papers of the Eleventh Regional Meeting*, (pp. 188–199). Chicago: Chicago Linguistic Society.

Gärdenfors, P. (1988). *Knowledge in Flux, Modeling the Dynamics of Epistemic States*. Cambridge, Massachusetts: MIT Press.

Gazdar, G. & Good, D. (1982). On a notion of relevance. In N. Smith (Ed.), *Mutual Knowledge*, (pp. 88–100). New York: Academic Press.

Good, L. (1950). *Probability and the Weighing of Evidence*. London: Griffin.

Green, M. (1995). Quantity, volubility, and some varieties of discourse. Linguistics and Philosophy, *18*, 83–112.

Grice, P. (1957). Meaning. Philosophical Review, *66*, 377–388.

Grice, P. (1975). Logic and conversation. In P. Cole & Morgan (Eds.), *Syntax and Semantics 3: Speech Acts*. New York: Academic Press.

Groenendijk, J. & Stokhof, M. (1984). *Studies in the Semantics of Questions and the Pragmatics of Answers*. Ph.D. thesis, University of Amsterdam.

Grosz, B., Joshi, A., & Weinstein, S. (1983). Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, (pp. 44–49). Cambridge MA.

Grosz, B., Joshi, A., & Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. Computational Linguistics, *21*, 203–226.

Hasida, K., Nogao, K., & Miyata, T. (1995). A game-theoretic account of collaboration in communication. In *Proceedings of the First International Conference on Multi-Agent Systems*. San Fransisco.

Hirschberg (1985). *A theory of scalar implicature*. Ph.D. thesis, University of Pennsilvania.

Hobbs, J. (1979). Coherence and coreference. Cognitive Science, *3*, 67–90.

Horn, L. (1984). Towards a new taxonomy of pragmatic inference: Q-based and r-based implicature. In D. Schiffrin (Ed.), *Meaning, Form, and Use in Context: Linguistic Applications*, (pp. 11–44). Washington: Georgetown University Press.

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

Jäger, G. (2002). Some notes on the formal properties of bidirectional optimality theory. Journal of Logic, Language and Information, *11*, 427–451.

Kadmon, N. (1987). *On Unique and Non-Unique Reference and assymetric Quantification*. Ph.D. thesis, University of Masachusetts.

Kempson, R. (1986). Ambiguity and the semantics-pragmatics distinction. In C. Travis (Ed.), *Meaning and Interpretation*, (pp. 777–103). Oxford: Blackwell.

Krifka, M. (2002). Be vague and short! and why the interaction of pragmatic rules in the interpretation of measure terms as conceived by bidirectional optimality theory allows for length and precision. In *Festschrifft Vennemann*.

Levinson, S. (1987a). Implicature explicated? Behavioral and Brain Sciences, *10*, 722–723.

Levinson, S. (1987b). Pragmatics and the grammar of anaphora. Journal of Linguistics, *23*, 397–434.

Levinson, S. (2000). *Presumptive Meanings. The Theory of Generalized Conversational Implicatures*. Cambridge Massachusetts: MIT Press.

Lewis, D. (1979). Scorekeeping in a language game. Journal of Philosophical Logic, *8*, 339–359.

Merin, A. (1999). Information, relevance, and social decisionmaking. In M. d. R. L. Moss, J. Ginzburg (Ed.), *Logic, Language, and Computation, Vol. 2*, (pp. 179–221). Stanford: CSLI.

Parikh, P. (2000). Communication, meaning, and interpretation. Linguistics and Philosophy, *23*, 185–212.

Savage, L. (1954). *Foundations of Statistics*. New York: Wiley.

Schulz, K. (2001). *Relevanz und 'Quantity' Implikaturen*. Master's thesis, University of Stuttgart.

Shannon, C. (1948). The mathematical theory of communication. Bell System Technical Journal, *27*, 397–424 and 623–656.

Soames, S. (1982). How presuppositions are inherited: A solution to the projection problem. Linguistic Inquiry, *13*, 483–545.

Sperber, D. & Wilson, D. (1986/1995). *Relevance; Communication and Cognition*. Oxford: Blackwell.

Stalnaker, R. (1978). Assertion. In P. Cole (Ed.), *Syntax and Semantics, vol. 9: Pragmatics*, (pp. 315–332). New York: Academic Press.

van der Sandt, R. (1992). Presupposition projection as anaphora resolution. Journal of Semantics, *9*, 223–267.

van Kuppevelt, J. (1996). Inferring from topics: Scalar implicature as topic-dependent inferences. Linguistics and Philosphy, *19*, 555–598.

van Rooy, R. (1999). Questioning to resolve decision problems. In P. Dekker (Ed.), *Proceedings of the Twelfth Amsterdam Colloquium*. Amsterdam.

van Rooy, R. (2001). Conversational implicatures. In J. van Kuppevelt & R. Smith (Eds.), *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*. Aalborg.

van Rooy, R. (2002). Utility, informativity, and protocols. In G. Bonanno (Ed.), *Proceedings of LOFT 5: Logic and the Foundations of the Theory of Games and Decisions*. Torino.

van Rooy, R. (to appear). Signaling games select horn strategies. Linguistics and Philosophy.

van Rooy, R. & Schulz, K. (2003). Exhaustification. In H. Bunt (Ed.), *Proceedings of the Fifth International Workshop on Computational Semantic*. Tilburg.

Zeevat, H. (1994). Questions and exhaustivity in update semantics. In H. Bunt (Ed.), *Proceedings of the International Workshop on Computational Semantics*. Tilburg.

Zeevat, H. (1999). Explaining presupposition triggers. In P. Dekker (Ed.), *Proceedings of the 12th Amsterdam Colloquiums*. Amsterdam.

Zeevat, H. (2000). The asymmetry of optimality theoretic syntax and semantics. Journal of Semantics, *17*, 243–262.

Zipf, G. (1949). *Human behavior and the principle of least effort*. Cambridge: Addison-Wesley.

Notes

[1]According to optimality theory there exists also a generation function, G, that assigns to each form f a *set* of interpretations that it could possible mean. For ease of exposition I will ignore this function, but all form-meaning pair combinations that play a role in the definitions will obey this constraint: for all $\langle f, m\rangle$ mentioned, $m \in G(f)$.

[2]Dekker & van Rooy (2000) have shown that this notion of optimality can be thought of as a special case of the notion of optimality used in Game Theory: it corresponds with the standard solution concept of a *Nash equilibrium* (in updated games). Also Parikh's (2000) game-theoretical analysis of successful communication is formally very close to Blutner's bidirectional OT. For discussion, see van Rooy (ms).

[3]More in detail, $\inf(m/[[f]])$ is $-log_2P(m/[[f]])$, where P is a probability function, and the probability of C conditional on B, $P(C/B)$, is determined as $\frac{P(B\wedge C)}{P(B)}$. The logarithm with base 2 of n is simply the power to which 2 must be raised to get n. Thus, if $P(C/B) = 1/4$, then $-log_2P(C/B) = 2$, because $2^2 = 4$, and if $P(C/B) = 1/8$, then $-log_2P(C/B) = 3$, because $2^3 = 8$. Thus, in case $P(C/B)$ gets lower, the value of $\inf(C/B)$ gets higher.

[4]In fact, Blutner (1998) argues that this reduction is in line with Atlas & Levinson's (1981) and Horn's (1984) reduction of Gricean pragmatics to the two contrary Q and I principles. Katrin Schulz convinced me that Blutner was wrong here. I will come back to this.

[5]Beaver's analysis of centering in OT extends the empirical coverage of the theory considerably. I will limit myself to original centering, however, and Beaver's reformulation of it.

[6]For simplicity, I will just assume the descriptive adequacy of centering theory, although I am aware that since the original statement of centering theory many alternatives have been proposed.

[7]Ignoring the more specific gender/number constraints.

[8]For a discussion of some other problems, see Zeevat (2000) and van Rooy (to appear).

[9]This analysis of assertions can be extended to questions. See van Rooy (1999, 2002) for details.

[10]A collection Q of subsets of W is a partition of W iff (i) the partition covers W: $\bigcup Q = W$, and (ii) the elements of Q do not overlap: $\forall q, q' \in Q : q \cap q' = \emptyset$.

[11]In fact, we do not have to limit ourselves to partitions, but I will do so to simplify matters.

[12]Thus, for all $q \in Q$ it holds that $P(q/B) = \frac{1}{card(\{q \in Q |\ B \cap q \neq \emptyset\})}$.

[13]Note that by quantification over probability functions, our ordering relation '$>$' induced by entropy does not generate a total ordering anymore.

[14]Although argumentative value is defined rather differently from entropy value, $EV_Q(\cdot)$, observe that in case of binary issues (is h true or $\neg h$?), the two notions of *irrelevance* coincide.

[15]This definition is not exactly the same as the one used by Merin (1999); he in fact uses Good's (1950) function that measures the *weigh of evidence*, a function that is continuously monotone increasing with respect to $AV_h(\cdot)$.

[16]Compare this also with the strongest meaning hypothesis of Dalrymple et al. (1998).

[17]The reason is that, in the end, the presumption of optimal relevance is not stated in terms of optimization of extend condition 1. It is only demanded that this extend has to be 'sufficiently' high. No independent measure of what counts as being sufficient is given, however. If 'sufficiently high' means 'having a positive utility', almost the entire notion of relevance comes down to minimizing processing effort.

[18]According to one reviewer, this analysis justifies something weaker than Stalnaker was claiming.

[19]Proponents of S&W Relevance Theory won't find this very surprising: Sperber & Wilson (1986) themselves explain such phenomena by appealing to the notion of 'processing effort' which my notion of utility by itself doesn't capture.

[20]See Gärdenfors (1988) for an analysis of revision of probability functions

[21]Levinson's (2000) I principle is formulated as follows: 'Say as little as necessary; that is, produce the minimal linguistic information sufficient to achieve your communicational ends'. According to Levinson (2000) this principle means the following from the hearer's point of view: 'Amplify the informational content of the speaker's utterance, by finding the most *specific* interpretation up to what you judge to be the speaker's m-intended point, unless the speaker has broken the maxim of Minimization by using a marked or prolix expression.' This suggests taking the maximally informative interpretation, and indeed, he explicitly defines p to be *more specific than* q if (a) p is more informative than q; and (b) p is isomorphic with q. Strangely enough, however, the I principle is also supposed

to account for the inference to stereotypical interpretations, which by definition are not the most informative at all. It is unclear to me how that is supposed to follow on Levinson's reading of 'specificity'. In this section I will assume that the I principle simply demands selection of the most informative interpretation.

[22]Or perhaps just the notion of utility, because it seems reasonable to assume that the second condition of our notion of relevance is already captured by Blutner's notion of *effort* in bidirectional OT.

[23]Schulz (2001) proposed this alternative way to interpret Grice.

[24]The result of this table can also be captured by the following exhaustivity operator that takes a number and a predicate as arguments and results in a proposition:

$$Exh(t)(P) \quad = \quad \{w \in P(t) | \neg \exists t' \in P(w) : P(t') > P(t)\}$$

Note that this exhaustivity operator says that one should interpret the sentence as relevantly as possible. In fact, Zeevat (1994) proposed something like this exhaustivity operator, but with '$>$' replaced by '$\models$'. Thus, according to Zeevat one should interpret a sentence *as informative* as possible.

[25]Although problematic, neither Gazdar (1979) nor Soames (1982) actually make this wrong prediction. Gazdar does not make it due to his assumption that the scalar implicatures are not allowed to be inconsistent with the clausal implicatures, and Soames not by weakening the force of scalar implicatures.

[26]See my 'Signalling games select Horn strategies' (to appear) for more on this.