*A Note on Pragmatic Principles of Least Effort**

ROBYN CARSTON

Abstract

Larry Horn has long argued for the reduction of the Gricean maxims of conversation to two, one that turns on saving the hearer's processing effort (the Q-Principle), the other oriented to reducing the speaker's effort (the R- Principle). I question the status of the latter as a communicative principle and, although it is presented as part of a neo-*Gricean* account, I suggest that it marks a clear departure from Grice's concerns.

1 A neo-Gricean framework

In his latest work, Larry Horn (2005) reiterates the position he has maintained since his important and influential 1984 paper, according to which, apart from the Quality (truthfulness) maxims, which he considers essential and unreducible (Horn 1984: 12; 2004: 13), the Gricean maxims should be reduced to two general principles. These are the Q-Principle and the R-Principle, the first of which is oriented to the interests of the hearer and the second to the interests of the speaker.

He invokes a number of forerunners to this position, including in particular Zipf (1949) and Martinet (1962). In developing an ecological account of human behaviour quite generally, Zipf emphasised the fundamental role played by a *Principle of Least Effort*: 'the primary principle that governs our entire individual and collective behaviour of all sorts, including the behaviour of our language …' (Zipf 1949: vii). In the realm of linguistic behaviour, he distinguished the opposing pressures exerted by the speaker's and the hearer's economies of effort, the one oriented toward minimal linguistic articulation, the other toward maximal explicitness. Martinet (1962) described a primary mechanism of language change as coming from the interaction of two factors: 'first, the requirements of communication, the need for the speaker to convey his message, and second, the

* This short piece is part of a longer paper entitled 'Relevance theory, Grice and the neo-Griceans', to appear in *Intercultural Pragmatics*, in which I respond to Horn (2005) on the different goals and orientations of the three approaches. I'm grateful to Tim Wharton, Deirdre Wilson and Vladimir Žegarac for interesting discussion of the issues.

principle of least effort, which makes him restrict his output of energy, both mental and physical, to the minimum compatible with achieving his ends.' (Martinet 1962: 139).

According to Horn: 'These same two antinomic forces – and the interaction between them – are largely responsible for generating Grice's conversational maxims and the schema for pragmatic inference derived therefrom' (Horn 1984: 11-12), and he sets up his Q- and R-Principles so as to more clearly reflect these 'two functional economies of Zipf and Martinet.' Even before looking into the details, this might strike one as a surprising endeavour, since, as the Martinet quote most clearly indicates, what is needed to make communication (conversation) work is pitched against the speaker's natural inclination to conserve his energy. *Prima facie*, then, we would not expect communicative principles (or conversational maxims) to somehow include the very force they are designed to overcome. I am not disputing the role of speakers' articulatory (and other) economies in language change, for which there is considerable evidence[1], but it is a very different matter to draw this into the precepts that govern effective communicative conduct.

Here are the two principles as they are standardly presented (Horn 1984, 1989, 2004):

The Q-Principle (Hearer-based):
MAKE YOUR CONTRIBUTION SUFFICIENT
SAY AS MUCH AS YOU CAN (modulo R)
Lower-bounding principle, inducing upper-bounding implicata
[Grice's first Quantity maxim and the first two Manner maxims]

The R-Principle (Speaker-based):
MAKE YOUR CONTRIBUTION NECESSARY
SAY NO MORE THAN YOU MUST (modulo Q)
Upper-bounding principle, inducing lower-bounding implicata
[Grice's second Quantity maxim, Relation maxim and the second two Manner maxims]

Note the neat symmetry: hearer-oriented vs. speaker-oriented, lower-bounding vs. upper-bounding, Gricean Quantity 1 vs. Gricean Quantity 2, and the split between the four Manner maxims. Although the one is hearer-based and the other speaker-based, they are both presented (like the Gricean maxims) as injunctions to the speaker, and each presumably has its hearer's corollary since 'The speaker and hearer are aware of their own and each other's desiderata [that is, the economies of

[1] The role of Horn's R-Principle (speaker's economy) in semantic change is explored in some detail in Traugott (2004).

both parties], and this awareness generates a variety of effects …' (Horn 2005: 19). I assume that the two formulations given for each principle are intended to have roughly the same import and that, therefore, the difference between 'your contribution', in the one, and what you 'say', in the other, is not to be taken as significant. When he's being more summary, Horn uses the second formulations (Horn 2005: 196), and those are the versions that matter in discussions of the implicatures arising from these principles (applied to what was said and what was not said).

The Q-principle has received a vast amount of attention (due mainly to the veritable industry that has developed around the phenomenon of scalar implicature). Here I'll focus instead on the R-Principle.

2 A speaker's economy of effort

The R-Principle is taken to correspond with Zipf's 'speaker's economy', that is, a force pressing in the direction of a single word or sound that would express all possible meanings, thereby sparing the speaker the mental effort that is involved in selecting the appropriate linguistic forms for the meanings she wants to convey (Zipf 1949: 20) and the physical effort of articulating them. At the same time, it is claimed to mop up a number of the Gricean maxims, including the relevance maxim. I find this quite puzzling. What need is there for a conversational maxim or principle that enjoins a speaker to minimise her effort? First, she's going to do it anyway (modulo her communicative goal, of course), as Zipf and Martinet tell us. Second, what the Gricean maxims are for (and so also the neo-Gricean maxims/principles derived from them) is to account for the derivation of conversational implicatures. Conversational implicatures are assumptions that a hearer has to attribute to a speaker in order to preserve the presumption that the speaker is observing the maxims. But while a hearer has reason to be concerned that a speaker is being relevant, informative, truthful, and orderly, why would he care to preserve an assumption that a speaker is minimising her effort? The opposition of forces that interests Horn seems to be between a general *non*-communicative principle of effort minimisation and whatever principles/maxims are responsible for successful communication.

Let's consider the various Gricean maxims which the R-Principle is claimed to subsume. First, the second Quantity maxim: 'Do not make your contribution more informative than is required'. Grice was unsure about this maxim, making a number of points against it, including the likelihood that it would be replaced by a properly developed maxim of Relation (relevance). It is clear from his discussion that he did not think of either Quantity-2 or the Relation maxim as being concerned with speaker effort. Rather, they were both intended to avoid 'confusing [the

hearer] in that it [overinformativeness] is liable to raise side issues; and there may be an indirect effect, in that the hearers may be misled as a result of thinking that there is some particular *point* in the provision of the excess of information' Grice (1975/89: 26-27, 34). In other words, these maxims are oriented towards sparing the *hearer* gratuitous processing effort. Now, what about the two Manner maxims that are claimed to fall under the R-Principle: 'Be brief' and 'Be orderly'. These are surely not concerned with the speaker's effort either (be brief so as to save yourself the physical energy of exercising your articulatory organs? be orderly in the way you present information because that will save you mental effort?). Recall that these submaxims fall under a general supermaxim of Manner: 'Be perspicuous'. A speaker is expected to be perspicuous for the sake of her addressee's comprehension, not to cater to any need of her own. Like all the other Gricean maxims, the manner maxims are aimed at achieving the speaker's intention of affecting the hearer in certain ways (getting him to entertain certain thoughts or to perform certain actions) (Grice 1975/89: 28).

The Q- and R-principles are claimed to have complementary bounding properties. The talk of bounds is reasonably clear in the Q cases: what is encoded or said sets a lower bound, e.g. 'at least three/some/possible' and the Q-based implicature provides an upper bound, e.g. 'at most three/not all/not certain':

(1) a. Utterance: Max ate some of the cakes.
b. Implicature: Max didn't eat all of the cakes.

It's considerably less clear how the alleged reverse pattern of bounds pans out in the R cases. What does the R Principle place an upper bound on – is it speaker effort, or information given, or (somehow) both? Consider some examples standardly cited as giving rise to R-based implicatures:

(2) a. Utterance: Hannah insulted Joe and he resigned.
b. Implicature: Hannah's insulting Joe caused him to resign.

(3) a. Utterance: I broke a finger yesterday.
b. Implicature: The finger is mine.

(4) a. Utterance: Can you reach the salt?
b. Implicature: I request you to pass the salt to me.

In what sense are the implicatures here lower-bounding? The idea seems to be that, on the basis of the speaker having said that *p*, there is an R-inference to 'more than *p*' and the hearer has to go on to figure out what this more is. In the case of 'and'-conjunctions, such as (2), it is a stronger relation between the conjuncts; in (3), it is

a narrowing to a specific domain of fingers; and in cases of indirect speech acts, such as (4), it is simply a different speech act from the (irrelevant) one directly expressed. But one could say much the same of the Q cases: something is explicitly said and what is communicated is stronger (more informative/relevant). In both kinds of case, there is a strengthening of communicated content: from 'at least some' to 'just some', from 'a finger' to 'my finger', etc. I find little support for the claim (Horn 1984: 14) that the effects of the two principles are mirror images of each other. (For fuller discussion of this point, see Carston 1998: section 3).

Another interesting feature of the examples governed by the R-principle is that, in each case, it seems that the utterance that apparently conforms to the principle of least speaker effort is also the low cost option from the hearer's point of view. He has immediate mental access to stereotypical scripts about the temporal sequence and consequence relations between events (for the 'and'-conjunctions) and to the most frequently encountered finger-breaking scenario. Can we find cases where the speaker's effort economy and the ease of processing for the hearer might come apart? Suppose you ask me if I know what has become of our old college friends, Ann and Bob, whom you lost touch with a long time ago, and suppose the first thought to occur to me is that they recently had a baby. I next recall that they got married a while ago. Although this is the order of my thoughts, it is very likely (I think) that I will choose to give you this information via an utterance of 'They got married and have had a baby', which reflects the order in which the events occurred, because this is the order from which you will most easily derive the appropriate (temporal, etc) implications. If this is right, I am tailoring my utterance in the interests of my hearer's effort expenditure rather than my own.

Consider the following pairs of possible utterances:

(5) a. Please pass the salt and pepper.
b. Please pass the condiments.

(6) a. Do you have any brothers or sisters?
b. Do you have any siblings?

(7) a. There's a lot of interesting animal life in this region.
b. There's a lot of interesting fauna in this region.

The second member of each of these pairs should cost the speaker less articulatory effort than the first while, most likely, the first in each case is the more frequent, the less marked, and hence the more accessible (other things being equal) to the hearer. Note that it's not obvious that the members of these pairs are each pragmatically associated with complementary denotations, as is the case for the examples usually cited by Horn to illustrate the 'division of pragmatic labor', such

as the 'pink'/'pale red', 'kill'/'cause to die', 'mother'/'father's wife' pairs (Horn 1984: 22-31; 2004: 16-17). The pairs in (5)-(7) seem to be synonymous. So which will a speaker choose? Other things being equal (as ever), the intuition is that she is likely to choose the more frequently used phrase over the briefer, less articulatorily demanding word. Furthermore, if she does choose the lexical form rather than the phrase, her utterance will tend to have a layer of effects (of a largely social sort: distance, formality) not present when the longer, more syntactically complex expression is used. So we may have a split here between speaker's and hearer's effort. I would suggest, though, that this sort of case really falls together with those that Horn presents ('kill'/'cause to die', etc), which have the apparently opposite properties in that it is the 'more complex, less lexicalised expression' that gives rise to some sort of special effect (Horn 2004: 17). However, unification of the two sets of cases is only obviously possible if the effort at issue is that of the hearer rather than that of the speaker. Then, as predicted by the relevance-theoretic (RT) account, in all instances, the extra processing effort required of the hearer is offset by particular cognitive effects which the less effort demanding variant would not have had (Sperber and Wilson 1986/95; Wilson and Sperber 2004).

As Horn often points out, Zipf wrote not only of speaker's economy, but also of the 'auditor's economy' which, he claimed, tends towards the establishment of as many different linguistic expressions as there are messages to communicate (Zipf 1949: 21; Horn 1984: 11), i.e. full explicitness. I don't think Zipf is correct about this, at least if the implication is that for any meaning *m*, the least effortful, hence most economical, utterance from the hearer's point of view would be one that fully encoded *m*. Even if this were possible, there is just no evidence that hearers prefer a given message or meaning to be as fully encoded as possible. A good speaker, as opposed to one who is found to be tedious, pedantic, even misleading, is one who does not over-encode, judging more or less correctly what a hearer can easily infer in a particular context: when a particular referent is highly salient, a pronoun is preferable to a name or a description; when a topic is established, a phrasal utterance (e.g. 'On the table') may be preferable to a sentential one (e.g. 'Mary put the book on the table'), etc. On the RT account, it is considerations such as these that inform a speaker's choice of form to utter.

Finally and very briefly, does the issue of speaker effort have any role to play within RT pragmatics? Horn and others (e.g. Saul 2002) often say that RT is an entirely hearer-based account, which 'does not address the question of how and why the speaker, given what she wants to convey, utters what she utters' (Horn 2005: 194). In fact, considerations of speaker effort are accommodated by the account, although they do not play anything like the central role that Horn favours.

Recall that the key theoretical concept in RT, on which the mechanics of communication and comprehension are based, is 'optimal relevance'. An utterance is optimally relevant to a hearer just in case: (a) it is relevant enough to be worth

the hearer's processing effort, and (b) it is the most relevant one compatible with the speaker's abilities and preferences (or goals) (Wilson and Sperber 2004: 612). Utterances quite generally come with a presumption that they have this property (this is the *Communicative Principle of Relevance*); for any given utterance, the addressee searches for an interpretation that satisfies specific expectations of relevance projected from the general presumption as applied to the specific case; and, for the most part, speakers aim at producing utterances which are optimally relevant to their addressees.

Speakers aren't always able to produce the utterance which would be the *most* relevant one at that time to their addressee, due to shortcomings in their knowledge and/or their facility with language - hence the proviso regarding the speaker's abilities given in clause (b) of the definition above. But what is of more interest in the present discussion about a speaker's economy of effort is the other proviso in the clause, that concerning the speaker's preferences. No doubt, speakers do often prefer to save themselves mental and physical effort, but this preference will usually be in interplay with other goals. A speaker may prefer not to risk offending her hearer and so speak in a roundabout way on a particular matter, she may be unwilling to disclose some highly relevant but perhaps incriminating information, she may want to impress with her learned vocabulary so not employ the most economical form of words, she may be talking with great energy in order to distract the hearer's attention from something, and so on. All of these and many other possible preferences (which are distinct from the fundamental communicative goal) may affect the degree of effort a speaker is willing to expend in encoding her thoughts and articulating linguistic forms - as little as possible in some cases, but considerably more in others. So for the hearer, following the relevance-based comprehension procedure, the issue of how much effort the speaker is willing to make is but one of numerous factors that may affect his specific expectation of relevance on a given occasion of utterance interpretation.

References

Carston, R. (1998). Informativeness, relevance and scalar implicature. In: Carston, R. and Uchida, S. (eds.), *Relevance Theory: Applications and Implications*, 179-236. Amsterdam: John Benjamins.

Grice, H. P. (1975/89). Logic and conversation. In: Cole, P. and Morgan, J. L. (eds.), *Syntax and Semantics 3: Speech Acts*, 41-58. New York: Academic Press. Reprinted in Grice, H.P. 1989. *Studies in the Way of Words*, 22-40. Cambridge: Harvard University Press.

Horn, L. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In: Schiffrin, D. (ed.), *Meaning, Form and Use in Context (GURT '84)*, 11-42. Washington: Georgetown University Press.

Horn, L. (1989). *A Natural History of Negation*. Chicago: University of Chicago Press.

Horn, L. (2004). Implicature. In: Horn, L. and Ward, G. (eds.), 3-28.

Horn, L. (2005). Current issues in neo-Gricean pragmatics. *Intercultural Pragmatics* 2 (2): 191-204.

Horn, L. and Ward, G. (eds.) (2004). *Handbook of Pragmatics*. Oxford: Blackwell.

Martinet, A. (1962). *A Functional View of Language*. Oxford: Clarendon Press.

Saul, J. (2002). What is said and psychological reality: Grice's project and relevance theorists' criticisms. *Linguistics and Philosophy* 25: 347-372.

Sperber, D. and Wilson, D. (1986/95). *Relevance: Communication and Cognition*. Oxford: Blackwell. Second edition 1995.

Traugott, E. (2004). Historical pragmatics. In: Horn, L. and Ward, G. (eds.), 538-561.

Wilson, D. and Sperber, D. (2004). Relevance theory. In: Horn, L. and Ward, G. (eds.), 607-632.

Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley Press.