

# Grice's Maxims: "Do the Right Thing"

Robert E. Frederking

Center for Machine Translation  
Carnegie Mellon University

ref@cs.cmu.edu

## Abstract

Grice's maxims are hopelessly vague, and in fact harmful, because they form a misleading taxonomy. While his cooperative principle may be useful at a high level of theoretical analysis, it too is vague, and should not be directly implemented in computational natural language systems. Answers are suggested to a number of this symposium's topics based on this position. Examples are presented to show that the maxims are too vague and too general, and that they are not really used by computational systems that claim to be based on them. The historical origins of the maxims in Kant's philosophy are revealed. A comparison is made with Relevance Theory, which seems to provide a better approach to the same phenomena. I conclude by suggesting that it may be too early in the history of computational linguistics to expect to find such broad principles.

## Introduction

I will argue in this position paper that Grice's maxims are hopelessly vague, and that while his cooperative principle may be useful at a high level of theoretical analysis, it should not be directly implemented in computational natural language systems.

From this point of view, there are clear answers to several of the questions this symposium will address:

- **Is the notion of conversational implicature still useful? What role if any do Grice's maxims and Cooperative Principle still play in computational and formal approaches?** The notion of conversational implicature, and the Cooperative Principle, have been useful and important to some researchers in thinking about how language works in real use. But however useful they are for guiding a researcher's thinking, they are not useful as an actual part of an implementation. The maxims, on the other hand, play no useful role whatsoever in any computational or formal approaches, even at a theoretical level. They are in fact harmful, because they form a misleading taxonomy.

- **But is Relevance a well-defined notion?** No. Like the other maxims, Grice's Relevance is a broad, general statement that is clearly true at some level, but is far too vague to be used directly in computational systems. I believe it is currently an open question whether some other approach to relevance (such as Relevance Theory, discussed below) may be amenable to precise definition.
- **What distinguishes conversational implicatures from other defeasible inferences in discourse (e.g., default inferences in text understanding)?** I would claim that the only thing that distinguishes conversational implicatures as a class is the fact that they can be seen as examples of Grice's principles. In other words, the categories he uses have no predictive or explanatory power. This is *not* to say that certain subclasses (such as scalar implicatures) do not have useful distinguishing features; only that the Gricean level of description is misleading.
- **Most models have focused on single classes of conversational implicature. What problems would arise in integrating them?** It is pointless to consider integrating different classes of conversational implicature, based on Grice's taxonomy. His taxonomy is not of a sufficiently concrete nature to be fruitfully applied to real implementations. There *are* significant issues in integrating different types of inference mechanisms in conversation, but Grice's categories are irrelevant to these issues.

In short, I claim that Grice's Maxims are similar to the maxim "Do the Right Thing," which any correctly working natural language system can be said to implement.

## Grice's maxims considered harmful

### A misleading taxonomy

Several researchers have tried to implement Grice's maxims in some fashion (for example, [Gazdar 79, Hirschberg 85]). This is not possible to do directly, due to the vagueness of the maxims, so they typically have implemented something more reasonable, and then claimed it was "Gricean". More often, researchers have tried to use Grice's maxims to describe particular systems or phenomena (for example, [Dale and Reiter 95, Joshi et al. 84, Passonneau 95]). Because the maxims have the form of a taxonomy, they lead researchers to think that the maxims taxonomize the space of conversational implicatures in some useful fashion. But using the maxims even in this way is counter-productive, because they are much too vague, and often overlap when applied to actual examples of conversational implicature. They tend to lead to confusion more than enlightenment.

For example, it is pointless to discuss (as Levinson does [Levinson 83, p. 109]) whether a particular phenomenon such as irony is based on the flouting of Relevance or Quality. Irony is a phenomenon that fits quite comfortably into both notions. It flouts both at once, and perhaps Manner, too. The desire to decide which maxim irony flouts is based on the false impression that there is some kind of significant difference between implicatures that fit into one category and those that fit in the other.

The maxims not only divide discourse phenomena up badly; they also group them together badly. Scalar and clausal quantity implicatures [Gazdar 79, Hirschberg 85] are a good example. These have both been described as subtypes of Quantity implicatures. But there does not seem to me to be any good reason to believe that these are two subclasses of the same phenomenon.

Clausal implicature typically occurs when an embedded proposition is neither affirmed nor denied by the full utterance. So the utterance of "If John sees me then he will tell Margaret" implicates that the speaker does not know whether John will see him. The standard explanation of this is that, based on the Cooperative Principle, if the speaker knew whether the first clause were true or false, he should have said so.

Scalar implicature, on the other hand, is based on the existence of sets of terms that have some salient partial ordering in degree of informativeness. So the utterance of "Paul ate some of the eggs" implicates that the speaker does not know that Paul ate all of the eggs. Again, the standard explanation of this relies on Quantity, that the speaker should have said so if he knew.

Clearly, these explanations are similar in character; unfortunately, as we will see below, speakers often provide more or less information than is necessary, so the generalization made by the maxim is not valid.

The other way these phenomena could be related is if the detailed, specific formal mechanisms for handling the two phenomena are similar. In fact, the formal mechanisms proposed to handle them differ. Hirschberg describes them as different phenomena before defining her mechanism for scalar implicature, and Gazdar states [Gazdar 79, p. 59] that his mechanism for handling scalar implicatures does not generate clausal implicatures, justifying his development of a separate formal mechanism for clausal implicatures. At the very least, no rigorous connection to the maxim is established, and it seems clear to me that there really isn't one. So, while each phenomenon appears to be well-defined in its own right, there does not seem to be any clear similarity in the way they are actually processed.

There is a subtle but important clue to the genesis of such an unhelpful taxonomy in Grice's original article [Grice 75]. Grice says that his categories are "echoing Kant" (p. 45). This clearly refers to Kant's theory of categories, which classified declarative statements along four dimensions: *Quantity*, *Quality*, *Relation*, and *Modality*. This makes it clear why Grice's taxonomy does not fit the discourse phenomena it was supposed to describe; it has been borrowed as a whole from a pre-linguistic, philosophical classification of statements! Taxonomies from one domain simply cannot be transferred wholesale to another and retain any usefulness. It is interesting that this attribution has never been quoted, to my knowledge.

### Trying to make use of a misleading taxonomy

Despite their vagueness, the maxims are clearly true in some sense. This, coupled with their vagueness, has allowed numerous researchers to read into them all sorts of specific true interpretations, rather than treating them as *maxims*, as Grice's name for them suggests<sup>1</sup>. As in the examples above, when we examine what actually exists in specific systems that are claimed to fulfill one or more of the maxims, we find much more specific mechanisms that apply to much more specific phenomena, and only bear a very tenuous connection to the maxims. The clarity of these individual phenomena and rules, despite any remaining controversies, is in sharp contrast to the haze of confusion surrounding the maxims. Given this, and

---

<sup>1</sup>Although Grice apparently did intend for them to be applied rigorously.

the fact that researchers often point out major problems with the maxims, it is difficult to understand the widespread, seemingly willful refusal to realize that the maxims simply are not correct.

As one example of being “soft on Grice”, Hirschberg [Hirschberg 85] redefines Quality very narrowly, and indicates that Quantity, Manner, and Relevance cannot really be defined precisely. Yet she then goes on to write logical formulae containing the maxims, as if they were rigorously definable. The basic problem is vagueness; Grice’s maxims are loaded with terms that are ill-defined, such as “as informative as required”. If one wants to be kind to Grice, this allows a huge amount of leeway for reinterpretation.

Another example is Levinson’s discussion [Levinson 83] of assymetric “and”. The fact that “and” can be used to mean “and then”, and that this is not a lexical ambiguity, seems to me to have been clearly established at this point. But this fact hardly makes the “Be orderly” submaxim of Manner a generally useful computational rule. There are in fact contexts in which it is quite acceptable to describe events out of order. If one is telling a story and says “John says he likes Mary, and Phil walks out the door,” there is a clear implication of sequentiality. However, if one says instead “John says he likes Mary, and Phil says he likes Mary,” there is no implication of sequentiality. The sequentiality seems to be provided by complex (and currently not well understood) phenomena involving real world knowledge and sequences of tense and aspect, not by some high-level principle of orderliness. What is needed, here and in general, is a careful investigation of specific phenomena, not a general pronouncement of a principle that is sometimes true.

As an even better example of the tendency to apologize for Grice, there is the Dale and Reiter discussion of the generation of referring expressions [Dale and Reiter 95, section 2.4]. They attribute the tendency of speakers to generate the shortest unique referring expression to the maxim of Quantity. They give a typical example of obeying Quantity: a speaker who says “Look at the pit bull” rather than “Look at the dog” implicates that the type of dog is important, perhaps because it is more dangerous. Unfortunately, as it turns out, speakers do *not* generate minimal referring expressions. Here is a similar example (similar at the level of the maxims) that does *not* obey Quantity: Suppose there is a room containing only alligators. English speakers would normally refer to the largest one as “the largest alligator” rather than “the largest animal” (or even better, “the largest thing”). They do this simply because “alligator” is the unmarked level of description that English normally uses. Why does this not gen-

erate a conversational implicature? According to the maxim of Quantity, a speaker generating clearly superfluous information should cause the hearer to produce implicatures, as in the previous example. This phenomenon clearly violates any direct interpretation of Quantity.

The truly surprising thing is that Dale and Reiter discuss this total failure of Quantity in the paper, and yet do not describe it as a failure. They refer to this as one type of “lexical preference” [Dale and Reiter 95, quoting [Reiter 91]]. In addition, in section 3.2, they mention that speakers generate other redundant information as well. According to the maxims, this should also cause implicatures; but it does not (except in those cases where it does...) Furthermore, the algorithm described in this paper does the *conventional* thing, that is, what people seem to do, according to psychological experiments. This seems to me to be exactly the right approach to designing noun phrase generators at the current time, but it is entirely non-Gricean, despite the title, and despite their claims that they are using “more precise” versions of his maxims.

It seems to me that Dale and Reiter should have gone on to conclude that the maxim of Quantity is simply wrong, or at least removed “Gricean Maxims” from the title. Later, they suggest that the Gricean maxims may be “approximations” to a general principle of “if a speaker utters an unexpected utterance, the hearer may try to infer a reason for the speaker’s failure to use the expected utterance” (section 3.3). But this *cannot* be a case of approximation to, or “simple interpretations” of, Grice’s maxim of Quantity, since their principle does not refer to *Quantity* at all. What they are actually suggesting are implicatures based on violating *conventions*, as opposed to implicatures based on any high-level principle. Which, again, I believe is probably correct, except that the maxims should be left out of it.

## Previous arguments along these lines

Kiefer [Kiefer 79] makes similar arguments against Grice’s maxims, as well as additional arguments that seem well-founded to me. Unfortunately, the fact that he combined this general attack with specific criticisms of Gazdar’s work [Gazdar 79] allowed Gazdar’s reply [Gazdar 80] to focus narrowly on the technical linguistic issues of his own work, rather than the broad criticism of Grice’s original theory. The same situation occurs with Cohen’s attack [Cohen 71] and Gazdar’s reply to it [Gazdar 79]. Disturbingly, Levinson’s book [Levinson 83, p. 122] gives the impression that these replies have successfully responded to the general attacks on Grice, which is not the case at all. Additional

clear criticisms of the Gricean approach can be found in works by Sadock [Sadock 78] and by Wilson and Sperber [Wilson and Sperber 81, Sperber and Wilson 86, Wilson and Sperber 88].

## Trying to clear things up

### The Cooperative Principle as an implicit constraint

As indicated above, Grice's Cooperative Principle explains many implicatures that occur in discourse, but it is far too broad. That is, one can view lots of true facts as subsumed by it, but one cannot start with the Cooperative Principle or the maxims and produce useful inferences, because they also predict things that do not occur, such as the "alligator" example above. Note that I am criticizing the Gricean-level maxims here, and not the much more specific systems such as clausal or scalar implicature.

A helpful way to think about the Cooperative Principle is as an *implicit* constraint, as opposed to an *explicit* one. By this I mean that there are true constraints that are not explicitly implemented by the systems they describe, but which the systems implicitly obey. For example, the moon obeys Kepler's Laws of Planetary Motion, but no one believes that the moon thinks about Kepler's Laws and decides how to move based on them<sup>2</sup>. For computational linguistics, explicit constraints are constraints that can be directly implemented by programs that correctly transduce between language and the information conveyed by language.

There are implications whenever people do something unusual. But it appears to me, as hinted at above, that there are specific classes of *conventionalized* implicatures that people use. Note that this is different from *conventional implicature* in that these classes of implicatures are not attached to specific lexical items, but rather occur in certain situations. As in the "pit bull" example above, one class might be when a speaker fails to use the normal lexical preference for an object. Other possible classes might be clausal implicatures and the various scalar implicatures.

Many and perhaps all of the specific conventionalized implicatures that people use do obey Grice's broad, implicit constraint. But it is pointless to put his constraint explicitly into a computational system if it also predicts implicatures where they do not occur. In the terminology of constraints, unlike Kepler's Laws, the Cooperative Principle is a *loose* constraint, which is exactly saying that it permits some things that do not happen.

---

<sup>2</sup>Some people think that the moon really does compute the law of gravity, which is why I refer to Kepler's Laws here.

If this view is correct, then part of learning to use a particular language is learning specific rules about when to make an implicature and when not to, or possibly learning to recognize categories of expected behavior, and making inferences when expectations are violated. One can view the system described by Green and Carberry [Green and Carberry 94] as a system of this type, where their precompiled discourse operators would be the end product of the learning of conventionalized implicature usage.

This raises an intriguing possibility: that the Cooperative Principle could be used as part of a language learning system, to constrain the space of possible implicatures learned. In order to be successful, inductive learning requires as many constraints and biases as possible. The Cooperative Principle, or possibly Relevance Theory (see below), could be used as one such constraint. The learning system's hypotheses would be constrained by the assumption that the other speaker's behavior was cooperative (or relevant).

### Further taxonomic considerations

I have described above the internal taxonomic problems with Grice's scheme, due to its maxims being vague and overlapping. In addition, there is an external taxonomic problem, in that it does not seem to be possible to distinguish conversational implicature from other non-logical, non-conventional, constraint-based inferences. Based on Grice's derivation of his tests [Grice 75], they in fact ought to apply to any such constraints. That is, he derives his tests from the fact that conversational implicatures are not entailments, that they are not based on the surface forms of utterances, that they are not part of conventional meaning, and that they are context dependent. If this is true, then they fundamentally cannot distinguish conversational implicature from other defeasible inferences. This means that Gricean implicatures are only distinguishable by the fact that they can be seen to be derived from cooperativity, which is not very interesting. Among others, Sadock [Sadock 78] has previously made similar points.

Even worse, Wilson and Sperber [Wilson and Sperber 81] make a fairly compelling argument that disambiguation and reference resolution can also be seen as relying on the Cooperative Principle. To use their example, if someone listening to a student playing violin says "John plays well," the normal interpretation is that he plays the violin well. Suppose someone says instead "John plays well – but not the violin." It is difficult to see how the first instance differs from "obeying the maxims" with regard to disambiguation, or how the second differs from the usual types of "flouting"

and cancellation. Thus the Cooperative Principle appears to be so broad that it even covers phenomena that have nothing to do with what is normally understood as implicature. Finally, even Levinson [Levinson 83, p. 132] says that speakers “may well” make use of inferences based on any constraints.

Given all this, what is the correct way to build a taxonomy? For a taxonomy to be useful for our computational purposes, it should be based on functional classes that correspond to explicitly used information processing constraints and mechanisms. These constraints must have *operational definitions*: definitions based on simple primitives that can be implemented in hardware. As an example of at least a better attempt, Wilson and Sperber [Wilson and Sperber 81] divide pragmatic implicatures into two subclasses, direct and indirect, based on whether new assumptions are needed to interpret an utterance as relating to the previous conversation. This is a much clearer approach to the “flouting” versus “non-flouting” distinction made by Grice.

### **The Cooperative Principle versus Relevance Theory**

I have described above how conversational implicature is difficult to distinguish from other inferences, and how Grice’s maxims seem to overlap in confusing ways. That said, it does seem that the Cooperative Principle is real in some way beyond a possible role in learning, and that inferences that seem to be derived from it do occur in language.

Sperber and Wilson have produced a promising alternative approach to this whole area in their work on Relevance Theory (RT) [Sperber and Wilson 86]. They start from the position [Wilson and Sperber 81] that Relevance does not follow from the Cooperative Principle, or any other sociological principle. It just arises from the nature of communication: a speaker demands resources from a hearer, creating an implication that what the speaker is saying is worthwhile for the hearer to attend to. Relevance results from having a large enough effect on the hearer’s cognitive environment with a small enough processing effort.

The call for participation to this workshop mentioned the possibility that Relevance was the key Gricean maxim. Remarkably enough, if one re-reads Grice’s Cooperative Principle in this context, it seems to be essentially describing Relevance, in both Grice’s and Sperber and Wilson’s senses:

Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged.

(Although Sperber and Wilson argue that their notion of relevance differs in not implying any agreement on a common purpose, or any knowledge of accepted norms [Sperber and Wilson 86, p. 161-163].)

RT can thus be viewed as a claim that a better version of Relevance is indeed the only maxim. Whether or not it stands up as a whole, RT in any event seem to have a much clearer definition of Relevance, have a consistent, clearly worked-out theory, and have avoided the sorts of problems caused by Grice’s maxims. The main drawback to RT (at least in its 1986 form) is that the crucial concepts of cognitive effect of utterances and processing effort in understanding utterances both belong to an unspecified detailed cognitive theory. This dependency on cognitive modelling is unavoidable, but until some experiments are done combining RT and a suitable computational cognitive model, it is hard to judge the validity of the theory. In their book, Sperber and Wilson demonstrate in a number of places a certain amount of naivete regarding computation, so I suspect computational implementations will have to come from other researchers. The only attempt at a computational implementation of Relevance Theory that has come to my attention as of this writing is [Poznanski 92]<sup>3</sup>, which I have not yet obtained a copy of.

### **Conclusion: let’s wait a while**

I have argued that while much good work is self-described as “Gricean”, it bears only a loose connection to his maxims. This is no accident, since Grice’s maxims taxonomize conversational implicatures in unfruitful ways, and lead to confusion if one tries to take them seriously. As I have indicated, mine is by no means the first criticism of the Gricean approach, but previous attacks have apparently been disregarded, without being refuted.

If Grice’s theory is unusable, what should take its place? While I believe that the work of Sperber and Wilson could form a much more promising basis for generalization than Grice’s maxims, I suspect that it may simply be too early in the history of computational linguistics for broad, deep theories to be formulated.

If we look at the history of everyone’s favorite science (physics), we see an interesting pattern. Kepler formulated his Laws of Planetary Motion more than 50 years before Newton developed the universal theory of gravity. Kepler did not understand gravity; he believed in exotic Pythagorean theories that have since been discredited. Similarly, when Michelson and Morley experimentally demonstrated that the speed of light was

---

<sup>3</sup>Thanks to Robyn Carston for the reference.

constant in all frames of reference, they had no explanation. They believed that light travelled through the ether. Einstein came along almost 20 years later with the special theory of relativity to explain what was happening. In both cases (and many others), correct, detailed mathematical descriptions of specific phenomena preceded the formation of correct general theories.

Similarly, I believe we should carefully and concretely describe, and computationally solve, many specific phenomena, continuing along the lines of Joshi [Joshi et al. 84], Dale and Reiter [Dale and Reiter 95], Hirschberg [Hirschberg 85], Green [Green and Carberry 94], Passonneau [Passonneau 95], and Rubinoff [Rubinoff 87], but without appealing to Grice. Only after we have a large body of well-understood computational discourse systems should we try to generalize.

## References

- L. Cohen 1971. The logical particles of natural language. In *Pragmatics of Natural Language*. Y. Bar-Hillel (ed.), Dordrecht: Reidel.
- R. Dale and E. Reiter 1995. Computational Interpretations of the Gricean Maxims in the generation of referring expressions. *Applied Artificial Intelligence Journal*, 9. To appear.
- G. Gazdar 1979. *Pragmatics: Implicature, Presupposition and Logical Form*. New York: Academic Press.
- G. Gazdar 1980. Reply to Kiefer. *Linguisticae Investigationes*, 3, pp. 375-377.
- N. Green and S. Carberry 1994. A hybrid reasoning model for indirect answers. In *Proceedings of the 32nd Annual Meeting of the ACL*, Las Cruces, NM.
- J. Hirschberg 1985. *A Theory of Scalar Quantity Implicature*. PhD thesis, University of Pennsylvania.
- A. Joshi, B. Webber, R. Weischedel 1984. Living up to expectations: computing expert responses. In *Proceedings of AAAI-84*.
- F. Kiefer 1979. What do conversational maxims explain? *Linguisticae Investigationes*, 3, pp. 57-74.
- S. Levinson 1983. *Pragmatics*. Cambridge University Press, Cambridge.
- R. Passonneau 1995. Integrating Gricean and Attentional Constraints. In *Proceedings of IJCAI-95*.
- V. Poznanski 1992. A relevance-based utterance processing system. University of Cambridge Computer Laboratory technical report no. 246.
- E. Reiter 1991. A new model of lexical choice for nouns. *Computational Intelligence*, 7(4): 240-251.
- R. Rubinoff 1987. Scalar Implicature and the Given/New Distinction. Unpublished paper from the 1987 Penn Linguistics Colloquium.
- J. Sadock 1978. On testing for conversational implicature. In *Syntax and Semantics 9: Pragmatics*, P. Cole (ed.), New York: Academic Press.
- D. Sperber and D. Wilson 1986. *Relevance: communication and cognition*, Harvard University Press.
- D. Wilson and D. Sperber 1981. On Grice's theory of conversation. In *Conversation and Discourse*, P. Werth (ed.), New York: St. Martin's Press.
- D. Wilson and D. Sperber 1988. Representation and relevance. In *Mental Representations*, R. Kempson (ed.), Cambridge University Press.