Links & Letters 1, 1994 85-106

Outline of relevance theory[1]

Deirdre Wilson*, Dan Sperber**
*Department of Linguistics University College London
**CNRS and CREA Ecole Polytechnique, Paris

ABSTRACT

In this paper, we outline a relevance-based approach to pragmatics, the theory of utterance interpretation. The main aim of a pragmatic theory is to explain how the hearer recognises the intended interpretation of an utterance. We argue that this interpretation is not decoded but inferred, by a non-demonstrative inference process in which considerations of relevance play a central role. We offer a definition of relevance, and argue that every utterance creates an expectation of relevance in the hearer, with the preferred interpretation being the one that best satisfies that expectation of relevance. The theory is illustrated by applying it to a wide range of examples.

1. Introduction

Pragmatics is the study of the general cognitive principles and abilities involved in utterance interpretation, and of their cognitive effects. In construct-ing an explanatory pragmatic theory, a variety of specific problems must be solved. Utterances may be ambiguous or referentially ambivalent, as in (1) :

(l) The football team gathered round their coach.

Pragmatic theory should explain how the hearer of caracter (l) decides which football team the speaker has in mind, and whether 'coach' was intended to mean *bus* or *games teacher.* Utterances have not only explicit content but implicit import, as in (2):

(2) a. *Peter:* Is George a good sailor?
b. *Mary:* ALL the Te English are good sailors.

Pragmatic theory should explain how (2b) is understood as implying that George is a good sailor. Utterances may be metaphorical or ironical, as in (3) and (4):

1. Deirdre Wilson would like to thank the faculty and students of the University of Minho, Portugal, and in particular Dr. Helio Osvaldo Alves and Mrs Helen Santos Alves, for their warm hospitality at the Linguistics Meeting in November 1985 at which an early version of this paper was first delivered.

(3) Their friendship blossomed.

(4) *Mart, of Peter, who has just tripped over his own feet:*
Peter's just like Rudolf Nureyev.

Pragmatic theory should describe and explain the differences between literal and non-literal interpretation. More generally, the style of an utterance may affect its interpretation - compare the mildly witty (4) with the explicitly critical (5):

(5) Peter is very clumsy.

Pragmatic theory should describe such stylistic effects and explain how they are achieved. In this paper, we outline a pragmatic theory - relevance theory which offers a unitary solution to these and other pragmatic problems; the theory is developed in more detail in our book *Relevance: Communication and Cognition* (Sperber and Wilson 1986; see also Sperber and Wilson 1987).

2. The code model of communication

It is clear from examples (1)-(5) that understanding an utterance involves more than merely knowing the meaning of the sentence uttered. The hearer of (l) must not only know the two possible meanings of the word 'coach', but also decide which meaning the speaker intended to convey. The hearer of (2) must not only know the meaning of the sentence uttered, but infer what was implicitly conveyed. The hearer of (3) or (4) must not only know the literal meaning of the sentence uttered, but decide whether the utterance was literally, metaphorically or ironically intended. The stylistic differences between (4) and (5) cannot be explained in purely semantic terms. The central aim of pragmatic theory is to describe the factors other than a knowledge of sentence meaning that affect the interpretation of utterances.

Until about twenty years ago, there seemed to be only one possible approach to pragmatics. It was almost universally assumed that communication in general, and verbal communication in particular, are achieved by encoding and decoding messages. On this account - the *code,* or *semiotic,* account - communication involves a set of unobservable messages, a set of observable signals, and a code: that is, a method of pairing signals with messages. The communicator, on deciding to convey a certain message, transmits the signal associated with that message by the code; the hearer, on receiving the signal, recovers the message associated with it by the code. In the case of verbal commmunication, the observable signals would be the phonetic (or graphic) representations of utterances, the messages would be the thoughts that the speaker wanted to convey, and the task of pragmatics would be to discover the code that hearers use to recover the intended message from the observable signal.

Many linguists have assumed without question that the code model of pragmatics is correct. It is easy to see why. There is no doubt that utterance interpretation involves an element of decoding: the grammar of a natural language just is a code which pairs phonetic and semantic representations of sentences, and there is no doubt that understanding an utterance involves recovering the phonetic representation of the sentence uttered and decoding it into the associated semantic representation. However, as examples (1)-(5) show, there is more to understanding an utterance than merely recovering the semantic representation of the sentence uttered: there is a gap between the semantic representations of sentences and the thoughts communicated by utterances.

Advocates of the semiotic approach to pragmatics assume that this gap can be filled by an extra layer of encoding and decoding. They assume, in other words, that pragmatics is an extension of grammar: that speakers of English know a pragmatic code which is used to disambiguate utterances in English, recover their implicit import, distinguish their literal and figurative meanings, and determine their stylistic effects. However, this assumption is very far from being justified.

The most general problem for the code model is its conception of what communication is designed to achieve. On the code model, the speaker's thoughts, encoded into an utterance, should be replicated in the hearer by a decoding process. The result of verbal communication should be an exact reproduction in the hearer of the thoughts the speaker intended to convey. However, the most cursory examination of ordinary conversation reveals that in the case of implicit import, figurative interpretation and stylistic effects, such reproduction is rarely intended or achieved. For example, the implicit import of (3) can be described in a number of different ways. What exactly is the implicit message it was intended to convey: that their friendship developed naturally, that it developed from small beginnings, that it grew into something beautiful, that like a flower it was destined to fade? The basic assumption of the code model - that a determinate subset of these messages must have been actually encoded and decoded - does not seem remotely plausible.

The existence of indeterminacies in interpretation suggests a fundamental inadequacy in the code model of communication. Where indeterminacy is involved, it seems that the most that communication can achieve is to bring about some similarity between the thoughts of communicator and audience. How could the code model describe those cases where similarity, rather than identity, is intended and achieved? The solution which comes to mind would consist in adding to the determinate output of the decoding process some blurring mechanism. Such an obviously ad hoc solution is hardly worth developing.

To the extent that the code model of pragmatics has been successful, its successes have been achieved by investigating a very restricted range of data. It

is obvious that utterance interpretation is highly context-dependent; yet the successes of the code model have generally been achieved by looking at utterances in which the role of context is either minimal or very easy to describe.

For example, although the pronoun 'I' refers to different people in different contexts, it almost invariably refers to whoever is speaking at the time. It is thus possible to write a decoding rule instructing the hearer of (6), on hearing the word 'I', to identify the speaker and interpret the pronoun as referring to Mary:

(6) *Mary:* I am unhappy today.

However, to be successful, the code model of pragmatics would have to show, not just that one pronoun can be dealt with along these lines, but that all can. Other pronouns are less amenable to the decoding approach.

Suppose that as I give a lecture, I make a slip of the tongue. You turn to your neighbour and whisper:

(7) That was interesting.

What decoding rule, analogous to the rule just given for 'I', could your neighbour use to decide that the pronoun 'that' referred to the slip of the tongue I had just made, rather than, say to the example I had just been discussing, the theoretical claim I had just made, or the fact that a strange bird had just flown past the window? The code model of pragmatics tends to ignore such cases, but an adequate pragmatic theory must deal with them.

Similarly, the code model of pragmatics tends to concentrate on a few, relatively restricted types of implicit import which are only minimally context dependent. For example, in most contexts, the speaker of (8) would implicitly convey (9):

(8) Some of my friends stayed away.
(9) Not all of my friends stayed away.

It would thus be possible to set up a decoding rule associating utterances of the form in (8) with implications of the form in (9), and to prevent the rule from operating in a restricted class of contexts.

Often, however, the implicit import of an utterance is highly context dependent. Consider (10):

(10) I'll be in Dublin tomorrow.

In different contexts, (10) would have widely different implications. For example, said by Mary to Peter, who has just asked her to dinner in London

tomorrow, it will imply that Mary can't come to dinner; said to Peter, who lives in Dublin and has just asked Mary when they can next meet, it will imply that they can meet the next day; and so on. Not only would it be hard to write a decoding rule assigning to each utterance of (10) the appropriate interpretation in the appropriate context: it would also be totally pointless. To see the implications of (10), all Peter needs is his knowledge of the world, and in particular his knowledge of the speaker and the situation, and his general reasoning abilities. Given these, he can *work out* the implications of (10) for himself. Might this not be true of (8)-(9) as well?

3. The inferential account of communication

It is certainly true that communication does not necessarily involve the use of a code. Consider (11):

(11) a. *Peter*: Did you enjoy your skiing holiday?
b. *Mary*: (displays her leg in plaster)

Here, Mary clearly communicates that her skiing holiday did not live up to expectations. Yet there is no code which states that displaying one's leg in plaster means that one's skiing holiday has not gone according to plan. To account for such examples, some alternative to the code model of communication is needed.

Intuitively, Peter does not need a code to understand Mary's behaviour in (11) because he can use his knowledge of the world and his general reasoning abilities to work out what she must have intended to convey. On this account -an *inferential* account - communication is achieved not by coding and decoding messages, but by providing evidence for an intended hypothesis about the communicator's intentions. Communication is successful when the audience interprets the evidence on the intended lines. Failures in communication result from misinterpretation of the evidence provided. Indeterminacy results from the fact that a single utterance may provide evidence for a range of related hypotheses, all similar enough to the thoughts the communicator wanted to convey.

In (11b), for example, Mary provides evidence that she broke her leg on holiday, and that as a result her holiday did not live up to expectations. However, from a logical point of view this is not the only hypothesis that Peter might have entertained. He might have assumed, for example, that Mary broke her leg before leaving, and as a result did not go on holiday at all.

This example brings out a fundamental difference between code and inferential models of communication. According to the inferential model, the interpretation of utterances, like the interpretation of evidence in general, is

always subject to risk. There are always alternative ways of interpreting a given piece of evidence, even when all the correct procedures for interpretation are applied. These procedures may yield a best hypothesis, but even the best hypothesis may not be the correct (i.e., the intended one). By contrast, decoding procedures, when correctly applied to an undistorted signal, guarantee the recovery not only of an interpretation, but of the correct one (i.e., the intended interpretation). The two approaches thus start from radically different assumptions about the nature of communication itself.

Inferential communication involves the formation and evaluation of hypotheses about the communicator's intentions. Little attention has been paid to the processes of pragmatic hypothesis formation. However, the work of Grice (1975, 1978) is a major contribution to the study of hypothesis confirmation or evaluation within an inferential theory of communication which Grice (1957, 1968) was also largely responsible for developing.

Grice suggested that speakers try to meet certain standards in their communicative behaviour, and that hearers use these standards in evaluating alternative hypotheses about the speaker's communicative intentions. He set out these standards as a co-operative principle and *maxims of conversation* addressed to speakers:

Co-operative principle: Make your 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.

Maxims of conversation

Quality Try to make your contribution one that is true.

(a) Do not say what you believe to be false.

(b) Do not say that for which you lack adequate evidence.

Quantity:

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

(b) Do not make your contribution more informative than is required.

Relation: Be relevant.

Manner: Be perspicuous.

(a) Avoid obscurity of expression.

(b) Avoid ambiguity.

(c) Be brief.

(d) Be orderly.

If speakers observe the co-operative principle and maxims, and hearers expect them to, any interpretation incompatible with them can be rejected. For example, if the speaker of (1) above could not truthfully have claimed that the football team gathered round their bus, then this interpretation cannot be correct. If the speaker of (2b) above could not have observed the maxim of

relevance without intending to communicate that George is a good sailor, then the information that George is a good sailor must figure in any acceptable interpretation of (2b).

The Gricean approach to pragmatics, while undoubtedly better equipped than the code model to deal with the full range of pragmatic data, leaves a number of important questions unanswered. First, there are problems about the formulation of the maxims themselves. What is meant by being 'as informative as required'? What is meant by 'relevance'? What is meant by 'clarity' and 'brevity'? Until we have some idea of what these terms mean, we cannot claim to have a theory at all. Second, there are problems about where the maxims come from. Are they universal? If so, why? Are they culture specific? If so, in what respects? Third, are these exactly the right maxims? Are there more? Could we do with less?

There are also more general questions about the nature and role of context, and about the process of pragmatic hypothesis formation itself. For an ambiguous utterance, the grammar generates a range of alternative interpretations. For an utterance that is referentially ambivalent, the range of possible interpretations is determined on the one hand by the grammar which indicates, for example, that 'I' must refer to the speaker - and on the other by encyclopaedic and contextual information - which indicate, for example, who the speaker is on any given occasion. But what is the source of hypotheses about the implicit import of utterances, about figurative interpretation and stylistic effects? These questions must be answered by an adequate pragmatic theory.

4. Cognition: relevance

We would like to suggest that the standards governing inferential comunication have their source in some basic facts about human cognition. Humans pay attention to some phenomena rather than others; they represent these phenomena to themselves in one way rather than another; they process these representations in one context rather than another. What is it that determines these choices? Our suggestion is that humans tend to pay attention to the most relevant phenomena available; that they tend to construct the most relevant possible representations of these phenomena, and to process them in a context that maximises their relevance. Relevance, and the maximisation of relevance, is the key to human cognition.

This has an important consequence for the theory of communication. A communicator, by the very act of claiming an audience's attention, suggests that the information he is offering is relevant enough to be worth the audience's attention. We would like to show that this simple idea - that communicated information comes with a guarantee of relevance - is enough on its own to yield an explanatory pragmatic theory.

But what is relevance? We claim that information is relevant to you if it interacts in a certain way with your existing assumptions about the world. Here are three examples of the type of interaction we have in mind.

Case A

You wake up with the following thought:

(12) a. If it's raining, I'll stay at home.

You look out of the window and discover:

(12) b. It's raining.

In this case, from your existing assumption (12a) and the new information (12b), you can deduce some further information not deducible from either the existing assumption or the new information alone:

(12) c. I'll stay at home.

To deduce (12c), you have to use both old and new information as joint premises in an inference process. Intuitively, the new information (12b) would be relevant in a context containing assumption (12a). We claim that it is relevant precisely because it enables such a joint inference process to take place. Let us say that assumption (12a) is the *context* in which the new information (12b) is processed, and that (12b) *contextually implies* (12c) in the context (12a). Then we claim that new information is relevant in any context in which it has contextual implications, and the more contextual implications it has, the more relevant it will be.

Assumptions about the world may vary in their strength: you may have more or less evidence for, more or less confidence in, your assumption that it is raining. New information may affect the strength of your existing assumptions, as in the following case:

Case B

You wake up, hearing a pattering on the roof, and form the hypothesis that:

(13) a. It's raining.

You open your eyes, look out of the window, and discover that:

(13) b. It IS raining.

Here, the new information (13b) strengthens, or confirms, your existing assumption (13a). It would also, intuitively, be relevant to you in a context containing assumption (13a). We claim that (13b) is relevant precisely because it strengthens an existing assumption of yours. New information is relevant in any context in which it strengthens an existing assumption; and the more assumptions it strengthens, and the more it strengthens them, the more relevant it will be.

If new information can achieve relevance by strengthening an existing assumption, it should also achieve relevance by contradicting, and eliminating, an existing assumption, as in the following case:

Case C

You wake up, as in case B, hearing a pattering on the roof, and form the hypothesis that:

(14) a. It's raining.

This time, when you open your eyes and look out of the window, you disccover that the sound was made by leaves falling on the roof, and that actually:

(14) b. It's not raining.

Let us assume that when new and old assumptions contradict each other, the weaker of the two assumptions is abandoned. Here, the new information (14b) would provide conclusive evidence against the old assumption (14a), which would therefore be abandoned. Intuitively, (14b) would be relevant in these circumstances. We claim that new information is relevant in any context in which it contradicts, and leads to the elimination of, an existing assumption; and the more assumptions it eliminates, and the stronger they were, the more relevant it will be.

These cases illustrate the three ways in which new information can interact with, and be relevant in, a context of existing assumptions: by combining with the context to yield contextual implications; by strengthening existing assumptions; and by contradicting and eliminating existing assumptions. Let us group these three types of interaction together and call them *contextual effects*. Then we claim that new information is relevant in any context in which it has contextual effects, and the greater its contextual effects, the more relevant it will be.

This comparative definition of relevance is inadequate in one respect, as the following example shows:

Case D

You wake up, thinking:

(15) a. If it rains, I'll stay at home.

Then EITHER:

You look out of the window and see:

(15) b. It's raining.

OR:

You look out of the window and see:

(15) c. It's raining and there's grass on the lawn.

Intuitively, (15b) would be more relevant to you than (15c) in the context (15a). Yet (15b) and (15c) have exactly the same contextual effects in this context: they both have the contextual implication (15d), and no other contextual effect at all:

(15) d. I'll stay at home.

If comparisons of relevance are based solely on contextual effects, then the difference in relevance between (15b) and (15c) is inexplicable.

This difference, we suggest, can be explained in terms of the intuition underlying Grice's Manner maxims. The intuition is that speakers should make their utterances easy to understand: in our terms, that speakers should make the contextual effects of their utterances easy to recover. Now it is clear that though (15b) and (15c) above have exactly the same contextual effects in the context (15a), you would have to work harder to recover them from (15c) than from (15b): since (15c) includes (15b) as a subpart, (15c) will require all the effort needed to process (15b), and more besides. This extra processing effort detracts from the relevance of (15c).

We thus propose the following comparative definition of relevance:

Relevance:

(a) Other things being equal, the greater the contextual effects, the greater the relevance.

(b) Other things being equal, the smaller the processing effort, the greater the relevance.

An individual with finite processing resources, who is aiming to maximise relevance, should pay attention to the phenomena which, when represented in the best possible way, and processed in the best possible context, seem likely to yield the greatest possible contextual effects in return for the available processing effort. Relevance, and the aim of maximising relevance, is the key to cognition.

5. Communication: the principle of relevance

If humans pay attention only to relevant information, a communicator, by claiming an audience's attention, creates an expectation of relevance. He creates an expectation, in particular, that the information he is attempting to convey, when processed in a context he believes the audience has accessible, will be relevant enough to be worth the audience's attention. But how relevant is that? What exactly is the expectation of relevance that each act of inferential communication creates?

On the contextual effect side, the expectation is one of adequacy. In the most straightforward cases of verbal communication, the speaker suggests that the proposition he intends to express, when processed in a context he expects the hearer to have accessible, will yield enough contextual effects to be worth the hearer's attention. How much is required in the way of contextual effects will vary from individual to individual and occasion to occasion. How the level of adequacy is fixed and varies is an interesting question, but intuitions about particular examples are clear enough.

On the processing effort side, as Grice's Manner maxims suggest, the expectation is of more than adequacy. A speaker who wants to achieve a certain range of contextual effects must make sure that they are as easy as possible for the hearer to recover: that is, he must make sure that his utterance puts the hearer to no unjustifiable processing effort. This is in the speaker's interest as well as the hearer's, for two reasons: firstly, the speaker wants to be understood and any increase in unjustifiable processing effort required of the hearer is an increase in risk of misunderstanding; secondly, any increase in processing effort detracts from overall relevance, and might cause the overall relevance of the utterance to fall below an acceptable level.

Let us say that an utterance (or more generally an act of inferential communication) which, on the one hand, achieves an adequate range of contextual effects, and on the other hand, achieves it for the minimum justifiable processing effort, is optimally relevant. Then Grice's maxim of relevance can be replaced by the following *principle of relevance:*

Principle of relevance

Every act of inferential communication creates an expectation of optimal relevance.

We believe that this single principle (or rather a more technical version developed in Sperber and Wilson 1986) is enough on its own to yield an explanatory pragmatic theory.

The fact that an utterance creates an expectation of optimal relevance does not mean that it will actually be optimally relevant to the hearer. I may tell you something in the mistaken belief that you do not already know it, or speak simply to distract your attention from relevant information elsewhere. In this case, you will be unable to find an interpretation which satisfies your expectation of relevance.

Let us say that an utterance on a given interpretation is *consistent with the principle of relevance* if a rational communicator might have expected it to be optimally relevant to the hearer (i.e., to achieve an adequate range of contextual effects as economically as possible). Then it is easy to show that every utterance has at most one interpretation which is consistent with the principle of relevance.

We will show this using our example of disambiguation, (1), with possible interpretations (16a) and (16b):

(1) The football team gathered round their coach.
(16) a. The football team gathered round their games teacher
b. The football team gathered round their bus.

Logically speaking, there are two routes that the disambiguation process might follow: one interpretation may be more accessible than the other, and be tested first for consistency with the principle of relevance; or both interpretations may be equally accessible, and be tested in parallel. We consider each possibility in turn.

Suppose that interpretation (16a) is more accessible than (16b), and is therefore the first to be tested for consistency with the principle of relevance. Suppose, moreover, that there is an easily accessible context in which this interpretation would have a manifestly adequate range of contextual effects, and that there would have been no obviously cheaper way of obtaining them. Then as long as a rational communicator could have foreseen this situation, interpretation (16a) is consistent with the principle of relevance, and is the only interpretation consistent with the principle of relevance, as the following argument shows.

Imagine that Mary, in uttering (1), had wanted to convey interpretation (16b), but had foreseen that interpretation (16a) would be both more accessible and consistent with the principle of relevance. By reformulating her utterance

to eliminate this unwanted interpretation - for example, by saying 'The team gathered round their bus', thus eliminating interpretation (16a) entirely - she could have spared her hearer the effort of first accessing and processing interpretation (16a), then accessing and processing interpretation (16b), and then engaging in some form of inference process to choose between them. In other words, she could have achieved the intended range of contextual effects at a much reduced processing cost, and at a much smaller risk of misunderstanding, by rephrasing her utterance. On this interpretation, although Mary's utterance in (1) may achieve an adequate range of contextual effects, it would put her hearer to some unjustifiable processing effort in recovering them, and is not consistent with the principle of relevance.

What would happen if interpretations (16a) and (16b) were equally accessible, and were thus simultaneously tested for consistency with the principle of relevance? Suppose that Peter has easy access to a context in which interpretation (16a) has an adequate range of contextual effects, while a comparable context for (16b) is much less accessible or not accessible at all. As long as Mary could rationally have foreseen this situation, interpretation (16a) is consistent with the principle of relevance, and is the only interpretation consistent with the principle of relevance. If Mary had intended to convey interpretation (16b), she could manifestly have spared Peter some processing effort by rephrasing her utterance to eliminate the unwanted interpretation (16a). For example, by saying 'The team gathered round their bus', she could have spared him the effort of accessing and processing both (16a) and (16b), and then engaging in some inference process to choose between them. On this interpretation, Mary's utterance (1) would put Peter to some unjustifiable processing effort, and is not consistent with the principle of relevance.

Finally, what would happen if interpretations (16a) and (16b) were equally accessible, and, moreover, yielded comparable contextual effects at comparable processing costs? Then there would be no way of choosing between the two interpretations, the ambiguity would remain unresolved and neither interpretation would be consistent with the principle of relevance since each could only be preferred, if at all, after an effort of comparison which Mary could easily have spared Peter. Thus the principle of relevance provides an account, not just of successes, but also of failures of disambiguation.

This example shows that, whatever the procedures used in disambiguation, the first interpretation - if any - tested and found consistent with the principle of relevance is the only interpretation consistent with the principle of relevance. A speaker who does not intend this interpretation should rephrase her utterance to eliminate it. This general principle applies to every aspect of utterance interpretation, as our remaining examples (2)-(5) will show.

6. Pragmatics and relevance

Example (2) above illustrates the fact that a speaker may communicate more than is explicitly expressed:

(2) a. *Peter:* Is George a good sailor?
b. *Mary:* ALL the English are good sailors.

Here, Mary implicitly communicates that George is a good sailor. How are such implications conveyed?

Grice suggests that the implicit import of an utterance is recovered by reference to the co-operative principle and maxims of conversation. The speaker implicitly communicates any assumption which must be added to the interpretation of an utterance to make it accord with the co-operative principle and maxims. Adopting his terminology, we will refer to such implicitly communicated assumptions as *implicatures*.

The situation in which (2b) is uttered is the following. Peter, by asking whether George is a good sailor, indicates that the information that George is (or is not) a good sailor would be optimally relevant to him. On hearing (2b), he can easily access the contextual assumption that George is English, recover as a contextual implication the information that George is a good sailor, and, by hypothesis, obtain an adequate range of contextual effects by processing this information in an immediately accessible context. Question: could Mary have rationally expected her utterance to be optimally relevant to Peter without also expecting him to supply the contextual assumption that George is English and derive the contextual implication that George is a good sailor? That is, could she have rationally expected her utterance to be optimally relevant when processed in some quite different way? Answer: no. In this situation, the interpretation just described would be the first accessible interpretation consistent with the principle of relevance, and hence the only interpretation consistent with the principle of relevance. Hence both the assumption that George is English and the conclusion that George is a good sailor are implicated by Mary.

In fact, if Mary was aiming at optimal relevance, she must have intended to communicate more than this: if this was all she intended to communicate, she could have communicated it more economically by saying simply, 'Yes'. To compensate for the extra effort of processing her indirect answer (2b); she must have expected to achieve some additional contextual effects not derivable from the direct answer 'Yes', and these will also be implicatures of her utterance. It is easy to see what these implicatures might be. For example, by supplying the names of other English people, Peter could recover a series of contextual implications to the effect that they are also good sailors; by supplying contextual assumptions such as 'If all the English are good sailors, then the English have

much to be proud of', 'If all the English are good sailors, then England deserves a good navy', he could recover a further series of contextual implications. To obtain an interpretation consistent with the principle of relevance, Peter must be able to derive enough such implicatures to offset the extra processing effort incurred.

As this example shows, implicatures may differ in their saliency. For the hearer of (2b), the implicature that George is a good sailor is strongly salient, but there is an indefinite array of further implicatures such as 'The English have much to be proud of', 'England deserves a good navy', which are only weakly salient. There is a necessary connection between strength (or saliency) of implicatures and determinacy of interpretation. An interpretation is determinate to the extent that its implicatures are strong, and implicatures are strong to the extent that there are no alternative assumptions which a speaker aiming at optimal relevance might have expected the hearer to access and use. In (2b) above, the implicature that George is a good sailor is very strong precisely because it forms an essential part of any interpretation consistent with the principle of relevance; the implicatures that the English have much to be proud of, that England deserves a good navy, and so on, are weak precisely because there are alternative, equally accessible interpretations of (2b), which would also be consistent with the principle of relevance, and in which these particular assumptions would play no role.

This example illustrates the importance of processing effort in utterance interpretation. By demanding extra processing effort - for example, by answering a question indirectly - the speaker can encourage the hearer to look for additional contextual effects in the form of additional weak or strong implicatures. This has important consequences for the analysis of metaphor and irony, and more generally of stylistic effects.

Grice analyses metaphor and irony as deliberate violations of the maxim of truthfulness. According to Grice, the hearer, faced with such a violation, reasons that the speaker must have been trying to communicate some logically related implicature which does satisfy the maxim of truthfulness. Thus, the patently false metaphorical utterance (3) might be interpreted as implicating (17), and the patently false ironical utterance (4) might be interpreted as implicating (18):

(3) Their friendship blossomed.
(17) Their friendship grew like a blossom.
(4) Peter is just like Rudolf Nureyev.
(18) Peter is not at all like Rudolf Nureyev.

This account is not compatible with relevance theory. In the first place, a speaker aiming at optimal relevance could not have said (3) merely intending to implicate (17), or (4) merely intending to implicate (18), since she could have spared her hearer some unnecessary processing effort by saying (17) or (18) directly.

In the second place, relevance theory has nothing comparable to Grice's maxim of truthfulness. According to the maxim of truthfulness, every utterance must explicitly express a belief of the speaker's. What follows from relevance theory is something much weaker: to be consistent with the principle of relevance, an utterance must achieve an adequate range of contextual effects, and achieve them as economically as possible. There are many utterances which satisfy this weaker condition without explicitly expressing a belief of the speaker's, as we will shortly show.

Inferential communication often involves a deliberate exploitation of resemblances. I may invite you for a drink by imitating the act of drinking; I may show you how to get to my house by drawing a diagram of the route. To understand me fully, you have to notice a resemblance between my action and the act of drinking, between my diagram and the intended route. Actions and objects resemble each other to the extent that they have properties in common. However, a representation can achieve its aim without sharing *all* its properties with the original. For example, when I imitate the act of drinking, there is no glass in my hand; when I draw you a route map, every landmark is clearly labelled. No rational addressee would assume that these properties of the representation are also shared by the original.

Where representation by resemblance is involved, the addressee must make some assumption about which properties of the representation are also shared by the original. In this, as in every other aspect of interpretation, the minimal assumption-that is, the first accessible assumption - consistent with the principle of relevance is the only assumption consistent with the principle of relevance. A bust of Napoleon may be made of white plaster, have no arms and legs, be found in a certain museum, and have been bought for a certain price. No rational addressee would attribute these properties to Napoleon. Nor should he: if a communicator aiming at optimal relevance could have intended to convey an adequate idea of Napoleon *without* intending to suggest that Napoleon was made of white plaster, lacked arms and legs, etc., then he must be credited with this minimal intention: it is the only intention which a communicator aiming at optimal relevance could have hoped to achieve.

Verbal communication may involve the exploitation of linguistic resemblances. Direct quotation, as in (19b), is a case in point:

(19) a. *Peter:* What is the last line of 'Rule Britannia'?
b. *Mary:* 'Britons never never never shall be slaves.'

In uttering (19b), Mary is not expressing a belief of her own: she is merely reproducing the words of a song. Her utterance is intended, not as a faithful expression of her own belief, but as a faithful representation of an original. As with the bust of Napoleon, not every property of Mary's utterance need be shared by the original: for example, Mary's utterance may have been spoken

rather than sung, but it does not follow that the last line of 'Rule Britannia' must be spoken rather than sung; Mary may have spoken in a Devon accent, but it does not follow that 'Rule Britannia' must be sung in a Devon accent, and so on. As always, a speaker will expect her hearer to attribute to the original only the minimal set of properties needed to achieve an interpretation consistent with the principle of relevance.

Utterances have not only form but content, i.e. logical and contextual implications. Verbal communication often exploits resemblances between the content of utterances: a good translation or summary, which shares logical and contextual implications with the original, is an example. However, utterances are not alone in having content. Thoughts also have content, (i.e., logical and contextual implications) and in virtue of their content, utterances may be used to represent not other utterances but thoughts. Seen from this perspective, it is clear that every utterance is used to represent a thought: the very thought the speaker intended to communicate.

How closely must the content of an utterance resemble the speaker's thought: that is, how many logical and contextual implications must they share? According to the maxim of truthfulness, full identity between the contents of representation and original is required. Let us say that in this case the utterance is a *literal* representation of the speaker's thought. According to relevance theory, the speaker guarantees not the truth of her utterance, but merely its optimal relevance. Sometimes this optimal relevance can only be achieved by a strictly literal interpretation of the utterance, but this is often not the case. Expecting optimal relevance rather than literal truthfulness, the hearer must in every case determine how the speaker intended to achieve this relevance, in particular what degree of resemblance she intended between the content of her utterance and the thought she intended to communicate. The hearer should take the utterance to be a literal representation of the speaker's thought only if this is the minimal assumption (i.e., the first accessible assumption, consistent with the principle of relevance).

The difference between the two approaches can be seen in their analyses of (20b) and (20c) as answers to the question in (20a):

(20) a. *Peter:* How far is Nottingham from London?
b. *Mary:* 120 miles.
c. *Mary*: 118 miles.

According to the maxim of truthfulness, Mary should not say that Nottingham is 120 miles from London unless she believes that it is exactly 120 miles from London. If she believes that the true distance is in fact 118 miles, then she would violate the maxim of truthfulness by answering as in (20b). However, there are many situations in which a speaker aiming at optimal relevance should prefer the non-literal (20b) to the strictly literal (20c). Suppose Peter, who

normally drives at about 60 miles an hour, is trying to decide when he should leave London for dinner in Nottingham. From both (20b) and (20c) he can recover the contextual implication that it will take him about two hours to drive to Nottingham, and that he should plan his journey accordingly. However, given that mental calculation is easier to do in round numbers, it will cost him less effort to recover these implications from (20b) than from (20c), and a speaker aiming at optimal relevance should prefer (20b) to (20c). Since there is an easily accessible less-than-literal interpretation of (20b) which is consistent with the principle of relevance, there is no need for the hearer to consider the literal interpretation at all.

Metaphorical utterances such as (3) fit straightforwardly into this pattern:

(3) Their friendship blossomed.

By processing (3) in the context of his encyclopaedic knowledge of blossoming, the hearer might derive a number of contextual implications. For instance, the implication that their friendship belonged to the plant kingdom, carry no plausible information and could hardly have been intended by the speaker to contribute to the relevance of his utterance. Other contextual implications, on the contrary, do contribute to the relevance of the utterance and can therefore be assumed to have been at least weakly implicated by the speaker, in the sense that the speaker intended the hearer to derive some such implications, if not exactly these. Thus, the hearer might conclude that the friendship being discussed grew from small beginnings, in a favourable environment, by a natural process, into something beautiful, that was perhaps destined to fade. As with most metaphors, there is a substantial element of indeterminacy in the interpretation of (3), and its associated implicatures will be relatively weak. For a speaker who wanted to achieve a range of effects along these lines, (3) would be the most economical way of achieving them. Since (3) has an easily accessible non-literal interpretation which is consistent with the principle of relevance, there is no need for the hearer to consider the literal interpretation at all.

On this approach, metaphorical utterances such as (3) and rough approximations such as (20b) are non-literal representations of the speaker's thought. Ironical utterances, by contrast, may be fully literal. Their distinguishing feature is that the thought they (literally or non-literally) represent is a thought attributed by the speaker to someone else.

This property is not unique to irony. All forms of indirect speech involve the attribution of a thought to some source other than the speaker. Consider (21):

(21) a. *Peter:* Have you read the manifesto?
b. *Mary:* Yes. We'll all be rich and happy if we vote for them.

In saying (21b), Mary may be reporting what the manifesto promises, not what she herself believes. At the same time, by varying her tone of voice or facial expression, or simply by exploiting contextual assumptions that she believes Peter has accessible, she may convey her attitude to the promise in the manifesto: she may communicate, for example, that she believes it, or that she rejects it with scorn. In the latter case, of course, her utterance will be ironical.

Irony, then, involves the implicit expression by the speaker of an attitude - scornful, mocking, contemptuous - to an implicitly attributed thought. Example (4) above fits straightforwardly into this pattern.

(4) *Mary, of Peter, who has just tripped over his own feet*
Peter's just like Rudolf Nureyev.

Here, it is not plausible to assume that Mary has explicitly expressed her own belief. Suppose that a friend of Peter's is constantly comparing him to Rudolf Nureyev. Then by repeating this opinion to him in a situation where it is clearly ridiculous, Mary can make fun of it and of anyone who would believe it. Or suppose that Peter has a streak of vanity, a tendency to enter a room as if he thought all eyes were on him. Then by representing the sort of opinion of Peter she implicitly attributes to Peter himself in a situation where it is clearly ludicrous, Mary can simultaneously make fun of it and of anyone who would believe it.

By saying (4), Mary achieves a range of contextual effects not obtainable from the strictly literal utterances (5) and (18):

(5) Peter is very clumsy.
(18) Peter is not at all like Rudolf Nureyev.

Since Mary could not normally dissociate herself from the belief that Peter is just like Rudolf Nureyev without also communicating that she believes (5) and (18), these will be strong implicatures of her utterance. However, Mary also communicates an indefinite array of weak implicatures - for example, that Peter looks ridiculous, that the idea that he is like Rudolf Nureyev is laughable, that anyone who would entertain this or any similar idea is a fool, and so on- whose recovery depends on the identification of her attitude of mockery or scorn. Relevance theory thus accounts for the differences between ironical and non-ironical utterances in a way that Gricean approaches notably fail to do.

More generally, relevance theory sheds light on the cognitive effects of style. Some stylistic effects are not deliberately achieved: for example, the speaker's choice of vocabulary may betray his social or political attitudes. Such attitudes may also be deliberately communicated. To take just one illustration, modern English speakers who prefer the form of words 'he or she' to the more

economical form 'he' communicate that, for them, choice of the more economical form would carry unwanted implications.

Often, the style of an utterance directly affects its propositional content. Compare (22a) and (22b):

(22) a. There was water everywhere.
b. There was water, water everywhere.

Within relevance theory, the traditional claim that repetition has an emphatic effect can be explained and made more precise. Since the repetition in (22b) demands additional processing effort, a speaker aiming at optimal relevance must expect it to achieve additional contextual effects. In interpreting (22a), the hearer must make some assumption about how much water there was. In interpreting (22b), he simply assumes that there was more water than could have been conveyed by the use of (22a): in other words, that there was a very great deal of water. The deliberate increase in processing effort is thus offset by an increase in implicatures.

Relevance theory can also shed light on the stylistic effects of what sociolinguists call 'register' variation - for example, between the relatively informal (23a) and the highly formal (23b):

(23) a. Peter bought a paper before leaving.
b. Peter purchased a newspaper prior to departure.

One factor known to affect processing effort is the frequency with which words are used. Thus (23a), with its relatively familiar vocabulary, requires less processing effort than (23b), with its relatively less familiar words. As a result, a speaker aiming at optimal relevance could prefer (23b) to (23a) only if he expected the additional processing effort to be offset by additional contextual effects. One obvious way of obtaining such effects would be to assume that the speaker of (23b) thinks that 'purchase' means something more than 'buy', 'prior to' means something more than 'before', and so on. Hence, (23b) will implicate that there were subtle differences between what Peter did and the commonplace act of buying a paper. Precisely because there are no clearcut semantic differences involved, the associated implicatures will be very weak, giving this formal style its simultaneously vague and portentous quality.

Some obvious counterexamples to Grice's maxim of brevity can be successfully analysed along these lines. Compare (24a) and (24b):

(24) a. I have no brothers or sisters.
b. I have no siblings.

By virtually any measure of brevity, (24b) is briefer than (24a); however, most English speakers would prefer (24a) to (24b), despite its extra length. In a relevance-based framework, Grice's notion of brevity is replaced by the notion of processing effort, which as we have seen, is affected by the relative frequency of words. Now 'sibling' is a very rare word indeed. The differences between (24a) and (24b) are straightforwardly explained on the assumption that the relative brevity of the word 'sibling' is not enough to offset the increase in processing cost resulting from its infrequency, so that (24a) is more economical overall. An anomaly in Grice's framework is thus removed.

7. Conclusion

In this paper, we have briefly sketched an explanatory pragmatic theory based on a single principle of relevance. Every act of inferential communication creates an expectation of optimal relevance, in the light of which hypotheses about the intended interpretation can be evaluated. The expectation of relevance has its source in universal human cognitive mechanisms. But what is the source of the hypotheses about the intended interpretation?

As already mentioned, the source of disambiguation hypotheses is the grammar, which assigns a range of possible semantic representations to every sentence uttered. The source of hypotheses about the intended reference of referential expressions is on the one hand the grammar, which determines a range of linguistically possible referents, and on the other hand the hearer's encyclopaedic and environmental knowledge, which determines a range of objects meeting the linguistically specified criteria. By choosing a unique semantic representation for every sentence uttered, and assigning referents to each of its referential expressions, the hearer recovers the explicit propositional content of the utterance.[2]

Implicatures have two sources. Some implicatures are contextual assumptions which the hearer was expected to use in processing the explicit propositional content of the utterance: like all contextual assumptions, such implicatures are derived from memory or from observation of the environment. Other implicatures are contextual implications which the hearer was expected to recover in processing the explicit propositional content of the utterance: like all contextual implications, such implicatures are derived by deductive inference from the explicit propositional content of the utterance and the context. The more salient the implicature, the stronger it is.

Metaphors are non-literal representations of the speaker's thought. Irony involves an implicit expression of attitude to an implicitly attributed thought.

2. For reasons of space, we have ignored the fact that more is involved in the recovery of explicit propositional content than disambiguation and reference assignment, for example, missing or ellipsed material must be restored and vaguenesses must be eliminated. We have also ignored the 'speech act' element. For fuller discussion, see Sperber and Wilson 1986, chapter 4, sections 2, 3 and 10.

Neither metaphor nor irony is deviant or a departure from the norm: indeed, an utterance will only be interpreted as literal if no non-literal interpretation will do.

As we have shown, the cognitive effects of metaphor and irony and style are analysable in terms of the notions of weak and strong implicature. More generally, we have suggested that the key to an explanatory theory of style lies in the correlation between the implicatures of an utterance and the processing effort it requires. Relevance theory thus promises new and satisfying answers to many questions raised, but left unanswered, by earlier accounts.[3]

References

GRICE, H.P. (1957). "Meaning". *Philosophical Review* 66: 377-88.

GRICE, H.P. (1968). "Utterer's meaning, sentence meaning and word meaning". *Foundations of Language* 4: 225-42.

GRICE, H.P. (1975). "Logic and conversation", Cole, P. and Morgan, J. (eds). *Syntax and semantics 3: Speech acts,* 41-58. New York: Academic Press.

GRICE, H.P. (1978). "Further notes on logic and conversation". Cole, P. and Morgan, J. (ed.) *Syntax and Semantics 9. Pragmatics,* 113-128. New York: Academic Press.

SPERBER, D. and WILSON, D. (1986). *Relevance: Communication and Cognition.* Oxford: Blackwell and Cambridge: Harvard University Press.

SPERBER, D. and WILSON D. (1987). "Précis of Relevance and Presumptions of relevance". *Behavioral and Brain Sciences* 10: 697-754.

3. For a fuller account of relevance theory, and of the issues discussed in this paper, see Sperber and Wilson 1986, and the précis and multiple review of *Relevance* in Sperber and Wilson 1987.