TESTING DIALOGUE PRINCIPLES IN TASK-ORIENTED DIALOGUES: AN EXPLORATION OF COOPERATION, COLLABORATION, EFFORT AND RISK

Bethan L. Davies

Abstract

This paper takes four behavioural principles which have been suggested as explanatory models for human conversation and tests them on a corpus of task-oriented dialogues (the HCRC Map Task Corpus). The principles chosen are Grice's Cooperative Principle, a folklinguistic notion of 'cooperation' (which we argue is often confused with the Gricean notion), Clark's Collaborative Theory, and Shadbolt's Principle of Parsimony. The aim of the study is to compare the explanatory power of each of these principles when they are applied to real language data.

Each of the principles was converted into a set of representative hypotheses about the types of behaviour which they would predict in dialogue. Then, a way of coding dialogue behaviour was developed, in order that the hypotheses could be tested on a suitably sized dataset. In particular, the coding system tried to distinguish between the levels of effort which participants used in their utterances. Finally, a series of statistical tests was undertaken to test the predictions of the hypotheses on the information generated by the coding system.

The strongest support was found for the Principle of Parsimony and its associate Principle of Least Individual Effort, at the expense of the Collaborative Principle and the Principle of Least Collaborative Effort. There is certainly evidence that speakers try to minimise effort, but this seems to be occurring on an individual basis – which can be to the cost of the overall dialogue and task performance – rather than on a collaborative basis. Some support was also found for Gricean Cooperation, although this is weakened by the difficulty in transforming the underspecified nature of Grice's work into a precise and unarguable set of predictions. However, a clear distinction can be drawn between Gricean Cooperation and the folklinguistic notion: even a broad definition of Grice is manifestly different from the predictions made for 'cooperation', and these indicators of 'cooperation' were not supported by the data.

1. Introduction

Over the last thirty years or so, a plethora of dialogue principles have been suggested to govern the management of dialogue. These vary from abstract theoretical concepts like Grice's (1975) Cooperative Principle, or Leech's (1983) complex of interdependent principles, to psychologically-oriented principles like Clark's Collaborative Theory (e.g. Clark 1996), or more computationally oriented concepts like Shadbolt's (1984) Principle of Parsimony. The aim of this study is to take four of these principles – Grice's Cooperative Principle, A folklinguistic notion of 'cooperation', Clark's Collaborative Principle, Shadbolt's risk-effort trade-off – out of their original settings, and test their ability to account for the behaviour of participants in a set of task-oriented dialogues.

These four principles have been derived from very different approaches to language, and the first challenge lies in interpreting them at the same level of language and to the same degree of specificity. Grice's Cooperative Principle (1975) is a very abstract notion which is open to being interpreted in different ways, particularly by those who are not familiar with Grice's other writings. We would argue that this has led to conflicting understandings of Grice's work and given rise to a folklinguistic notion of cooperation which is at odds with an explication more in keeping with Grice's work as a whole (Davies, to appear). One of the aims of this work is to clarify the Gricean position, and thus demonstrate this differentiation via the contrasting predictions which the two positions would make. For the other two principles, the shifts are rather different. In the case of the Collaborative Theory, it is more a question of expanding the focus from very specific discourse elements (in the context of psycholinguistic experiments) to a wider concept of what collaboration would mean in dialogue as a whole. Therefore, the attempt is to reinterpret Collaboration in a more general way. Finally, for the Principle of Parsimony the shift is from higher-order planning (the more general) to instances of talk (the more specific). This model was developed in the context of natural language processing, and was concerned with the planning needed in the belief systems of computer agents. The question here is whether the core concept of Shadbolt's model can explain the decisions taken by real speakers in real discourse contexts, or whether it is too much of an idealisation and simplification.

The data chosen for this study were drawn from the HCRC Map Task Corpus (Anderson et al. 1991a). These dialogues involve the exchanging and negotiating of information in order to complete a relatively complex task. They also produce an output in the form of a route drawn on a map by one of the participants, which is an independent indication of task success. This allowed an investigation of the relationship between speaker strategies and task outcome. The analytical method involved the development of a coding system which categorised talk in terms of the strategies used (or not used), and also in terms of the effort expended.

In the following section, a brief review is given of each principle with an indication of how they will be operationalised. Sections 3 and 4 deal with methodological issues and describe the Typology of Move Attributes, the coding system used to analyse the data. In section 5, a set of testable hypotheses is given for each principle based on the discussion in section 2. The results of the empirical tests are presented in section 6, and this is followed by a discussion of these results and their implications in the final section.

2. Setting the Scene

2.1 The Cooperative Principle and 'Cooperation'

One of the arguments we will be making in this article, and have made more fully elsewhere (e.g. Davies 1998; to appear) is that the type of linguistic behaviour suggested by Grice's Cooperative Principle (Grice 1975) should be carefully distinguished from folklinguistic notions of Cooperation. Therefore, in this study, two different principles will be tested: Gricean Cooperation, which will attempt to generalise and operationalise the Cooperative Principle (hereafter CP) and general Gricean view, and Cooperation, which will take a non-technical interpretation of 'helpfulness' and 'effort'.

To justify this approach, we will first demonstrate the way in which these two have been confused and the potential misinterpretations to which this can lead. The initial problem lies in the fact that (as with many terms in linguistics) there is a restricted technical definition of a word which also has a more general non-technical meaning. Thus, when the word 'cooperation' is used in a linguistic context, it is not always clear which sense is intended, and unless the reader already has knowledge of the technical notion then the more general meaning will be accessed. Indeed, it could be argued that even for the initiated, the more general meaning will be more easily accessible unless the alternative meaning is clearly indicated. The following quotations provide examples of this:

> "[implicatures rely on] some very general expectation of interactional cooperation" Levinson (1983: 50)

> "One of the defining features of conversation is that it is cooperative in nature" Fais (1994: 231-242)

Other examples create more potential confusion by using concepts like 'harmony' and 'effort' alongside the term 'cooperation', implying a link between the two.

> "Grice's theory rests on the assumption that people are intrinsically cooperative and aim to be as informative as possible in communication, with informativeness referring to a maximally efficient information transfer." Eelen (2001: 2)

> "… speakers cooperate … When studying transcripts of genuine conversation one is struck by the general atmosphere of cooperativeness and harmony" Stenström (1994: 1)

Or, the CP is taken to mean the avoidance of miscommunication – the idea being that it exhorts us to provide perfect information for our addressees. This is also seen to some extent in the quotation from Eelen (2001) above.

> "Grice's principle assumes that people cooperate in the process of communication in order to reduce misunderstanding." Finch (2000: 159)

This assumption of effort and perfection is then taken by some as a reason to reject the Gricean notion:

> "It seems to us to be matter of common experience that the degree of cooperation described by Grice is not automatically expected of communicators. People who don't give us all the information we wish they would, and don't answer our questions as well as they could are no doubt much to blame, but not for violating principles of communication." Sperber & Wilson (1986: 162)

While it is far from being clear what exactly is meant by the CP – an issue which we will examine shortly – it is relatively straightforward to reject concepts like 'helpfulness', 'harmony' and extreme degrees of effort. The CP explains how addressees treat potentially meaningless utterances – talk which would otherwise not make sense. The insights which the CP offers refer to utterances where the speaker has produced talk which requires more effort on the part of the addressee to interpret: this is scarcely 'helpful' behaviour or effortful on the part of the speaker. And, indeed, implicatures can be generated and interpreted by speakers and addressees who are feeling less than 'harmonious' towards one another.

If the CP is examined in the context of Grice's other writings, then the overall framework within which he was operating soon becomes clear. The underlying motivation behind his view of philosophy as a whole was the assumption that rational action was at the core of all human behaviour (e.g. Grice 1986). The notion of cooperation does not resurface in any of Grice's other work, whereas rationality recurs on a regular basis (e.g. Grice 1989), and indeed, it occurs in Grice (1975):

> "… one of my avowed aims is to see talking as a special case or variety of purposive, indeed rational behaviour" Grice (1975: 45)

> "A dull, but no doubt at a certain level, adequate answer is that it is just a well-recognised empirical fact that people DO behave in these ways … . I am, however, enough of a rationalist to want to find a basis that underlies these facts, undeniable though they may be; I would like to be able to think of the standard type of conversational practice not merely as something that all or most do IN FACT follow but as something that it is REASONABLE for us to follow, that we SHOULD NOT abandon." Grice (1975: 48, original emphasis)

The problem then becomes what Grice meant by the term 'rationality'. This is not clear in his writings, and as Markie (2000: 740) says, the "application of the term 'rationalist' can say very little about what two philosophers have in common". In this study, we have taken a broad view of what this might entail, but have taken into account the moral requirement to complete the task appropriately, notions of efficiency, and the idea that the application of reason should allow for learning.[1]

This is contrasted to the definition we take for the folklinguistic notion of cooperation, which can be viewed in terms of a desire to be 'helpful' to one's co-participants. Brown (1995: 16) terms this in the following way:

> "… a system which requires effort on the part of the speaker in constructing a helpful message and also on the part of the hearer in working out what the speaker might have meant."

So, helpfulness here is linked to a notion of 'effort' on the part of both speaker and hearer. Therefore, operationalising this notion of cooperation will take the measurement of effort as its starting point, but will also take into account what other types of behaviour could be considered helpful in this context.

[1] The philosophical background to this issue is discussed in detail in Davies (to appear), and the implications for operationalising the CP are discussed further in Davies (1998).

2.2 The Collaborative Theory and the Principle of Least Collaborative Effort

The Collaborative Theory has been developed through the work of Herb Clark and his co-workers (e.g. Brennan & Clark 1996; Clark & Brennan 1991; Clark & Krych 2004; Clark & Schaefer 1987a,b; Clark & Wilkes-Gibbs 1986; Schober & Clark 1989; Schober 1995; Wilkes-Gibbs 1986); its central tenet is that language is a joint production, and is not reducible to the contributions of two individuals. Essentially, the sum of a conversation adds up to more than the sum of its parts. This theory is based on psycholinguistic evidence from a number of experimental papers (see Clark 1992 for a key collection), and its full exegesis is given in Clark (1996). The following is a brief account of the theory where we concentrate on the aspects of most importance to the work reported here. A fuller discussion can be found in Davies (1998, in prep).

According to the Collaborative Theory, each utterance should be considered as a presentation and needs to be accepted by the addressee before it can be deemed to be added to the speakers' common ground. Ratified participation in this process is essential – only the understanding of full participants is taken into account in this process; overhearers, etc. do not share the same degree of common ground because although they may have access to the same set of utterances, it is not their understanding which is being monitored by the process (Clark & Schaefer 1987b; Schober & Clark 1989; Wilkes-Gibbs & Clark 1992). It is this process of building common ground which is termed 'collaboration'.

As common ground is increased through the process of collaboration, speakers can be less explicit – that is, say less – when engaged in tasks because a certain level of shared knowledge is being assumed. In the tangram matching tests used by Clark and his co-workers, this was shown by the interactants using shorter referring expressions and taking fewer turns to complete the task. This decrease in the words and turns used was seen by Clark et al. as a decrease in effort: speakers said less because they saw the opportunity to conserve effort. However, as collaboration is a joint effort Clark et al. also argue that the minimisation of effort is a joint activity too, hence the *Principle of Least Collaborative Effort*.

This is demonstrated in the conversations about tangrams by speakers *refashioning* referring expressions, and gradually coming to an agreement about a referring expression rather than one individual investing a lot of effort to produce a perfect referring expression in one utterance. Therefore, Clark argues, the work is divided and minimised:

A: Um, third one is the guy reading with, holding his book to the left.
B: Okay, kind of standing up?
A: Yeah.
B: Okay.

Clark & Wilkes-Gibbs (1986: 22)

In Schober's (1995) experimental work, a similar pattern was found. He set up a task where the perspective of the Director and the Matcher could be different: the experimental materials could have the same perspective, or could be offset 90^o or 180^o from the Director's. Over a series of trials, he found that Directors tended to move from Matcher-centred to perspective-neutral descriptions. Thus it was argued that this represented least collaborative effort because Matcher-centred or Director-centred descriptions would maximise work for one participant and minimise it for another, whereas neutral descriptions minimised the work for both.[2]

These results can be linked to the Principle of Mutual Responsibility:

> "The participants try to establish, roughly by the beginning of the next contribution to their discourse, the mutual belief that they have understood what the contributor meant, to a criterion sufficient for their current purposes." Clark & Wilkes-Gibbs (1986: 33)

Joint effort can also be minimised by participants deciding the extent to which something needs to be understood: the necessity of understanding small talk at a party is rather less than the need to understand the instructions for your driving test. This is what is meant by *"to a criterion sufficient for current purposes"*. Wilkes-Gibbs (1986, 1997) investigated this by setting up an experimental task where participants were given high or low criterion goals. In each case, the basic task was the same: the participants were given a map of the same city centre, with half the squares blocked out. Between them, they had to work out the route between two given points. The difference was in the instructions given to the high criterion (HC) and low criterion (LC) participants. HC participants were told that they should make sure they knew the route well enough to describe the route from A to B to someone intended to drive between those two points, whereas LC participants only needed to estimate how long the same route would take to drive at 1pm. All participants were then asked to take the same test: each individual had to replicate the route on a full version of the map.

The experiment was set up with three conditions: HC pairs, LC pairs and mixed pairs. The HC pairs talked for significantly longer than either the LC pairs or the mixed pairs, which would indicate more collaboration and thus greater effort on the part of the matched HC speakers. This distinction in the effort invested would seem to demonstrate an orientation to some notion of 'a criterion sufficient for current purposes', and a minimisation of effort where it is perceived to be possible.

However, this differentiation in levels of effort did not appear to change the task result: there was no significant difference in task performance (i.e. accuracy of the map) between the HC and LC pairs, but the mixed pairs did significantly worse than the matched pairs – the HC participants in mixed pairs did particularly badly. This was taken by Wilkes-Gibbs to indicate that more effort, in itself, would not improve task result: being paired with someone who had an equal level of commitment was seen as being more important. Wilkes-Gibbs explains the similarity in effort levels between the LC pairs and the mixed pairs in terms of the HC participants being more flexible, and thus willing to accommodate to their partner's needs.

It should be noted at this point that we consider there to be problems with the explanation offered for both this and the refashioning and perspective examples above. The 'flexibility' of the HC participants in the mixed pairs wouldn't appear to account for the poor performance of those pairs, particularly in terms of those needs-sensitive HC participants. For the previous examples, it would seem to be true that there is a shift

[2] Matcher-centred descriptions maximise work for the Director, because they have to take into account the altered perspective of the Matcher, but minimise the work for the Matcher. Director-centred descriptions minimise the work for the Director, but maximise the work for the Matcher, because they now have to process the effect of the difference in perspective.

in the way the work is done: from ‘perfect’ referring expressions to refashioning; from offset-specific descriptions to perspective neutral ones.[3] However, what is not clear is that there is *necessarily* a decrease in overall collaborative effort, as we have no concrete way of measuring this. In the next section we will consider an alternative explanation: a Principle of *Least Individual Effort*, which we will argue can account more effectively for these examples.

The notions which are taken forward into the hypotheses to represent collaboration – and more particularly least collaborative effort – are concerned with effort and its relationship with task success. According to the Clarkian view, we would expect to see a reduction in effort as the speakers gain familiarity with the task (*c.f.* Clark & Wilkes-Gibbs 1986), and no relationship between absolute effort and task success (*c.f.* Wilkes-Gibbs 1986).

2.3 The Principle of Parsimony, The Risk-Effort Trade-Off and the Principle of Least Individual Effort

When we engage in a dialogue, we reason about how we intend to proceed. Usually, we will have more than one option open to us, which will take more or less effort for us to formulate, and, conversely, imply more or less risk that the intended goal will be achieved first time. Shadbolt (1984: 342) argues that the decision we make is based on the Principle of Parsimony:

> “… a behavioural principle which instructs processors to do no more processing than is necessary to achieve a goal.”

In other words, an interactant will try to choose the approach which will be the least effortful – and thus the most risky – that is still likely to succeed.

To use an example from the type of dialogues with which we are concerned here, the difference between a high risk posture and a low risk posture can be illustrated by the way in which a speaker deals with a new feature in the context of a route instruction:

Low risk:
Do you have a burnt cottage?
Go to the left past the burnt cottage.

High risk:
Go to the left past the burnt cottage.

The low risk approach takes more effort initially, but it checks a precondition for the planned instruction. Therefore, it is more likely to succeed first time. The high risk approach makes the assumption that the location of a particular landmark is shared knowledge. This strategy is lower effort, but takes the risk that a potentially effortful repair sequence will have to be entered into. The trade-off here is the opportunity to save some effort (introducing the feature) against the possibility of having to engage in a potentially more effortful repair sequence. The risk-effort trade-off, then, is the

[3] Although see Davies (1998, in prep) on the question of whether refashioning does count as a change in effort.

judgement that the speaker makes in terms of the likelihood of a particular risk being worthwhile.

This concept was developed through research in computational linguistics, and it concentrates on high level strategies used by a computer agent in reasoning about beliefs such as *degree of assumed shared knowledge about the task, degree of specificity used in referring to an object in the task, degree to which you assume your addressee shares your area of focus, and degree to which you provide feedback to your discourse partner*, rather than the instantiation of those strategies at the level of the move. The overall model was extended by Carletta (1992, Carletta & Mellish 1996), who developed a computational system which generated a simple conversation between two participants in the Map Task domain. This concentrated on a similar set of strategies to Shadbolt, but was engaged in producing moves not just high level planning. Therefore, the work described here is a first attempt to shift these concepts from a limited domain to the analysis of real talk.

There were two immediate problems in relation to how the notion of risk was quantified and conceptualised in both their studies which we needed to address. Firstly, both their systems were limited to only a binary categorisation for risk (high or low), whereas any purported relationship between effort and risk would in reality be on a continuous scale rather than being discrete categories. In our study, measures of both risk and effort are on continuous scales.[4] Secondly, an important – and problematic – assumption in relation to the conceptualisation of risk is made. In both systems, risks may cause dialogue problems and misunderstandings, and lead to extra effort, but these problems are always recognised by the agents and successful remedial action is taken. In real human dialogue, such misunderstandings are not always noticed by the interactants. Even if they are, the interactants may be unable – or unwilling – to address them. So, the definition of risk that we use is slightly different to that used by Shadbolt or Carletta:

> **Risk:** When a risk is taken, the speaker takes a chance that the communication may fail. This miscommunication may, or may not be, resolved.

We also wish to link the notions of Parsimony and the risk-effort trade-off to a suggested Principle of Least Individual Effort, in contrast to the Principle of Least Collaborative Effort suggested by Clark. This is because we interpret the message of the Principle of Parsimony as an exhortation to individuals: it is individual speakers who decide how to balance effort and risk against each other, rather than pairs or groups of participants. It also makes a clear differentiation between the prediction of Parsimony/Least Individual Effort and Collaboration/Least Collaborative Effort.

We have already suggested above that we believe the evidence offered in support of Least Collaborative Effort by Clark et al. could equally be argued to be support for Least Individual Effort. For example, whilst we have no way of measuring the overall effort involved in the refashioning exchanges reported in Clark & Wilkes-Gibbs (1986), nor calculating the difference in processing effort for both interactants shifting from

[4] Although we do categorise effort for a particular strategy into discrete levels, the overall judgement for a dialogue in on a continuous scale.

offset-specific descriptions to perspective neutral ones in Schober (1995), we can see what is happening to individuals' contributions. Directors in the tangram task avoid the work of producing 'perfect' referring expressions, and Directors in Schober's task avoid the effort of thinking in terms of a different perspective and use a perspective to which both speakers can relate. In both cases, the Directors (who arguably have more say in the way in which the task is approached) choose strategies which would seem to minimise their personal effort, at the expense of the effort of their co-participant. With respect to the importance of equal commitment to task success, it is also relatively easy to point to the alacrity with which HC participants seem to abandon their high effort approach in contrast to LC participants' apparent commitment to retaining their low effort one.

In operationalising these ideas, the predictions will be concerned with two main areas. Firstly, it will consider the risk-taking behaviour of the interactants, and whether this has any relationship with task success. And secondly, it will also take up an investigation of effort – and in particular the relative importance of joint and individual effort. This will highlight the distinction between the concepts of Least Collaborative Effort and Least Individual Effort.

3. Methodology

In this section we will outline the important features of the HCRC Map Task Corpus, and describe the coding and analysis which was undertaken in this research.

3.1 The Map Task Corpus

The data used in this study is part of the HCRC Map Task Corpus (Anderson et al. 1991a) which consists of 128 task-oriented dialogues collected from 64 speakers. These participants were divided into groups of four, a 'quad'. Each person was involved in four dialogues in their quad, and each quad generated eight dialogues in total. The task they undertake involves one speaker (the Instruction Giver) describing the route on their map to the other (the Instruction Follower), who has a slightly different map. Each person is a Giver twice and a Follower twice; they 'give' the same route twice (to different Followers), but are Followers on different maps each time. In this analysis, we used four quads (32 dialogues, 16 speakers) which amounted to approximately four hours of speech.

3.1.1 The Task

The maps show the same fictional location, but they are not identical. The route (which only the Giver has) is based around a number of small named pictures (known as features or landmarks), but not all of these are on both maps: about eight out of eleven are shared. As all these features are important to the route, the interactants must engage in information exchange if they are to complete the task successfully. The instructions given to the speakers informed them that their partner had a map drawn by another explorer which might, therefore, be different. They were also told that the route drawn on the Giver's map was the only 'safe' one, and that they should try to ensure the route which the Follower drew was as accurate as possible. These instructions were intended to suggest that there may be some differences between the maps (although not the type nor the extent of difference), and also to encourage the participants to become involved in the negotiation process necessary for an accurate route to be drawn.

3.1.2 Task-Oriented Conversations

While the experimental situation used produces data which is not strictly naturally-occurring, it in fact meets our requirements quite well. One of the ongoing issues for researchers interested in the analysis of conversational data is the problem of the Observer's Paradox: how can one record natural data when ethical and legal requirements demand that the subjects know that they are being observed? Labov's answer to this was to engage the subject by getting them to talk about a near-death situation. Interestingly, a task such as this seems to work equally well. Participants rapidly became absorbed by the cognitive demands of the task and overcame their initial nervousness very quickly. Audio recordings show the data to be very natural, and some participants commented how rapidly they had forgotten about the alien environment. Of course, the talk genre they produced is of a primarily transactional rather than an interactional type. But then, we are interested in how participants manage, transfer and negotiate information. There are interactional aspects to this – not answering questions or directly refusing your partner's suggestions would probably lead to partial or total breakdown in the conversation – and such elements are taken into account in our analysis.

If one were interested in primarily interpersonal aspects of talk, then the use of task-oriented data would probably be inappropriate: data collection must be fit for purpose. And whilst the analysis of casual conversation is often seen as a 'gold standard' in pragmatics, there is an increasing interest in both other genres and other methodologies. For example, talk in the workplace is seen as an increasingly important site for analysis (e.g. Bargiela-Chiappini & Harris 1997; Connor & Upton 2004), and methodologies such as Discourse Completion Tasks (DCTs) are seen as a legitimate tool for data collection in work on politeness (though see Beebe & Cummings 1996 for a discussion of the limitations of this approach). Arguably, business talk is a type of task-oriented dialogue – it is often concerned with the transferral and negotiation of information. DCTs are essentially a very constrained role play: the set up used for data collection here is also a role play, but without the constraints and lack of discourse context for which DCTs have been criticised. Thus we would argue that task-oriented data is of legitimate interest to linguists, provided that their aims fit with the constraints of the data.

3.1.3 Advantages of the Data

From our point of view, it is this transactional nature of the corpus which makes our approach possible. To analyse the choices interactants make, we need to know (as far as possible) their state of knowledge and their likely goals at a given point. Such a degree of insight is rarely possible when observing casual conversation. In addition, the communication channels are constrained by the experimental environment. Although the interactants could see each others' faces, the board between the speakers effectively barred accidental non-verbal gestures (participants were asked not to use gestures) meaning that information content had to be carried largely by the verbal channel.

A further advantage of the Map Task dialogues is the existence of the route drawn by the Follower on their map: this is an independent indication of task success. We used a metric (Incorrect Entity Score) based on the Follower's success in navigating the features correctly. This was weighted according to both the relative difficulty of the feature (whether it was shared, only on the Follower's map, only on the Giver's map), and the degree of error ('good miss', 'bad miss'). The IE score produced a number between 0 – 22, with larger numbers representing more – and more serious – errors. We developed this metric because we see the landmarks as pivotal to the route, whereas absolute accuracy is not.[5]

However, the main advantage of this corpus of task-oriented dialogues over many others (e.g. Grosz and Sidner, 1986; Clark & Wilkes-Gibbs 1986; Clark & Schaefer 1987a,b; Schober & Clark 1989; Clark & Brennan 1991) is that there is no one participant who has all the necessary information: there is not an expert or a novice (Wilkes-Gibbs 1986 is an exception to this – although even her participants alternate these roles). It is not sufficient for the Giver to *describe* the route to the Follower: without the knowledge of the unshared features (on either map), it is likely that the route drawn by the Follower will negotiate some aspects of the route incorrectly. This means the dialogue is more of an equal enterprise. The input of the Giver and Follower is equally important, and the responsibility for a good task result is joint, rather than being the main responsibility of the Giver. We would argue that this makes it an ideal corpus on which to base an investigation into dialogue principles.

3.2 Approach to Data

When the concept of a dialogue coding system is introduced, most people assume that its concern is the identification and labelling of overall dialogue structure (e.g. Carletta et al. 1997; Houghton & Isard 1987; Kowtko et al. 1992; Sinclair & Coulthard 1975) or of structures within a dialogue (e.g. Conversation Analysis) rather than a scheme which attempts to identify the presence *or absence* of certain types of discourse strategies. The Typology of Move Attributes (hereafter Typology) is different because it attempts to code instances where an interactant has engaged in specific subtypes of moves (e.g. clarificatory questions, 'new' questions or clashes in information status; short replies or full replies), but also where an interactant is judged to have failed to engage in a particular strategy where it would have been appropriate. In other words, we are coding what is 'not there' as well as what is.

This may be seen as problematic on two fronts. Firstly, it is unashamedly evaluative, and this goes against what is often seen as a basic tenet of linguistics: that we are to *des*cribe not *pres*cribe – and evaluation could be seen as a type of prescription. However, this can also lead us to what Coupland, Giles & Wiemann (1991) term the 'Pollyanna Principle' – that linguists always describe how good humans are at language, without also explicitly discussing that they also (on occasion) fail to interpret correctly, say things others can't understand and make poor linguistic choices. It is clear in our data that some of the dialogues produce more successful task outcomes (more accurate routes) than others, and therefore as the medium used is language, it is inevitable that some dialogues will be labelled as 'more effective' than others. We would also argue that we are not judging language in a way which should worry a descriptive linguist – the linguistic strategies which we evaluate are essentially realisations of higher order planning, and are not concerned with issues of 'standards', i.e. grammaticality, lexical choice or register.

[5] For further discussion of this, and a description of empirical work which tests the validity of this metric and the deviation score approach used by Anderson & Boyle (1994), please see Davies (1998).

Secondly, we are assuming that ‘we know’ what an interactant should be doing at a particular point in a dialogue. This is obviously far more problematic, but is aided by the type of dialogue: we know (to a reasonable extent) the overall goals of the interactants and their state of knowledge with respect to these goals. Also, our coding system was very much data-driven and did not start with any preconceived categories.

The Typology was based on analysing one quad (eight dialogues), where we identified ‘problem points’: points at which the addressee was forced to clarify, question or object in some way to an utterance made by the speaker. Then the aspects of the previous utterances which caused the need for the clarificatory behaviour were identified. These were then categorised into groups, and formed the basis for the Typology. So the core of the Typology was driven by the identification of behaviours which caused problems in particular contexts.

This process generated the set of attributes which could be considered to be ‘absent’ (i.e. negatively coded) and also the core of those which could be considered to be ‘present’ (i.e. positively coded). Several other strategies were added to the ‘positive’ list, such as suggestions or new questions used by the Follower. These were considered to be valuable contributions but would not normally be seen as ‘necessary’ and thus rarely defined by their absence.

Thirdly, we are assuming that others would *agree with* our assessment of what others should be doing at a particular point in a dialogue even assuming we agree on the state of knowledge and so forth. This essentially demands that the coding scheme must be reliable: well-defined, rigorous and usable by other coders (Carletta 1996; Isard & Carletta 1995). To this end, we undertook a small reliability study whose results are reported in Davies (1998).[6]

4. The Typology of Move Attribute Types

The approach we have taken is not based on one model. The basic unit that we code is ‘the move’ and this is drawn from Birmingham School Discourse Analysis; we retain the tripartite distinction IRF from here, although we do not employ the notions of ‘exchange’ or ‘act’. Arguably, the notion of choice can also be linked to the IRF model’s roots in systemic functional linguistics, and we have explored this in a more formal way in the context of the COMMUNAL project (Fawcett & Davies 1992; Lin et al. 1993). However, not all our categories are move-specific, or relate to a notion of ‘move structure’: we take the idea from Conversation Analysis that a method of dialogue analysis does not have to be only about structure, and that the analyst can draw attention to whatever aspects of a text they believe to be relevant. In addition, we have also used the idea that utterances are interpreted against what went before, and that certain utterance types are expected and preferred in certain discourse situations (*c.f.* conditional relevance, adjacency pairs and preference organisation). Our debt to Dialogue Games (Houghton & Isard 1987; Power 1979) is more conceptual than structural – it too places emphasis on the intentions and goals of the speaker which sits well with the higher order principles we are interested in investigating. The idea of evaluating utterance choices in relation to task success can be traced back to work by

[6] The majority of the move attributes were shown to be reliable, and those that did not quite reach the accepted statistical level were often affected by the small number of datapoints in that category.

Anderson and Boyle (Anderson et al. 1991b; Anderson & Boyle 1994; Anderson et. al. 1994) who found individual linguistic choices like the use of feature introductions and the quality of response moves in Map Task data could be linked to the accuracy of the route. The extension undertaken by this project is to broaden the scope of evaluation from a narrow set of indicators to a wider, data-driven set.

4.1 Encoding Effort

As we have indicated earlier in the discussion of dialogue principles, measuring speaker effort is not a straightforward task. In this study we equate 'effort' with the perceived amount of work involved in planning and producing an utterance, and have identified four different levels of effort:

1. The social needs of the dialogue [minimum effort]
2. The responsibility of supplying the needs of your partner [moderate effort]
3. The responsibility of maintaining correct mutual beliefs [medium effort]
4. The responsibility of initiating new subtasks [high effort]

The social needs of the dialogue

This is the minimum you need to do to keep the conversation going, and includes such move types as minimal responses and acknowledgements. These brief utterances are classed as low effort because they do not require much planning, much examination of joint beliefs, nor much consideration of the contribution of the utterance to the overall dialogue and joint task.

The responsibility of supplying the needs of your partner

Beyond the social needs required to simply keep an utterance going, one can provide Responses and Follow-ups which take more consideration of your partner's intentions and goals in formulating that particular utterance. This involves more effort because it involves (typically) longer utterances, and more processing of joint beliefs. But it is still largely prompted by the actions of the other speaker.

The responsibility of maintaining correct mutual beliefs

This level of effort refers to the work involved in both querying the assumptions of your partner (in respect of mutual knowledge) and trying to ensure that your assumptions of mutual knowledge are well-founded. We make the assumption that going against the predicted move in an exchange will require more work than simply producing the expected move-type. This is because, in the case of false assumptions, the speaker must have undertaken a certain amount of work (e.g. reasoning about beliefs) to decide that such a query is necessary.

The responsibility if initiating new subtasks

The previous levels of effort all consider the actions of a speaker within the context of a particular subgoal: that is, they mainly deal with situations where a speaker is reacting to the instruction or question offered by the other participant, rather than moving the discourse on to the next subgoal. This we perceive to be greater effort because it involves reasoning about the task as a whole, as well as planning and producing a particular utterance.

4.1.1 Using Effort Levels

In using the Typology to code dialogues, we will refer to positive codings (i.e. finding an instance of the behaviour in an utterance), and negative codings (finding an instance where we believe a particular behaviour should have been used). Rather than simply making a tally of these codings, we used the effort levels described above to weight the incidence or absence of particular behaviours, as shown in Table 1:

Effort Level (Least first)	**Positive Weighting**	**Negative Weighting**
Level 1 – Minimum Effort	+1	-4
Level 2 – Moderate Effort	+2	-3
Level 3 – Medium Effort	+3	-2
Level 4 – High Effort	+4	-1

Table 1. Effort levels and weightings in the Typology.

In terms of Traum's (1994) discourse obligations, speakers are more obligated to engage in lower effort behaviours than higher effort ones: if you do not at least respond, the conversation will end, but you can choose whether or not to query an instruction or offer a suggestion about what to do next. This is reflected in the weighting system where behaviours with a high discourse obligation have a low positive weighting and a high negative weighting, and vice versa.

Using this system provides a positive and negative score (sum of codings) for a dialogue. This represents a principled attempt to account for the effort invested and provides a basis for the empirical testing of dialogue principles.

4.2 A Summary of Move Attribute Types

A separate table is given for positive and negative codings. Each table is divided into those attribute types specific to particular moves and those which can be applied to any type of move. Each attribute type is also categorised in terms of the effort levels given above.

SUMMARY OF POSITIVE CODINGS		
INSTRUCT		**Positive Weighting**
+NEW-QUESTION	Asks question *not directly prompted* by previous utterance	+4
+RELEVANT-INFO	Introduces new, unsolicited information ('new' in terms of focus, potentially relevant to route section)	+4
+NEW-SUGGESTION	Makes unsolicited suggestion about where route might go nest (need not be *correct*)	+4
+QUERY	Question (function not form) prompted by previous utterance either because of information problem or checking *self* understanding (check if +KNOWLEDGE-MISMATCH is appropriate)	+3

+OBJECTION	Statement (function not form) prompted by previous utterance, concerned with information problem (check if +KNOWLEDGE-MISMATCH is appropriate)	+3
+CHECK	Question which solicits *other* understanding of information already offered	+2
RESPONSE		
+REPLY-MIN –REPLY-FULL	Insufficient or inappropriate information	+1 & -3
+REPLY-YN	Yes-No reply to Yes-No question	+1
+REPLY-FULL	Reply to WH-question, or full reply to Yes-No question	+2
(+INFO-INTEG)	Additional information offered (Move should be coded as REPLY-FULL) [RARE]	+4
FOLLOW-UP		
+ACK-SHORT	Appropriately brief follow-up	+1
+ACK-FULL	Full follow-up	+2
(+INFO-INTEG)	Additional information offered (Move should be coded as ACK-FULL) [RARE]	+4
FEATURE-SPECIFIC CODINGS		
+FEATURE-INTRO	Highlighted (re-)introduction of a feature	+2
+FEATURE-LOC	Attempt to locate position of feature	+3
+FEATURE-UNIQUE	Attempt to uniquely identify feature (e.g. in terms of location)	+3
HIGHER-LEVEL CODINGS		
+KNOWLEDGE-MISMATCH	Move points out mistaken assumption (should be move-coded as +QUERY or +OBJECTION)	+3

Table 2. A Summary of Positive Codings.

SUMMARY OF NEGATIVE CODINGS		
INSTRUCT		**Negative Weighting**
-NEW-QUESTION	Not applicable	N/A
-RELEVANT-INFO	Failure to introduce useful knowledge when necessary	-1
-NEW-SUGGESTION	Failure to make a suggestion (This behaviour is potentially helpful rather than necessary, and therefore failure is rare)	-1
-QUERY	Failure to indicate information problem.	-2
-OBJECTION	Not applicable: defined on difference in function which can only be identified if strategy is realised – use -QUERY	N/A
-CHECK	Failure to check other's understanding of information offered (mainly at topic/segment boundaries)	-3

RESPONSE		
–REPLY-FULL	No response given when required	-3
+REPLY-MIN –REPLY-FULL	Reply too short, or inappropriate	+1 & -3
(-INFO-INTEG)	More information necessary [RARE]	-1
FOLLOW-UP		
-ACK-SHORT	No follow-up given when necessary	-4
-ACK-FULL	Inappropriately brief follow-up. (can occur with +ACK-SHORT)	-3
(-INFO-INTEG)	More information necessary [RARE]	-1
FEATURE-SPECIFIC CODINGS		
-FEATURE-INTRO	New feature introduced, but not highlighted (i.e. treated as shared information)	-3
-FEATURE-LOC	Failure to start negotiation process for unshared (typically) feature	-2
HIGHER-LEVEL CODINGS		
-KNOWLEDGE-MISMATCH	Move fails to point out mistaken assumption (should be move-coded as -QUERY)	N/A

Table 3. A Summary of Negative Codings.

Coding the data in this way provided a resource of information about the corpus. Various profiles of each dialogue could be drawn from the coding results, looking at such issues as the relationship between negative score and task success, or effort invested and task success. A range of nonparametric tests were used to investigate these and other relationships, as explained in section 6.

4.3 An Example of Coding

While it is impossible to give examples of all the possible coding types (see Davies 1998 for a full description), the following extract should give an indication of how the coding system works. In particular, it demonstrates the use of negative coding alongside positive coding. This particular excerpt[7] is taken from about the middle of a conversation. It's the first time these two participants have had these particular roles, but it will be the second Map Task conversation in which they have been involved.

G59: Then you go up vertically, til just above the alpine garden
Then you mak… go right to the end of the alpine garden
#G -CHECK
Then vertically again to the top of the youth hostel.
#G -FEATURE-INTRO[youth hostel]

[7] This example is taken from conversation EAQ4C4. Turn numbers are as in original; the transcription has been slightly simplified in terms of the representation of pauses, hesitation phenomena and false starts, etc. No route segment coding is used. G and F are abbreviations for Giver and Follower, respectively.

F60: There is no youth hostel.
#F +OBJECTION

G61: Then about three inches
#G -REPLY-MIN

F62: Where was the youth hostel?
#F +NEW-QUESTION

G63: The youth hostel was … upwards and to the right of the alpine garden, if you go up from four inches from where you stopped at the alpine garden.
#G +REPLY-FULL

F64: Right
#F +ACK-SHORT

G65: Vertically
#G [continuation of REPLY-FULL]

F66: Right, I think I know where it is
#F +ACK-FULL
right. Ehm, where would I be going after that?
#F +NEW-QUESTION

G67: After where?
#G +QUERY

F68: After the, eh, this youth hostel.
#F +REPLY-FULL

G68: After the youth hostel
#G +ACK-FULL
you'd be going an inch to your right and then a diagonal line sloping at approximately, eh, what'd it be now, hundred and thirty five degrees from vertical.
#G +REPLY-FULL

F69: Right.
#F +ACK-SHORT

This piece of talk starts with the Giver trying to explain a large chunk of the route, including one new feature (*youth hostel*) to the Follower. This is not very successful, and the remainder of the extract shows how the issues caused by this first utterance are resolved.

In terms of negative coding, we see a lack of a CHECK and a FEATURE INTRODUCTION in G59 – it was risky for the Giver not to check the Follower's understanding of this first part of route description (to the *alpine garden*) before moving onto the next section (to the *youth hostel*). This was then compounded by a failure to check whether the Follower had the feature *youth hostel* on their map. The final negative coding in this extract is also used on an utterance made by the Giver. The OBJECTION made by the Follower in F60 demands explicit attention, but the Giver's utterance does not provide an obvious answer – it appears to be a further instruction – thus it is coded as a failure to provide a response (-REPLY-MIN).

For positive coding, there are some examples of acknowledges and replies, but the ones relating to initiate moves are probably more interesting. We see an OBJECTION in F60, a QUERY in G67 (because it tests self-understanding), and two NEW-QUESTIONS in F62 and F66. While the latter of these is probably unproblematic, the coding of F62 as NEW-QUESTION rather than QUERY may need further explanation. QUERIES and OBJECTIONS must relate directly to the previous utterance – that is, they must obey conditional relevance (Schegloff 1968) – in this case, because the OBJECTION has been ignored, F62 is not conditionally relevant on G61, and represents a new attempt to solve the problem.

It should be pointed out that negative coding only represents the taking of a risk, and does not entail (or represent) a mistake being made by the Follower in the drawing of their route. Indeed, in this particular case, the efforts of the Follower resolved the problem successfully and no error was made.

5. Operationalising the Principles

In order to use the data generated by the dialogue coding, a practical method of testing the various dialogue principles had to be developed. This involved defining a set of hypotheses for each principle which then could be empirically tested. As such a translation of the principles is far from straightforward, we justify each of the decisions we have taken in some detail.

5.1 Gricean Cooperation

Grice's work comes from a philosophical tradition based on intuition and reflection, and was never explored by its originator in an empirical framework. The suggestions made for this Principle are thus those probably most open to criticism and disagreement out of the four discussed. While we accept these limitations on the analysis and the interpretation of the following results, our intention is to demonstrate the clear distinction between Gricean Cooperation and the folklinguistic notion of Cooperation (argued for in section 2.1) rather than produce an inarguable interpretation of Grice.

We take as our analysis the possible interpretations of dialogue as rational activity or dialogue as efficient activity, as we have justified above. Indeed, although it may be argued that Grice's primary focus was on rationality, it could equally be argued that efficiency may be an instantiation of rational action.

1. Speakers will avoid unnecessary effort

Although speakers should have a moral commitment to doing the work necessary to the task, they are not expected to do any *more* than that. The CP makes it possible for speakers to decrease their effort, and thus meet this ideal.

2. Speakers will improve at tasks

Agents should have the ability to *learn*. This could be seen as the application of reason [i.e. rationality] to a particular problem set. In terms of task-oriented dialogues, we would expect agents to produce better task results over time. This can also be linked to notions of efficiency: the agent learns the minimum that is required to do the task.

3. Speaker effort will decrease

This hypothesis is linked to the previous two. As speakers learn, they will determine what effort is absolutely necessary to the task, and what is extra. They can then adjust their behaviour accordingly. Therefore, they can minimise their effort for the task.

5.2 Collaboration and the Principle of Least Collaborative Effort

These two are considered together because firstly, the Principle of Collaboration makes no useful prediction in itself, and secondly, because the Principle of Least Collaborative Effort is entirely reliant on the concept of Collaboration.

1. Speakers will collaborate

This is the only prediction made by the Principle of Collaboration: that both speakers will be involved in the task.

2. Dialogues will get shorter the more times the participants do the task

Participants build up common ground about the nature of the task. Therefore, they need less grounding and thus less Collaboration. The measure in this case will be that used by Wilkes-Gibbs (1986) – the number of turns taken to complete the task – rather than the measure used by Schober (1995), as we have no equivalent to his ‘offset’ conditions.

3. There will be a decrease in average effort for later dialogues

Because the Map Task is a repeat of form rather than content and therefore does not exhibit such a marked increase in common ground, it was considered that a measure of *average* effort was more appropriate than a measure of *absolute* effort. Therefore this could be seen as an *alternative* hypothesis to the one above (H2).

4a Speakers with equal commitment (whether high or low) should be associated with more task success

According to Wilkes-Gibbs (1986), language is a joint product dependent on the criteria (high or low) of both speakers. If speakers do not have matched criteria, then the input of the speakers will default to the lower criterion: thus lowest Collaborative Effort. These mismatched pairs tend to do worse in Wilkes-Gibbs’ experiments. For the purposes of this study, high criterion will be considered to be equal to high effort and low criterion to low effort.

4b There is no relationship between increased collaboration and task success

If speakers collaborate more, there is no guarantee that they will gain a better task result (Wilkes-Gibbs 1986). Wilkes-Gibbs gives three indicators of high collaboration:

1. High numbers of turns
2. High numbers of words
3. Low mean number of words per turn

Statistically speaking, this is difficult to test because this statement requires the acceptance of the null hypothesis ('There is no significant difference'), and such a finding has no true statistical status. Therefore, should the null hypothesis be accepted, this can only be seen as weak, contributory support.

5.3 Principle of Parsimony / Risk-Effort Trade-Off / Principle of Least Individual Effort

For this Principle, we make the assumption that speakers would be trying to find the most efficient point in the risk-effort trade-off.

1. Risks would be taken – some failures

If speakers are trying to work out where the best trade-off occurs, then they are bound to take risks which do not pay off. Otherwise they would never find out which actions are necessary and which aren't. Such risky behaviour is bound to lead to some task failures.

2. Risks would decrease over time – fewer failures

As speakers work out what behaviour is acceptably risky, and what behaviour isn't, then we would expect the *bad risks* to decrease, thus decreasing the failure rate.

3. Task success would improve as speakers negotiate trade-off more successfully over time

The previous hypothesis leads directly to this task-level hypothesis. If speakers work out the best point on the risk-effort trade-off, then their failure rate should also be minimised and they should produce better task results.

This is an equivalent prediction to Grice H2 and Cooperation H5, although the motivation is different.

4. Behaviour would modify as speakers try out different risk-effort combinations, and eventually settle on a set of useful combinations

Speakers should try out various strategies until they find one which satisfies their constraints. This also makes the assumption that the behaviour found later in the task (as participants gain experience) would better represent their 'best-fit' on the risk-effort scale. It should be noted that the risk-effort approach would suggest that speakers are equally likely to start from either a high or low risk posture, and they may adjust down or up respectively. This is in contrast to Grice H1 where the speakers are predicted to decrease risk levels as their experience of the task increases.

The Principle of Least Individual Effort will use the same hypotheses as the Principle of Least Collaborative Effort, as we have argued in sections 2.2 and 2.3 that the former can provide a more convincing explanation for the phenomena described by Clark and his co-workers.

1a Speakers with equal commitment (whether high or low) should be associated with more task success

Where the commitment level of participants is mismatched, the needs of the participants (particularly those with more commitment) will not be met, which will lead to less effective dialogue and thus a poorer task result.

1b There is no relationship between increased collaboration and task success

The need for equal commitment takes precedence over the input of individuals: the effort of one cannot replace the lack of effort by another.

5.4 Cooperation

The definition of Cooperation used here assumes that increasing the effort used means increasing the level of Cooperation, as described above.

1. High effort would be associated with task success

More effort, and thus more Cooperation, should lead to task success. If you invest effort in being helpful, then that effort should pay off. [Note this is reverse prediction to that made by H4b, Collaboration]

2. Low effort would be associated with low task success

Not being helpful – not investing effort – should have a detrimental effect on task success.

3. Few risks would be taken

A Cooperative stance would argue against risk: why risk failure when you are trying to be helpful – it wouldn't be appropriate. Fewer risks should mean better task success. Again, this is the null hypothesis, and thus we can be less certain of the result – unless the experimental hypothesis (that risks are taken – Parsimony H1) is accepted.

4. Expect modification of behaviour

Cooperative participants try to be helpful. With experience, they should learn what approaches are the most helpful, and they should converge on these. There is no explicit prediction of the direction of modification, although one might suggest that more helpful strategies are likely to be more effortful ones, but that is conjecture (similar to hypotheses Parsimony H4, Grice H3).

5. Expect better task success with experience

This is related to the previous hypothesis. As participants work out what is helpful, and orient towards it, then task success should improve. Note, however, that speakers following Cooperation learn in order to improve the task result, whereas those following Parsimony learn in order to decrease effort.

5.5 Summary

It should be seen from this that a number of the Principles make similar (or even identical) predictions, although the motivation behind the prediction is often different or even contradictory. In this case, interpreting the results will rely on seeing how interrelated hypotheses support each other (or not), and thus how the evidence for a Principle overall is supported, rather than relying simply on the *number* of hypotheses apparently supported for each Principle.

6. Experimental Tests and Results

All the statistical tests used on this data were non-parametric: this is because either the data were categorical or the measures were not believed to meet the criterion of interval/ratio data. The following types of test were used:

- Correlation [Spearmans] between task success (based on Incorrect Entity Score) and an alternative independent measure.
- Unrelated test [Wilcoxon Mann Whitney (Siegel & Castellan 1988)] to investigate any significant difference between the first half of each quad [first giving of a map by each Giver] and the second half of each quad [second giving of a map by each Giver]. This was used to see the effect of learning/experience on behaviour.
- Chi Square test

The tests undertaken for each dialogue principle and their results will be described in turn. Although this will require some repetition (because certain tests relate to more than one principle), this will enable the reader to see the overall level of support for a particular principle more easily.

6.1 Gricean Cooperation

1. Speakers will avoid unnecessary effort

This can only be investigated by looking for changes in behaviour, to see what speakers judge as necessary or unnecessary. We would expect to find a movement towards less effort over time, but also accompanied by task improvement (or, at least, no deleterious effect).

Two particular types of attributes were used to investigate this: checking routines [CHECK] and checking shared knowledge of new landmarks introduced into the conversation [FEATURE-INTRO]. These were chosen because although they are important to the task in regards to the status of shared knowledge between the two interactants, they are not directly prompted by the other interactant or the task – i.e. the route has to be explained (instructions used), but features don’t have to be checked in advance and you can choose whether or not to check something before moving on to the next instruction.

Some change in behaviour in these two attributes was noticed over time. For checks, there was no significant difference between the positive totals for the two halves of the dataset, but the failure to use checks did decrease over time (Wilcoxon-Mann-Whitney $z = 2.83$, $p = 0.0024$, one-tailed test).

For FEATURE-INTRO the results were more complex. Although Wilcoxon-Mann-Whitney showed no difference in either positive or negative totals, a Chi-Square test did show an interaction between feature type (shared, unshared), time (first giving, second giving) and the failure to use FEATURE-INTRO. There was a marked decrease in the failure to use feature-intro *in unshared features* over time (df = 1, $\chi^2 = 6.1$, $p < 0.01$, one-tailed test). So, in each case, although about the same number of each of these strategies are used, the places in which they are employed are chosen more effectively.

2. Speakers will improve at tasks

Wilcoxon-Mann-Whitney was used to test for a significant decrease in the Incorrect Entity Score between the two halves of the data set. A significant result was found ($z = 2.11$, $p = 0.0179$, one-tailed test).

3. Decrease in effort over time

Three tests were undertaken on this:

1. Significant decrease in dialogue length from first part of quad to second part of quad.
2. Significant decrease in total positive score as above.
3. Significant decrease in average positive score (per utterance) as above

All three measures were used as we believe that the length of a dialogue is not necessarily indicative of degree of effort in itself (or of collaboration – see H4b Collaboration, following). Even so, none of these were found to be significant. However, as there has been evidence of behaviour modification over time (see H1 above), a further investigation was undertaken to see whether effort was being focused differently (which would also support this principle).

3a Decrease in risk over time

In this set of tests, we equate ‘risks’ with negative coding: this is justified since poor task success (high Incorrect Entity Score) correlates with a high total negative score (see H3, Cooperation).

Therefore, to test whether there was a significant decrease in risk over time, Wilcoxon Mann-Whitney was used to compare the total negative score for dialogues in the first half of each quad against the second half of each quad. A significant decrease in the scores was found ($z = 2.09$, $p = 0.0183$, one-tailed test).

Therefore, as with H1 above, we conclude that although there is no overall decrease in effort, the same resources are being targeted more effectively.

6.1.1 Summary

The hypotheses developed for the Principle are reasonably well supported. There is evidence of changes in behaviour, improvement in task success and a refocusing of effort (which appears to have been effective). These hypotheses are all interdependent, and they all needed to demonstrate support for the reasoning behind the hypotheses to be upheld. Rational/Efficient interactants learn to invest effort effectively.

6.2 Collaboration

1. Both speakers contribute

As both interactants contribute turns and words (which is Wilkes-Gibbs' 1986 definition of collaboration), this hypothesis is demonstrably supported. However, this tells us little about how or why speakers engage with each other.

2. Dialogues will get shorter the more times the participants do the task AND 3. Decrease in effort over time

These hypotheses were not supported (see H3, Gricean Cooperation). The data did not show an absolute decrease in effort. This is probably because the type of task used is a repetition of form rather than a repetition of content, unlike the tangram tasks in Clark & Wilkes-Gibbs (1986). However, this does suggest that this evidence for collaboration does not generalise easily.

4a Speakers with equal commitment (whether high or low) should be associated with more task success

Clark and his co-workers argue that absolute effort is not related to task success because dialogue is a joint production and thus is reliant on the equal input of both parties. This was tested on our data by investigating the relative similarity in use of the high effort attribute types. We argue that speakers who assign themselves as high-criterion will use a lot of high effort strategies, whereas those who assign themselves as low criterion won't. For speakers in a dialogue to show equal commitment, they should use about the same number of these high effort strategies; in unequal dialogues, the speakers will use different amounts. A score to represent this was calculated as follows:

1. Total number of high effort attributes used for each of the Giver and the Follower.
2. Each speaker's total is then represented as a proportion of the overall total (these two fractions should add up to one).
3. The absolute difference between the two proportions is then calculated.

This process should give a number in value from zero to one, where the higher the figure the greater the difference in the number of attributes used by each speaker. Table 4 gives a constructed example of how this works.

No.	Giver	Follower	Total	Giver prop.	Follower prop.	Diff	Order
1	12	6	18	0.67	0.33	0.34	2
2	9	9	18	0.5	0.5	0	1
3	2	16	18	0.11	0.89	0.78	3

Table 4. Ordering the proportion of high effort strategies used by Giver and Follower.

Therefore, the calculation represents any difference in the effort invested by the two speakers. The dataset was ordered from smallest difference (most equal commitment) to largest difference (most unequal commitment), and was then tested for correlation with task success. The result was highly significant ($r_s = 0.820$, $p \leq 0.0005$, one-tailed test). Therefore, this hypothesis can be considered to be strongly supported.

4b There is no relationship between increased collaboration and task success

A number of correlations were undertaken to test this hypothesis as there are various ways in which 'effort' or 'degree of collaboration' could potentially be measured. The first four are based on Wilkes-Gibbs' measures and the remainder use information provided by the coding. In each case, the measure is tested for correlation with Task Success (based on Incorrect Entity Score) across the whole dataset.

1. Number of turns in a dialogue
2. Number of words in a dialogue
3. Average words per turn (for Giver)
4. Average words per turn (for Follower)[8]
5. Total positive score
6. Average positive score per utterance
7. Total of high effort attribute types – because these represent 'unnecessary' effort on behalf of the speaker
8. Percentage of utterances coded as using high effort attributes in a dialogue

None of these tests produced a significant result; in fact, the first test comes close to showing a negative correlation (r_s = -0.313, $p \leq 0.1$, 2-tailed test). This is interesting, because these various tests show a range of ways of thinking about effort – amount of talk, length of utterances, amount of work put into an utterance, number of effortful utterances – yet none of them have shown any marked degree of positive correlation. While we would have argued that Wilkes-Gibbs' measures are rather limited, those based on information from the coding are more focused and detailed.

However, this result can only be considered as weak support, as it involves acceptance of the null hypothesis and not a significant difference.

6.2.1 Summary

At a superficial level, there is some support for the Collaborative Principle and the Principle of Least Collaborative Effort. However, the main support comes from the upholding of the importance of equal commitment to a dialogue. We have argued earlier in this paper that we view Wilkes-Gibb's (1986, 1997) explanation of these phenomena to be problematic and this point will be taken up in the final section.

6.3 Cooperation

1. High effort leads to greater task success AND 2. Low effort leads to low task success

These two hypotheses have been taken together because for Cooperation, they must be seen as two sides of the same coin. This distinguishes Cooperation from Gricean Cooperation and the Principle of Parsimony.

[8] Wilkes-Gibbs' measure is an average overall turns, not by speaker. However, this is not appropriate for the data, as the Giver and Follower do not typically contribute equally to the dialogue (overall average percentage of talk by the Follower = 28.97%), and their mean turn length is also quite different (Giver = 11.02 and Follower = 4.97). Combining such different populations would not be appropriate statistically.

- High effort is not associated with task success

A number of tests were undertaken (see H4b, Collaboration above) involving overall positive score, length of the dialogue and high effort strategies. None were found to be at all significant.

- Low effort is associated with poor task result

A number of tests were undertaken (see H3, below) involving overall negative score, average negative score, and the relationship between incidences of negative coding in dialogue segments and task errors in that segment. All tests were found to be significant.

Overall, this pair of hypotheses are not supported, because they have to be seen as interdependent.

3. Few risks would be taken

The incidence of negative scores is taken to be evidence that participants in the conversations take risks. This is due to the way in which the Typology was developed (section 3.2) and is also demonstrated by our finding a statistically significant relationship between task errors and negative score. We have investigated this link in terms of both relationships between negative score and poor task success, and also negative score and incidences of dialogue failure.

Does total negative score correlate with task success?

The results from Spearman's Rank Correlation were highly significant: $r_s = 0.820$, $p \leq 0.0005$ (one-tailed test).

Does average negative score correlate with task success?

This test was also highly significant: $r_s = 0.573$, $p \leq 0.0005$ (one-tailed test).

Are occasions of dialogue failure associated with negative score?

The purpose of this test was to investigate the relationship between incidences of negative coding and task errors in smaller subsections of the dialogue, rather than just at the more general level of the dialogue as a whole. In order to do this, segments of dialogue were categorised as 'error' or 'no error', and simultaneously categorised as 'high negative score' and 'low negative score'. The dialogues were divided into segments based on the route structure and the categorisation as 'error' or 'no error' was based on the Incorrect Entity score for that part of the route. In order to assign a segment as 'high' or 'low' negative score, the score for each segment was calculated and then divided into two conditions (low score, high score) along the median. Three Chi Square tests were then performed: on the whole dataset, on segments concerning unshared features only, and on segments containing shared features only. This was to ensure that feature type did not act as a confounding variable.

Significant results were found for all three tests:

- All features (df =1, χ^2 = 57.52, p < 0.0005, one-tailed test)
- Shared features (df = 1, χ^2 = 17.20, p < 0.0005, one-tailed test)
- Unshared features (df = 1, χ^2 = 21.02, p < 0.0005, one-tailed test)

Therefore, as we have shown that incidences of negative coding do appear to be associated with errors, it would seem reasonable to conclude that failures of this type do constitute ‘risks’. As no dialogue scored zero negative codings, we must conclude that interactants do take risks, and thus this hypothesis is not supported.

4. Expect modification of behaviour over time

Change in behaviour was tested on checking routines (CHECK) and the use of feature introductions (FEATURE-INTRO). In both cases, there was a significant decrease in the *failure* to use these attributes, but no difference in their overall incidence (a fuller explanation is given under H1, Grice). This could be seen as only partial support for Cooperation, because the changes are always downwards rather than being in either direction as the Principle predicted.

5. Expect better task success with experience

Wilcoxon-Mann-Whitney was used to test for a significant decrease in the Incorrect Entity Score between the two halves of the data set (first giving/second giving). A significant result was found (z = 2.11, p = 0.0179, one-tailed test).

6.3.1 Summary

The support for task success improvement and behaviour modification is undermined by the lack of support for the other hypotheses: they must be seen as interdependent. It is clear that speakers do take risks, which is not predicted by Cooperation, and whilst low effort does lead to poor task success, the converse does not follow. It could also be suggested that the finding that equal commitment is important (Collaboration, Least Individual Effort) also weakens the case for Cooperation: the effort of one cannot overcome the low criterion of another. This is scarcely ‘helpful’ behaviour.

6.4 Parsimony: Risk-Effort Trade-Off and the Principle of Least Individual Effort

1. Risks would be taken – some failures

The incidence of negative scores is taken to be evidence that participants in the conversations take risks. This is due to the way in which the Typology was developed (section 3.2) and is also demonstrated by our finding a statistically significant relationship between task errors and negative score (see H3, Cooperation).

2. Risks would decrease over time – fewer failures

Again, the assumption is made that incidences of negative codings are equivalent to risks. Wilcoxon-Mann-Whitney was used to test for a significant decrease in the Total Negative Score between the two halves of the data set. A significant result was found (z = 2.09, p = 0.0183, one-tailed test). Given the statistically significant correlation between negative score (and thus risk taking) and task success, we can assume that fewer risks will lead to fewer failures.

3. Task success will improve (as speakers negotiate trade-off more successfully over time)

Wilcoxon-Mann-Whitney was used to test for a significant decrease in the Incorrect Entity Score between the two halves of the data set. A significant result was found ($z = 2.11$, $p = 0.0179$, one-tailed test).

4. Behaviour will be modified as speakers try out different risk-effort combinations, and eventually settle on a set of useful combinations

Change in behaviour was tested on checking routines (CHECK) and the use of feature introductions (FEATURE-INTRO). In both cases, there was a significant decrease in the *failure* to use these attributes, but no difference in their overall incidence (a fuller explanation is given under Gricean Cooperation). This could be seen as only partial support for Parsimony, because the changes are always downwards. As speakers are equally likely to start from a high or low risk posture, then they should be considered equally likely to adjust their effort levels downwards or upwards respectively.

5. Decrease in Effort – Refocusing [Hypothesis 3 & 3a, Gricean Cooperation]

Although no overall difference in effort was noted (see Gricean Cooperation), there is evidence in terms of *Behaviour modification* and *Decrease in risk* that suggest effort is being refocused to be used more effectively. This outcome would seem as much evidence for the risk-effort trade-off as Gricean Cooperation, so it is also included here.

6. Low Effort = Low Task Success [risks cause errors]

A number of tests were undertaken (H3, Cooperation) involving overall negative score, average negative score, and the relationship between incidences of negative coding in dialogue segments and task errors in that segment. All tests were found to be significant. A decrease (in at least some types of) effort is associated with risk. This significant result reiterates the relationship between risk, effort and task success. Therefore, it is also included here.

6.4.1 Least Individual Effort

1a Speakers with equal commitment (whether high or low) should be associated with more task success

The relationship between the similarity in number of high effort strategies used by two participants and task success was investigated. A highly significant result was found (see H4a, Collaboration).

1b There is no relationship between increased collaboration and task success

Six different tests were undertaken which took into consideration a number of ways of calculating the overall effort expended in a dialogue (i.e. taking into account the contribution of both participants). None were found to be significant (see H4b, Collaboration).

6.4.2 Summary

Good support was found for this principle. All the suggested hypotheses were found to be supported: task success improves over time, and is associated with both changes in behaviour and a decrease in risks taken as speakers work out what risks are worth taking (risk-effort trade-off) and which aren't. This is also associated with refocusing of effort: overall effort may not decrease, but speakers seem to be using it more effectively (finding a more optimum point on the risk-effort trade-off), since they are improving their performance on the task. In contrast to Cooperation, the support found for improvement in task success and behaviour modification finds greater weight through support for the other interdependent hypotheses.

Particular support was also found for the Principle of Least Individual Effort: equal commitment of the two participants was found to be more important than any measure of overall effort and thus the actions of individuals have a substantial effect on dialogue outcome.

7. Discussion

While we have argued that each Principle has a set of interdependent hypotheses, the Principles themselves have been set into an oppositional relationship to each other. In this discussion, we will consider the implications of the results found in terms of descriptions of dialogue behaviour and their potential motivations.

First, we will list the findings of the study:

- Task success improves over time
- Behaviour modifies over time (CHECK, FEATURE-INTRO)
- Participants take risks
- Participants take fewer risks over time
- Overall effort does not decrease, but there is refocusing of effort
- Low effort leads to poor task success
- High effort shows no relationship with task success
- Equal commitment leads to task success

This represents some support for each of the Principles – which in itself demonstrates the difficulty in teasing apart the behaviours predicted – but the Principle of Parsimony is better able to motivate the findings, and integrate them into an overall explanation, as we will demonstrate below.

While Collaboration could arguably explain the taking of risks in terms of the self-selection of speakers as high or low criterion, it cannot so easily account for the shifts in behaviour with regard to risk The decrease in risk over time would have to be interpreted as a change in speaker criterion, but this would not seem to make sense in terms of the Collaborative Theory: why should speakers shift in their perception of the Principle of Mutual Responsibility, which states that participants should "try to establish … the mutual belief that they have understood what the contributor meant, to a criterion sufficient for their current purposes" (Clark & Wilkes-Gibbs 1986: 33). The concept of criteria does seem intuitively sensible – that certain speech events are worth understanding to a greater degree of certainty than others – and we would also suggest that individual speakers will vary in their judgement of the criterion of a particular speech event. But it makes no sense to suggest that speakers' judgements will regularly change partway through a task.

Related to the change in risks taken, we also see adjustments in how effort is employed – in the use of particular strategies, and as an overall profile of the dialogue. So, unlike the predictions of the Collaborative Theory, we did not find a reduction in effort – collaborative or otherwise. More importantly, we can also link these findings with an issue not really considered within the collaborative model: task success. While the participants in the tasks set by Clark and his co-workers often did improve over repetition, this was not commented on or related to the collaborative model. These measures of task success were used purely to demonstrate the grounding process through distinguishing the levels of understanding achieved by different types of participants. However, from the viewpoint of the risk-effort trade-off, these changes in behaviour can be seen as adjustments made as part of the learning process, which lead to a more effective gauging of effort-risk, and thus better task results.[9]

In addition to finding results consistent with the risk-effort trade-off model, the detailed profile provided by our coding system also offers a potential refinement of it. What we find is not just an inverse relationship between sheer effort and risk, but rather between *well-targeted* effort and risk: overall effort does not decrease, but risks do. This can be linked back to the distinction we drew between our definition of risk and that used by Shadbolt (1984) and Carletta (1992): risk in our definition entails the potential for failure to misunderstand, whereas for the original model it entails only the necessity for repair work which will resolve the problem. Because humans cannot guarantee that problems will be noticed (or successfully repaired, even if they are) then the importance of focusing effort in the right places is magnified. It also suggests the amount of effort invested is relatively inelastic: the participants do not simply invest extra effort to solve or avoid problems, instead they try to invest their budget more wisely. This would perhaps suggest that Shadbolt's original conception of agents being as likely to start a task with a low risk posture as a high risk posture is somewhat idealistic: our evidence suggests that most human interactants are effort-averse and the difference would appear to be in their reaction to negative evidence (i.e. their ability/willingness to learn and change strategy).

Overall, this finding of being effort-averse would seem to be compatible with Wilkes-Gibbs' (1986) finding that in her mixed pairs, the LC participants did not appear to be willing to increase their effort invested to match that of their HC partners. And there are other similarities with her findings: the lack of relationship between absolute effort (by any of a number of possible types of measurement) and task success, and in contrast, the significant relationship between equal commitment by the two participants and task success. Arguably, our results are potentially more interesting from the point of view of studying human behaviour because they represent choices made entirely

[9] One might argue that repetitions of dialogue-led tasks such as these offer opportunities for learning (and 'improvement') that are not generalisable to naturally occurring dialogue. However, in addition to bringing into question all of Clark's work as well as that described here, such a viewpoint would also ignore the argument that language socialisation is essentially a process of learning, as evidenced by the use of such models as communities of practice to explain language variation and change (e.g. Eckert 1999), which has been imported from Education (e.g. Lave & Wenger 1991).

independently by the participants rather than different situations created by the experimenter.

However, we disagree with their explanation of their findings which seems somewhat rose-tinted (see sections 2.2 & 2.3). In their analysis of the mixed pair data, the reluctance of the LC participants to increase their effort input coupled with the tendency for the HC participants to downgrade their sights is interpreted as indicative of the flexibility typical of HC participants, rather than the unwillingness of the LC participant to accept a greater demand on their resources. This view seems particularly inappropriate given the detrimental effect on the relative task success of these HC participants: if the flexibility of these participants is seen to be indicative of their higher criteria in relation to 'current purposes', then surely it should at least not lead to a negative effect on their performance? The impact of the criterion of the individual on this jointly constructed process thus cannot be denied: conversation may be collaborative in terms of the joint construction of common ground, but the decisions about effort are made on an individual basis. And all the evidence we see here points to participants generally trying to minimise their own effort, rather than considering the effort of all involved.

The question of effort also links in to the predictions made by Cooperation and Gricean Cooperation – or, more particularly, the difference between them. The folklinguistic notion of Cooperation relies on an assumption of helpfulness, as characterised by Brown (1995) (see section 2.1), whereas a Gricean view draws on concepts of rationalism and (more arguably) efficiency. While the nature of both of these Principles makes them harder to operationalise than Parsimony/Least Individual Effort and Collaboration/ Least Collaborative Effort, it is clearly evident that they are not equivalent. If talk is based on a Principle of Cooperation, then we would expect effort to be correlated with task success and there to be little or no risk-taking, particularly in the second half of each of the quads. In contrast, the Gricean view predicted only that risk-taking would decrease over time, as would the amount of effort invested by the speakers as they learnt to use it more effectively. The findings were much closer to those predicted by our interpretation of Grice: speakers do take risks, and they do decrease over time; sheer effort is not correlated with task success; and although there is no absolute decrease in effort, there is evidence that speakers refocus effort to use it more effectively (which is a kind of effort conservation).

While both of these principles predicted an improvement in task success and changes in behaviour (which were both supported by the data), Gricean Cooperation can motivate this through its other findings: task success improves because speakers learn which behaviours are more effective and they modify their strategies accordingly. In contrast, the folklinguistic notion of Cooperation relies on the idea of speakers doing more helpful things to improve task success, and not just fewer unhelpful ones. The fact that speakers continue to undertake behaviours which jeopardise task success and can also behave in such a way that impedes their partners (i.e. equal commitment) also mitigates against an analysis of conversation as a Cooperative enterprise. Thus it would seem clear that this Principle can be largely rejected.

Gricean Cooperation is rather harder to evaluate, for although it is reasonably well supported *in the interpretation we have offered* – and which we have motivated through a broader perspective on Grice's work – the move from an abstract principle in Ordinary Language philosophy to a set of empirically testable hypotheses will always involve scope for discussion. However, our point here was not so much to present an unchallengeable operationalisation of the CP, but to provide a principled differentiation from a folklinguistic notion of Cooperation. This we have demonstrated, and shown the CP to be better supported empirically than the folklinguistic notion. In its current form, the hypotheses are a subset of those offered for the Principle of Parsimony, albeit with somewhat differently focused explanations. This is perhaps not surprising, given that risk-effort could be viewed as a possible realisation of efficiency.

However, there is little question that our data provides the most support to the Principle of Parsimony and Least Individual Effort. It provides a more consistent and complete explanation at this level of discourse than that offered by Collaboration and Least Collaborative Effort in the contexts of both their data and the data presented here. The folklinguistic notion of Cooperation has been shown to be just that: an assumption about the way in which people talk, which is not grounded in data. And Gricean Cooperation, although subject to rather too much discussion to claim full support, has been clearly differentiated from the non-technical notion with which it is too-often confused.

Of course, the issue here is not just what we have found, but how we have set about the problem of generating that information. The problem of operationalisation mentioned above did not only apply to the more abstract principles. Reinterpreting Parsimony/Least Individual Effort and Collaboration/Least Collaborative Effort in a different type of dataset and a different coding scheme was not an automatic process, and thus inevitably leaves room for alternative explanations. However, if progress is to be made in our understanding of how talk works, then novel approaches need to be employed. In this case, the development lies not only in applying these principles to real talk, but also in designing an analysis method based as much on evaluation as description. The pay-off here is a broad profile of information about what real speakers do in a real speech context, albeit one which is constrained by its task-orientation. Its purpose is to open out the discussion about how people 'do' talk: the clear evidence for Parsimony and Least Individual Effort is an interesting first step, but is an issue which deserves investigation in a wider range of speech contexts.

References

Anderson, A.H., Bader, M., Bard, E.G., Boyle, E., Doherty, G.M., Garrod, S., Isard, S.D., Kowtko, J.C., McAllister, J., Miller, J., Sotillo, C.F., Thompson, H.S. & Weinart, R. (1991a) The HCRC Map Task Corpus. *Language and Speech* **34.**4. 351-366.

Anderson, A.H., Boyle, E.A. (1994) Forms of introduction in dialogues: Their discourse contexts and communicative consequences. *Language and Cognitive Processes* **9**. 101-122.

Anderson, A.H., Clark, A. & Mullin, J. (1991b) Introducing information in dialogues: Forms of introduction chosen by young speakers and the responses elicited from young listeners. *Journal of Child Language* **18**. 663-687.

Anderson, A.H., Clark, A. & Mullin, J. (1994) Interactive communication skills in children: Learning how to make language work in dialogue. *Journal of Child Language* **21**. 439-463.

Bargiela-Chiappini, F. & Harris, S. (1997) *Managing Language: The discourse of corporate meetings*. Amsterdam: John Benjamins.

Beebe, L.M. & Cummings, M.C. (1996) Natural speech act data versus written questionnaire data: How data collection method affects speech act performance. In Gass, S.M. & Neu, J. (eds.), *Speech Acts Across Cultures*. Berlin: Mouton de Gruyter. pp.65-86.

Brennan, S.E., Clark, H.H. (1996) Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition* **22. 6**. 1482-1493.

Brown, G. (1995) *Speakers,Llisteners and Communication: Explorations in discourse analysis*. Cambridge: Cambridge University Press.

Carletta, J. (1992) *Risk-taking and Recovery in Task-oriented Dialogue*. Unpublished PhD thesis, University of Edinburgh.

Carletta, J. (1996) Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics* **22. 2**. 249-254.

Carletta, J., & Mellish, C. (1996) Risk-taking and recovery in task-oriented dialogue. *Journal of Pragmatics* **26**. 71-107.

Carletta, J., Isard, S., Doherty-Sneddon, G., Isard, A., Kowtko, J.C. & Anderson, A. H. (1997) The reliability of a dialogue structure coding scheme. *Computational Linguistics* **23. 1**. 13-31.

Clark, H.H. (1992) *Arenas of Language Use*. Chicago: University of Chicago Press.

Clark, H.H. (1996) *Using Language*. Cambridge: Cambridge University Press.

Clark, H.H, Brennan, S.E. (1991) Grounding in communication. In Resnick, L., Levine, J., & Teasley, S. (eds.) *Perspectives on Socially Shared Cognition*. Washington, DC: American Psychological Association. pp. 127-149.

Clark, H.H. & Krych, M.A. (2004) Speaking while monitoring addressees for understanding. *Journal of Memory and Language* **50**. 62-81.

Clark, H.H. & Schaefer, E.F. (1987a) Collaborating on contributions to conversations. *Language and Cognitive Processes* **2**.19-41.

Clark, H.H. & Schaefer, E.F. (1987b) Concealing one's meaning from overhearers. *Journal of Memory and Language* **26**. 209-225.

Clark, H.H. & Wilkes-Gibbs, D. (1986) Referring as a collaborative process. *Cognition* **22**. 1-39.

Connor, U. & Upton, T.A. (eds.) (2004) *Discourse in the Professions: Perspectives from corpus linguistics*. Amsterdam: Benjamins.

Coupland, N., Giles, H. & Wiemann, J.M. (eds.) (1991) *'Miscommunication' and Problematic Talk*. London: Sage.

Davies, B.L. (1998) An Empirical Examination of Cooperation, Effort and Risk in Task-oriented Dialogue. Unpublished PhD thesis, University of Edinburgh.

Davies, B.L. (to appear) Grice's Cooperative Principle: Meaning and rationality.

Davies, B.L. (in prep) Least Collaborative Effort or Least Individual Effort: Examining the Evidence.

Eckert, P. (1999). *Linguistic Variation as Social Practice*. Oxford, U.K.: Blackwell.

Eelen, G. (2001) *A Critique of Politeness Theories*. Manchester: St. Jerome's Press.

Fais, L. (1994) Conversation as collaboration: Some syntactic evidence. *Speech Communication* **15**. 231-242.

Fawcett, R.P. & Davies, B.L. (1992) Monologue as a turn in dialogue: towards an integration of Exchange Structure and Rhetorical Structure Theory. In Dale, R., Hovy, E., Rösner, D. & Stock, O. (eds.) *Aspects of Automated Natural Language Generation*. Berlin: Springer Verlag.

Finch, G. (2000) *Linguistic Terms and Concepts*. London: Macmillan Press.

Grice, H.P. (1975) Logic and conversation. In Cole, P. & Morgan, J.L. (eds.) *Syntax and Semantics, Vol. 3: Speech Acts*. New York: Academic Press. pp. 41-58.

Grice, H. P. (1986) Reply to Richards. In Grandy, R. & Warner, R. E. (eds.) *Philosophical Grounds of Rationality*. Oxford: Clarendon Press. pp. 45-106.

Grice, H. P. (1989) *Studies in the Way of Words*. Cambridge, Mass: Harvard University Press.

Grosz, B. & Sidner, C. (1986) Attention, intention, and the structure of discourse. *Computational Linguistics* **12**. 175-206.

Houghton, G. & Isard, S.D. (1987) Why to speak, what to say and how to say it: Modelling language production in discourse. In Morris, P. (ed.) *Modelling Cognition*. Chichester: Wiley. pp. 249-267.

Isard, A. & Carletta, J. (1995) Transaction and action coding in the Map Task corpus. *Tech. Rep. HCRC/RP-65*. Edinburgh, Scotland: Human Communication Research Centre, University of Edinburgh.

Kowtko, J.C.& Isard, S.D., Doherty, G.M.(1992) Conversational games within dialogue. *Tech. Rep. HCRC/RP-31*. Edinburgh, Scotland: Human Communication Research Centre, University of Edinburgh.

Lave, J. & Wenger, E. (1991) *Situated Learning: Legitimate peripheral participation*. Cambridge, U.K.: Cambridge University Press.

Leech, G. (1983) *Principles of Pragmatics*. London: Longman.

Levinson, S. (1983) *Pragmatics.* Cambridge: Cambridge University Press.

Lin, Y.Q., Fawcett, R.P. & Davies, B.L. (1993) Genedis: The discourse generator in COMMUNAL. In Sloman, A., Hogg, D., Humphreys, G., Ramsay, A. & Partridge, D. (eds.) *Prospects for Artificial Intelligence.* Amsterdam: IOS Press.

Markie, P. J. (2000) Rationalism. In *Concise Routledge Encyclopedia of Philosophy*. London: Routledge.

Power, R. (1979) The organisation of purposeful dialogues. *Linguistics* **17**. 107-152.

Sacks, H., Schegloff, E.A. & Jefferson, G. (1974) A simplest systematics for the organisation of turn-taking in conversation. *Language* **50**. 696-735.

Schegloff, E.A (1968) Sequencing in conversational openings. *American Anthropologist* **70**. 1075-95.

Schober, M.F. (1995) Speakers, addressees, and frames of reference: Whose effort is minimised in conversations about locations? *Discourse Processes* **20**. 219-247.

Schober, M.F. & Clark, H.H. (1989) Understanding by addressees and overhearers. *Cognitive Psychology* **21.** 211-232.

Shadbolt, R.N. (1984) *Constituting Reference in Natural Language Dialogue: The problem of referential opacity*. Unpublished PhD thesis, University of Edinburgh.

Siegel, S.& Castellan, N.J. (1988) Nonparametric Statistics for the Behavioral Sciences. London: McGraw Hill.

Sinclair, J. McH & Coulthard, M. (1975) *Towards an Analysis of Discourse: The English used by teachers and pupils*. London: Oxford University Press.

Sperber, D. & Wilson, D. (1986) *Relevance*. Oxford: Blackwell.

Stenström, A. (1994) *An Introduction to Spoken Discourse*. London: Longman.

Traum, D.R. (1994) *A Computational Theory of Grounding in Natural Language Conversation.* Unpublished PhD thesis, University of Rochester.

Wilkes-Gibbs, D. (1986) *Collaborative Processes of Language Use in Conversation*. Unpublished PhD thesis, Stanford University.

Wilkes-Gibbs, D. (1997) Studying language use as collaboration. In Kasper, G. & Kellerman, E. (eds.) *Communication Strategies: Psycholinguistic and Sociolinguistic Perspectives*. London: Longman. pp. 238-274.

Wilkes-Gibbs, D. & Clark, H.H. (1992) Coordinating beliefs in conversation. *Journal of Memory and Language* **31**. 183-94.

Bethan Davies
Department of Linguistics & Phonetics
University of Leeds
LS2 9JT

B.L.Davies@leeds.ac.uk