An Empirical Examination of Cooperation, Effort and Risk in Task-Oriented Dialogue

Bethan L. Davies

Ph.D.
University of Edinburgh
1997
Abstract

This thesis presents a discussion of proposed structuring principles for dialogue, and tests them empirically using data from the HCRC Map Task Corpus.

The concepts of Cooperation (as described by Grice, and as used more generally in Linguistics), Coordination, Collaboration, Parsimony, Risk and Effort are examined, and empirically testable hypotheses are developed, with which we are able to evaluate the claims for these principles in the context of task-oriented dialogue.

In order to test our hypotheses, we categorise the utterances in our database in terms of Risk and Effort. Unlike Discourse Analysis, Conversation Analysis or Dialogue Games, our approach is evaluative. The intent in our dialogue coding is not only to label what the speakers did, but also to assess it in terms of its appropriateness at that point in the dialogue: the system codes not only what people do, but also what they don't do. Therefore, our system marks both the presence and absence of dialogue attributes.

The hypotheses derived from the structuring concepts were statistically tested on the data produced by the coding system. The results produced by the empirical tests showed a relationship between dialogue errors and task errors, but not between increased effort and increased task success. The importance of *matched effort* was also demonstrated, as dialogue pairs who invested similar amounts of effort produced better task results. Dialogue pairs also produced better results over time, which we argue is due to the focusing of effort. Participants work out where their effort should be channelled so that they can increase risk-taking where problems have not occurred, and decrease risk-taking where problems have occurred.

These results suggest that interactants' behaviour follows a Principle of Least Individual Effort, which we argue subsumes the Principle of Parsimony and thus the Risk-Effort Trade-Off. We reject the Principle of Least Collaborative Effort because although the empirical result of high effort not being associated with task success supports this principle in theory, we argue that its motivation is not supported in practice.

The empirical work also distinguishes between what we term 'Gricean Cooperation', and folklinguistic notions of Cooperation found in the literature. In general terms, Gricean Cooperation predicted the same type of effort-minimising behaviour as the Principle of Least Individual Effort, and was thus supported by the empirical work. However, the concept of 'helpfulness' suggested by more general uses of Cooperation made predictions which were in conflict with those of the Principle of Least Individual Effort, and were found not to be supported.
Acknowledgements

Thanks to my main supervisor, Steve Isard, who has always made me clarify my ideas, and has provided useful criticisms and perceptive comments on the drafts of this thesis. Any errors that remain are, of course, my own, and are probably due to my stubbornness. Thanks also to my second supervisor, Chris Mellish, who has always been very encouraging about my work.

Thanks to all those who have contributed to my work or my sanity over the last five years. There are far too many people to individually mention, but a general ‘thank you’ should go to all the folk at the Human Communication Research Centre (particularly the Dialogue Group), the Departments of Linguistics (not forgetting Ethell!), Cognitive Science and Artificial Intelligence. Especially those who persuaded me that the Pub was an appropriate place to discuss the finer points of Linguistics...

Special thanks should go to:

- My office mates and fellow dialoggers at HCRC: Matthew Aylett, Jean Carletta, Jacqui Deverell, Amy Isard, Jacqueline Kowtko, Andrew Merri- son, Jim Miller, Cathy Sotillo, who between them have provided much tea, sympathy, biscuits and laughter - as well as serious linguistic discussion, of course.

- My new Leeds compatriots, Catriona McPherson and Diane Nelson, for helping me keep things in perspective (mainly through food, wine, and trips to Tesco, IKEA and the Turkish baths...) .

- Catriona (again) and Andrew (again) for ‘willingly’ taking part in the replication study from hell...

- The Computer Support Team, who have answered all my queries, and have helped me through the joys of \LaTeX{}.

- Robin Fawcett and Ted Lawson, who got me into this in the first place, and to Geoff Lindsey, who helped me out when the going got tough.

- My friends and family - whom I hope will forgive me for disappearing for 18 months whilst I was writing up.

And a final big, big thanks to Bill Hewitt, for cooking more dinners, solving more computer problems and writing more awk scripts than anyone could expect. And for (mostly) forgiving me when the thesis got in the way of real life...

I acknowledge financial support from the Engineering and Physical Sciences Research Council.
Declaration

I hereby declare that I composed this thesis entirely myself and that it describes my own research.

Bethan L. Davies
Edinburgh
April 8, 1997
Table of Contents

Abstract ii
Acknowledgements iii
Declaration iv
List of Figures xii

Chapter 1 Introduction 1
  1.1 The Approach: Dialogue Analysis and Statistical Methods ... 2
  1.2 The Major Contributions ........................................... 3
    1.2.1 The Typology of Move Attributes ............................. 4
    1.2.2 Evaluation of Dialogue Principles ........................... 4
    1.2.3 Empirical Test of Dialogue Principles ....................... 5
  1.3 The Organisation of the Thesis ................................... 7

Chapter 2 Review of Approaches to Dialogue 10
  2.1 Introduction ...................................................... 10
  2.2 Philosophical Approaches to Dialogue .......................... 10
    2.2.1 Speech Act Theory ........................................... 10
    2.2.2 Searle - Speech Acts ........................................ 12
    2.2.3 The Cooperative Principle and Maxims ....................... 13
  2.3 Sociolinguistic Approaches to Dialogue ......................... 14
    2.3.1 Discourse Analysis .......................................... 15
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3.2</td>
<td>Conversation Analysis</td>
<td>17</td>
</tr>
<tr>
<td>2.4</td>
<td>Psycholinguistic Approaches to Dialogue</td>
<td>21</td>
</tr>
<tr>
<td>2.4.1</td>
<td>Task Success in the Map Task Corpus</td>
<td>21</td>
</tr>
<tr>
<td>2.4.2</td>
<td>Clark's Collaborative Model</td>
<td>23</td>
</tr>
<tr>
<td>2.4.3</td>
<td>Output-Input Coordination Model</td>
<td>35</td>
</tr>
<tr>
<td>2.5</td>
<td>Computational Approaches to Dialogue</td>
<td>37</td>
</tr>
<tr>
<td>2.5.1</td>
<td>Dialogue Games</td>
<td>37</td>
</tr>
<tr>
<td>2.5.2</td>
<td>Communicative Postures</td>
<td>41</td>
</tr>
<tr>
<td>2.5.3</td>
<td>Discourse Obligations</td>
<td>44</td>
</tr>
<tr>
<td>2.6</td>
<td>Summary</td>
<td>46</td>
</tr>
</tbody>
</table>

Chapter 3  Cooperation as a Structuring Principle  47

3.1  Introduction | 47
3.1.1  Competing Principles for Dialogue Structure | 47

3.2  Structuring Principles for Conversation | 48
3.2.1  The Cooperative Principle | 49
3.2.2  Dialogue as Cooperation | 54
3.2.3  Dialogue as Joint Action | 56
3.2.4  Dialogue as Risk-Effort Trade-off | 68
3.2.5  The Principle of Parsimony and the Principle of Least Individual Effort | 70

3.3  Definitions of Cooperation | 71
3.3.1  Gricean Cooperation versus Cooperation | 73

3.4  Cooperation in an Empirical Context | 74
3.4.1  Gricean Cooperation | 74
3.4.2  Coordination | 75
3.4.3  Collaboration and the Principle of Least Collaborative Effort | 76
3.4.4  Principle of Parsimony | 78
3.4.5  Cooperation | 80
5.3 Effort in the Typology of Move Attributes ................................................. 110
  5.3.1 A Definition of Effort ................................................................. 110
  5.3.2 Effort in the Typology ................................................................. 113
  5.3.3 Using Effort Groupings ................................................................. 114
  5.3.4 A Note on Examples ................................................................. 115
5.4 1 - Minimum Effort ................................................................................. 116
  5.4.1 PARTNER-BELIEF - Belief in partner’s contribution ..................... 117
  5.4.2 REPLY-MIN - Provide a Response ...................................................... 120
  5.4.3 REPLY-YN - Answer Queries ........................................................... 124
  5.4.4 ACK-SHORT - Minimal Follow-up ..................................................... 125
5.5 2 - Moderate Effort .................................................................................. 126
  5.5.1 Checking Routines ............................................................................. 128
  5.5.2 REPLY-FULL - Be Appropriately Informative ..................................... 132
  5.5.3 ACK-FULL - Follow-up Information Offered ..................................... 133
  5.5.4 ACK-REPEAT - Self-Understanding Checks ....................................... 136
  5.5.5 FEATURE-INTRO - Check Existence of New Features .................. 137
5.6 3 - Medium Effort ....................................................................................... 141
  5.6.1 QUERY - Queries .............................................................................. 141
  5.6.2 OBJECTION - Objection to information Given ................................... 144
  5.6.3 KNOWLEDGE-MISMATCH - Querying Different Beliefs ................. 145
  5.6.4 FEATURE-LOC - Don’t Refer to Unnegotiated Unshared Features .... 147
  5.6.5 FEATURE-UNIQUE - Don’t Assume Features are Unique ............... 149
5.7 4 - High Effort .......................................................................................... 151
  5.7.1 High Effort INITIATE moves ............................................................... 152
  5.7.2 INFO-INTEG - Negotiation of Information ........................................ 157
5.8 An Outline of Move Attributes .................................................................... 159
5.9 Summary .................................................................................................... 161

viii
Chapter 6 Empirical Tests

6.1 Introduction .............................................. 162
6.1.1 Coding .................................................. 162

6.2 Hypotheses ................................................. 163
6.2.1 Hypotheses based on the comparison of whole dialogues:
  effect of high effort ........................................... 163
6.2.2 Hypotheses based on the comparison of whole dialogues:
  effect of low effort ......................................... 163
6.2.3 Hypotheses based on changes in the dataset over time:
  changes in dialogue .......................................... 164
6.2.4 Hypotheses based on changes in the dataset over time:
  changes in task success ....................................... 164

6.3 Empirical Tests ............................................. 164
6.3.1 Hypotheses based on the comparison of whole dialogues:
  effect of high effort ........................................... 165
6.3.2 Hypotheses based on the comparison of whole dialogues:
  effect of low effort ......................................... 170
6.3.3 Hypotheses based on changes in the dataset over time:
  Changes in dialogue behaviour ............................... 173
6.3.4 Hypotheses based on changes in the dataset over time:
  Changes in task success ....................................... 179

6.4 Discussion .................................................. 179
6.4.1 Gricean Cooperation ...................................... 180
6.4.2 Coordination .............................................. 181
6.4.3 Collaboration ............................................. 181
6.4.4 Cooperation .............................................. 182
6.4.5 Parsimony: Risk-Effort Trade-Off and the Principle of Least
  Individual Effort .............................................. 183

6.5 Summary ..................................................... 185

Chapter 7 Conclusions and Further Work ................. 187
List of Figures

4.1 Example Map from the HCRC Map Task Corpus
   This map was used in dialogues EAQ4C1, EAQ4C7, EAQ8C1 &
   EAQ8C7 in our subset of the Corpus. ......................... 87

4.2 Example Map from the HCRC Map Task Corpus
   This map was used in dialogues EAQ3C4, EAQ3C6, EAQ7C4 &
   EAQ7C6 in our subset of the Corpus. ......................... 89
Chapter 1
Introduction

This thesis is concerned with the choices speakers make when engaging in a task-oriented dialogue. The data we use is taken from the HCRC Map Task Corpus (Anderson et al., 1991a). Speakers have a multitude of decisions to make when producing an utterance, for example: what they say, how they say it, how it relates to the previous utterance. Much general linguistic writing on this level of language espouses Grice's Cooperative Principle (Grice, 1975) as being the main conversational Principle which governs our language usage. However, we would suggest that this approach is problematic in two respects. Firstly, there is a confusion between what Grice intended by the Cooperative Principle, and what it has since been taken to mean. This in itself has led to a proliferation of alternative terminology, for example: Coordination, Collaboration, Parsimony. Secondly, Grice's work, and much of the comment which has followed, has been based on intuitive analysis of either made-up data or small excerpts of real data, rather than a more empirical approach\(^1\). The aim of this work is to take an empirical approach to the question of 'Cooperation', by analysing the arguments for the various suggested Principles, and applying their expectations to our data set.

The specific empirical questions we attempt to address here are what governs the choices speakers make - whether any of the conversational Principles above explain the behaviour - and what effect these choices have. This involves three separate areas of investigation. Firstly, we must elucidate what choices speakers make. We know that there are numerous possibilities in terms of utterance choices, but

\(^1\)The work by Clark on the concepts of Collaboration, and to a lesser extent, Coordination, are an exception to this. See for example Clark and Wilkes-Gibbs (1986), Clark and Schaefer (1987a), Clark and Schaefer (1987b), and for discussions of his work in this thesis, Sections 2.4.2 and 3.2.3. However, in Clark (1996), which is his most recent work, and summarises his approach to language, he rarely cites experimental findings to support his point of view.
no empirical study has been undertaken to show either what decisions speakers make, or whether these patterns of decisions change over time. Secondly, we must consider how these choices relate to various suggested conversational Principles mentioned above. Finally, we must develop a method of comparing the impact of the various decisions which speakers can make.

In this study we focus on the level of the 'move' (See Sections 2.3.1 & 5.2 for explanations of this term). Interactants have to make decisions about the general move-type they will employ, the specific move-type\(^2\) and the content which will go into that move. We label the decisions speakers make by categorising their moves according to their general and specific type, some aspects of their content, and their appropriateness for that point in the dialogue. This labelling system is termed the Typology of Move Attributes, and is described in Chapter Five.

The information generated by this process enables us to study the behaviour of dialogue participants in two respects. Firstly, it provides empirical evidence of speaker choices which could be compared to the expectations of several conversational principles (e.g. Cooperation, Coordination, Collaboration, Parsimony). Secondly, the profile of participant choices could also be compared with the task result dialogue pairs achieve, thus providing some idea of the effect of these choices.

1.1 The Approach: Dialogue Analysis and Statistical Methods

The work described here is based on the close linguistic analysis of task-oriented dialogue data. This is combined with statistical techniques, in order that we can empirically test the comparisons between speaker behaviour and the expectations of the various conversational Principles. The thesis therefore consists of two main aspects: a description of the dialogue analysis, and the background to - and results from - the empirical tests.

The dialogues are drawn from the HCRC Map Task Corpus. This task involves two speakers who each have slightly different maps of an island. One (the Giver) has a route on their map, the other (the Follower) does not. Their goal is for the Follower to replicate the Giver’s route on the Follower’s map. Because of the mismatches in the information which they have access to, this task provides many opportunities for negotiation, discussion and miscommunication. It also

\(^2\)There is, after all, more than one way to ask a question, or reply to one.
constrains the extent of the dialogue: the analyst will almost always know the intentions behind an utterance, because they have access to all the information about the task and the maps. The design of the Map Task Corpus is discussed more fully in Section 4.2.

The analysis applied to these dialogues is based mainly on a transactional, rather than an interactional approach to dialogue (Brown and Yule, 1983). In this respect, it has more in common with Birmingham School Discourse Analysis or Dialogue Games than Conversation Analysis. However, structurally it is more like Conversation Analysis as it is more concerned with the attributes of an utterance than its structural properties. The attributes are based on the choices speakers make: types of Initiates, Responses and Follow-ups are categorised - the fact of the move-type is only one part of the decision made. This is in keeping with the aims of our work, as we are trying to evaluate the contribution of an utterance to a conversation, not parse the conversation's structure. The dialogue analysis produces a series of 'tags' or 'labels' which then provide a profiles of the decisions which each speaker makes in the dialogue.

This information is then used to investigate how the behaviour documented compares with that predicted by the various conversational Principles. A series of empirical tests are undertaken, and their significance verified by statistical means. Thus we can then present the empirical, as well as philosophical, evidence for each of the discussed conversational Principles.

1.2 The Major Contributions

This thesis makes three main contributions to this area: one related to dialogue analysis methodology, one to the analysis of Dialogue Principles and one related to the empirical results. The contributions are:

- The development of a dialogue analysis which is evaluative rather than just descriptive: the Typology of Move Attributes.

- A detailed evaluation of the research on Dialogue Principles. In particular, we analyse Clark's model of Collaboration (e.g., Clark and Wilkes-Gibbs, 1986; Wilkes-Gibbs, 1986), and the notion of Cooperation (e.g., Grice, 1975; Brown, 1995).

- An empirical evaluation of Dialogue Principles, using the Typology and a corpus of task-oriented dialogues. This produced support for the Principle
of Parsimony and the Risk-Effort Trade-Off, at the expense of other such Principles as Collaboration and Cooperation.

1.2.1 The Typology of Move Attributes

There are many current modes of dialogue analysis, for example, Birmingham School Discourse Analysis (e.g., Coulthard, 1994), Conversation Analysis (e.g., Sacks et al., 1974) and Dialogue Games (e.g., Kowtko et al., 1992). What all of these have in common is a descriptive approach to dialogue: their remit is to provide a description of dialogue structure or attributes. There is no attempt to consider what choices speakers could have made, or whether they made an appropriate choice for the point in the conversation.

Because we are interested in dialogue efficacy (task success) and the concept of Dialogue Principles, both of these points are of great importance. Both these issues involve the consideration of speaker decisions and their appropriateness. Therefore, the model we propose concentrates on this element of choice.

Our aim is to be able to describe and investigate the dialogue behaviours which the interactants in our data produce. Therefore, each move is considered according to what the speaker chose to do (the descriptive element), and whether we judge them to have chosen an appropriate move (the evaluative element). By taking this approach we hope to illustrate some of the differences between those dialogue pairs who achieve good task results, and those who don't.

The Typology focuses at the level of the move. We use the terminology Initiate, Response and Follow-up as proposed by proponents of Birmingham School Discourse Analysis (Coulthard, 1994), but do not attempt to make the moves part of a higher structure (e.g., exchanges)\(^3\). We concentrate on the broad move type used (I, R or F), and the specific move type within those categories.

1.2.2 Evaluation of Dialogue Principles

Various dialogue Principles have been suggested as the motivating force behind dialogue choices: Cooperation, Coordination, Collaboration, Parsimony. Here, we offer a critique of each of the above concepts, and discuss their strengths and weaknesses when applied to dialogue. We concentrate particularly on the following issues:

\(^3\)This is because exchange structure itself is not unproblematic, and as it was not particularly relevant to our aims here, it was decided to exclude it from the model.
Firstly, there is a tension in the meaning of the term ‘Cooperation’. We argue that there is a difference between Grice’s use of the Cooperative Principle, and what it has been taken to mean in the general linguistic description of dialogue. Here, we suggest a more faithful representation of Grice (‘Gricean Cooperation’), and also attempt a definition of the term ‘Cooperation’ as it is more frequently used. This in itself is problematic, as ‘Cooperation’ is a term used fairly freely, but defined rarely. It is also a gradeable concept, and there is the question of what degree of ‘Cooperation’ is to be considered ‘Cooperative’.

Secondly, we evaluate Clark et al’s notion of Collaboration, or ‘conversation as joint action’ (e.g. Clark 1996). We suggest alternative analyses of Clark’s experimental evidence, particularly in relation to the Principle of Least Collaborative Effort, which we argue has no sound basis. Its theoretical foundations (as suggested by Clark) are flawed, and we suggest that a Principle of Least Individual Effort would be better supported by Clark’s empirical work. This argument is pursued in our own empirical work, described below.

Finally, we describe the Risk-Effort Trade-Off and the Principle of Parsimony. This is related to Clark’s work via the Principle of Least Individual Effort, which we argue is equivalent to the notion of Parsimony.

1.2.3 Empirical Test of Dialogue Principles

In addition to the theoretic analysis of Cooperation, Coordination, Collaboration and Parsimony, we also tested the predictions of these Principles empirically. This involved producing practical definitions of each Principle, and then a series of hypotheses representing the behaviours each Principle would predict, and why. This is a useful contribution because it enables a direct comparison of these Principles.

These empirical tests make the assumption that the various Principles should be treated as competing theories of dialogue structure, rather than treating them as all potentially contributing to any given dialogue. Our main reason for this is that the originators of most of these Principles would appear to take this type of approach, and it is their ideas we are principally interested in testing.

From a theoretic angle, one of the most interesting aspects of the empirical testing is that much of the dialogue behaviour predicted by the different Principles is similar, but the justification that each of those Principles gives for those be-

\[\text{See Chapter 3 for further justification of this decision.}\]
behaviours is subject to variation. Essentially, researchers analysing dialogue are viewing the same 'output', but disagree on the 'input' which produces it. Therefore, we are mostly investigating the motivations for dialogue choices, rather than the behaviour itself. This suggested difference in motivation is particularly noticeable in those hypotheses which predict changes in interactant behaviour over time. For example, all the Principles suggest that dialogue pairs will improve their task result as they gain experience with the task, but they vary in their explanation of why this improvement should happen.

The data for this investigation consisted of dialogues coded with the Typology of Move Attributes, and a series of task success scores for those dialogues. This gave a profile of dialogue behaviours which could be compared with the level of task success achieved. The design of the Map Task Corpus also means that the performance of the dialogue pairs could be compared over time (see Section 4.2), which is important to our empirical study.

As far as we are aware, this research is the first attempt to empirically test a series of Principles in this way. The various notions of Cooperation and the notion of Parsimony are based on intuition rather than empirical work. In the case of Coordination and Collaboration, Clark and his co-workers do develop their work on an empirical basis, but they do not compare the predictions of their Principles with those of other researchers.

**Support for Principle of Parsimony**

The empirical tests carried out provided strongest evidence for the Principle of Parsimony, and the Principle of Least Individual Effort. Parsimony, and its associated mechanism, the Risk-Effort Trade-Off, had up to this point been based on intuition rather than dialogue data. Therefore, this study has provided new empirical support for this Principle.

It has also provided evidence against traditional notions of Cooperation and some aspects of Collaboration (at least in terms of justification). The Principle of Coordination was argued to be a descriptive, rather than motivational, principle. Grice’s Cooperative Principle proved difficult to define in empirical terms, because of its essentially philosophical nature. However, we suggest that is is closest in underlying motivation to the Principles of Parsimony and Least Individual

---

5It should be noted that both these systems (the Typology, the task success measure) have been tested for their reliability. This is reported in Chapter Four.
Effort. It can certainly be distinguished from the type of behaviours predicted by traditional notions of Cooperation.

In summary, our experimental evidence suggested the following:

Speakers attempt to minimise their effort, and in doing so, they make errors because of the risks they take. They are also not necessarily helpful towards their fellow interactants: speakers appear to only invest the effort that they believe to be necessary. If your partner requires lower effort than you are willing to invest, then you are likely to decrease your effort level. However, if your partner is requesting a higher effort investment, which you do not believe is necessary, then you will be unlikely to increase your input. Therefore, speakers will adjust their effort, but are more likely to adjust downwards rather than upwards. This effect is demonstrated in Wilkes-Gibbs’ (1986) empirical data, which we discuss in Sections 3.2.3.2 & 3.2.3.3.

However, in terms of performance, task success does improve. From the evidence that our study has generated, we would argue that this improved level of communication is due more to speakers working out what is necessary to the task (what they cannot afford to leave out) rather than their using a greater level of ‘effortful’ and ‘helpful’\(^6\) utterances.

### 1.3 The Organisation of the Thesis

This thesis takes philosophical concepts which have been used to describe dialogue, and tests them in an empirical environment. The aim is to investigate whether any of the suggested concepts (Cooperation, Coordination, Collaboration, Parsimony) are effective in modelling human dialogue. In order to undertake this study, several tools are needed:

- Practical Definitions of Dialogue Principles
- Information about dialogue behaviours

The development of these, and their place in the thesis is described below.

\(^6\)This is one way of defining cooperativeness, and it is the method we use here. The more effortful and helpful an utterance, the more cooperative it is deemed to be. See Section 3.3 for a justification of this.
Chapter Two  This Chapter provides a review of the relevant literature. We describe the pertinent aspects of research from philosophy, ethnomethodology, linguistics, psycholinguistics and computational linguistics.

Chapter Three  We move on from the general review to a more detailed evaluation of each of the Dialogue Principles with which we are concerned (Cooperation, Coordination, Collaboration, Parsimony). From these discussions, a series of empirically-testable hypotheses are developed for each of the Principles. These represent the behaviours which we believe the Dialogue Principle in question would predict in a task-oriented dialogue. This set of hypotheses represents our practical definitions of the Principles, and form the basis for the empirical tests described in Chapter Six.

Chapter Four  Here, we set out the methodology of the empirical study. Firstly, we outline our materials. This involves an account of our data (a subset of the HCRC Map Task Corpus), the design of the Corpus itself, and a description of its appropriateness for this current work. Secondly, we report the experimental work which was undertaken to demonstrate the reliability of our research tools (Task success measure, Typology of Move Attributes).

Chapter Five  This Chapter describes the practical aspect of this project. It is the method of analysis described here which provides our information on dialogue behaviours.

First, we motivate the development of the Typology of Dialogue Attributes, and outline how the analysis method is used to 'code' data. Secondly, we describe each of the categories in the Typology, and illustrate their intended usage through examples. These are taken from our dataset, as described in Chapter Four.

Chapter Six  In this Chapter we test the hypotheses which were developed through the analysis of Dialogue Principles which we presented in Chapter Three. For each hypothesis, we justify the empirical test used, describe the methodology, and report the results.

At the end of this Chapter, we synthesise all these empirical results, and evaluate the performance of each of the Dialogue Principles in turn.
Chapter Seven  Here, we consider the conclusions which can be drawn from this work, and evaluate its contribution. Also, we suggest how this line of research could be developed, particularly in relation to the support of our current results through further investigation.
Chapter 2

Review of Approaches to Dialogue

2.1 Introduction

The purpose of this Chapter is to outline the differing perspectives on dialogue offered by various disciplines: philosophy, sociolinguistics, psycholinguistics and computational linguistics. The intention is not to provide a comprehensive review, but rather to locate the present work in a larger context. A more critical consideration of the literature relevant to the notions of Cooperation, Coordination, Collaboration and Parsimony is undertaken in the following Chapter.

2.2 Philosophical Approaches to Dialogue

In this Section, we will consider the work of Austin (1962), Searle (1969) and Grice (1975). Their research was developed within the context of a philosophical rather than a linguistic framework, but it has since been integrated into linguistics, and is the starting point for modern pragmatics.

2.2.1 Speech Act Theory

Austin’s work should be viewed in the context of contemporary approaches to language. Logical positivists (e.g. Bertrand Russell) saw most language as being debased or deficient in some respect: if an utterance was not analytic or empirically testable then it was dismissed as being ‘meaningless’. The logical positivists wished to improve language, and remove these problems. Austin, however, noted that even with these ‘deficiencies’, speakers seemed to be able to grasp the in-
tended meaning. His interest was in how we convey our intended meaning, even through all the ambiguities. Therefore, language is seen as a tool, a way of achieving things.

Speech Act Theory (Austin, 1962) itself is concerned with the meanings and effects of particular utterances in particular situations.

"...the total speech act in the total speech situation is the only actual phenomenon which, in the last resort, we are engaged in elucidating..."

Austin (1962:147, emphasis in original)

Austin's initial focus was on a particular set of utterances, called 'performatives’, which performed an action merely by being uttered:

1. I name this ship the HMS Cardiff
2. I pronounce you man and wife

These ‘performatives’ would only take effect if the conditions for their use were fulfilled. For example, you cannot name a ship if you are not the appointed namer, the ship already has a name, or there are no witnesses, and so forth. These conditions which had to be fulfilled by the utterance and the situation were called Felicity Conditions.

However, Austin found that making a division between performatives and constatives (statements) was untenable. Speakers perform actions using language without using performatives, and conversely speakers use what might be considered to be performative verbs, but the associated action isn’t achieved. Therefore, the idea of ‘performing’ was generalised to the concept of illocutionary force: the speaker’s intention in uttering the act. The other aspects of the utterance (its form, and its effect) were characterised in terms of locutionary and perlocutionary acts.

Thomas (1995:49) summarises these three simultaneous acts as follows:

**Locution** the actual words uttered

**Illocution** the force or intention behind the words

**Perlocution** the effect of the illocution on the hearer

11
For example, the utterance "You can't do that!" could variously have the illocutionary force of forbidding or protesting, depending on the context, but the perlocutionary effect could be to check the addressee's action, annoy them, or even incite them.

From our point of view, the important aspects of Speech Act Theory are the move towards an explanation of natural language, and the recognition that language is not necessarily predictable. Speakers intend certain actions, but the resulting action is dependent of the reaction of the Hearer.

2.2.2 Searle - Speech Acts

Searle (1969) developed Austin's work on Speech Acts, and attempted to formalise Austin's approach. Searle's aim was to describe a rule-based system, similar to the systems which have been developed for other levels of language, such as syntax. He believed that it would be possible to create a set of rules for each speech act, which would differentiate that act from all others.

He presented the rules for the act of promising:

**Propositional act** Speaker (S) predicates a future act (A) of Speaker (S).

**Preparatory condition** S believes that doing act A is in H's best interest and that S can do A.

**Sincerity condition** Speaker intends to do act A

**Essential condition** S undertakes an obligation to do act A.

Thomas (1995:94)

A promise is a promise if and only if all these conditions are satisfied: the act of promising would then 'fire'. The description is based both on the Speaker's intention and the Hearer's wants. If an interactant 'promises' to do something which they know not to be possible, then the act would not fire because the preparatory condition would not have been fulfilled. Similarly, if a Speaker 'promises' to come to a party that evening, but the Hearer doesn’t want them to come, then the speech act would not meet the preparatory condition, and would therefore fail - the Hearer might consider this to be a 'threat' rather than a 'promise'.

This approach is problematic because it has proved difficult to define sets of rules which described anything more than the straightforward, typical examples.
of speech acts. The descriptions cannot account for some acceptable examples of speech acts, yet allow some blatantly unacceptable ones. Attempts to avoid these problems lead to a top-heavy system of ad-hoc preparatory conditions for each speech act, many of which are circular rules because they rely on the speaker’s knowledge of the differences between speech acts (e.g. compliment versus congratulate)

**Propositional act** Some event, act, etc., E related to H.

**Preparatory condition** E is in H’s interest and S believes E is in H’s interest.

**Sincerity condition** S is pleased at E.

**Essential condition**

Counts as an expression of pleasure at E

[congratulate]

OR

Counts as a commendation of E or tribute to H

[compliment]

Thomas (1995:98)

Here, the definition is reliant on one’s understanding of expression of pleasure versus commendation/tribute. Therefore you must already understand the concept in order to interpret the definition of the speech act verb\(^1\).

However, Searle’s work is important because it highlights the concept of joint meaning: the effect of a speech act is dependent on both participants. The jointness of language production is a growing theme in current research on dialogue: see the discussion of Clark et al.’s work later in this Chapter (Section 2.4.2). It is also important to our own work, and is considered in relation to the discussion of Dialogue Principles in the next Chapter.

### 2.2.3 The Cooperative Principle and Maxims

Austin and Searle both considered the difference between what people say and what people mean - hence the definition of (in Austin’s terms) locution, illocution and perlocution. Grice’s (1975) concern was with how interactants negotiate

\(^1\)See Thomas (1995) for a comprehensive discussion of this.
that gap. How can a Speaker use an indirect speech act to generate a particular implicature and how can this be understood by the Hearer?²

The Cooperative Principle and associated maxims (quantity, quality, relation, manner) attempt to formalise the way in which we generate an implicature using an indirect speech act. Grice has tried to account for the ways in which an utterance can be indirect, and thus the strategies we might have for generating the intended meaning³. His basic argument is that if we all make the same assumptions - operate according to the same conventions - then a Speaker can safely judge what a Hearer will interpret from a given utterance in a given context.

**The Cooperative Principle** basically states that our utterances should be appropriate to the situation:

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

Grice (1975:45)

We generate implicatures by flouting one or more of the maxims, maybe to gain a particular effect. Our co-interactant must assume that we are operating according to the Cooperative Principle - that our utterance is appropriate to the circumstances - otherwise, such an indirect speech act would not be worth the effort of interpreting.

In the following Chapter, we consider the implications of Grice's view in more detail. In particular, we will focus on the relationship between Grice's Cooperative Principle and the notion of 'cooperativeness' in general (Sections 3.2.1 & 3.2.2).

### 2.3 Sociolinguistic Approaches to Dialogue

In this Section we will discuss the basic ideas behind the disciplines of Discourse Analysis and Conversation Analysis. In the terms of this research, these areas have been drawn on in the construction of the Typology of Move Attributes which is presented in Chapter Five.

²Although it should be made clear at this point that Grice's interest is with the generation of implicature, not with the process of understanding, nor the reasons for employing it.
³Grice does not really consider the effect of indirect speech acts: why we use them rather than straightforward language.
2.3.1 Discourse Analysis

Birmingham School Discourse Analysis (e.g., Berry, 1981b; Coulthard and Montgomery, 1981; Coulthard, 1994; Stubbs, 1983) was developed from a combination of Systemic Functional Linguistics and Speech Act Theory. Its theoretical basis was derived from the rank and scale view of language described in Halliday (1961): discourse was seen to be a higher level of language than the clause, but consisting of much the same type of structures. The basic units of structure in this case were based in part on the types of speech acts which Austin (1962) and Searle (1969) had identified. Halliday (1961) argues that the analysis of any level of language can be expressed through the interrelationship between analytical units presented in the terms of a rank scale. Each unit has its own internal structure, which is expressed by limited combinations of the units at the rank immediately below.

Discourse Analysts, in a similar way to Searle, argue that discourse is rule-based. Discourse has a grammatical structure like syntax: it is simply a question of correctly identifying that structure. They aim "... to describe conversation as a distinctive, highly organized level of language." (Taylor and Cameron 1987:5). Sinclair and Coulthard (Sinclair and Coulthard, 1975; Coulthard, 1994) were the first to attempt this in their model of classroom discourse. They proposed a hierarchy of five ranks to explicate discourse structure: lesson, transaction, exchange, move and act. Lessons consist of a series of transactions, and there does not appear to be any standard relationship between these transactions. The transactions themselves consist of a series of exchanges, usually beginning and ending with a boundary exchange. The structure of the exchange is the most formalised of the ranks:

\[
\begin{array}{ccc}
I & (R) & (F) \\
\end{array}
\]

An exchange structure prototypically consists of three moves: Initiation (I), Response (R) and Feedback (F). The bracketting around ‘R’ and ‘F’ indicates their optionality. Sinclair and Coulthard identified 22 different types of act which could make up the structure of these moves, many of which are limited to filling particular moves. For example, ‘accept’ or ‘evaluate’ acts typically fill Response moves.

This system has many problems. Firstly, it is too genre-specific - even the term ‘Feedback’ seems inappropriate to all but a pedagogical context. Also, there are
problems with the description of exchange structure. The IRF model is not very predictive, because it does not suggest when 'R' and 'F' elements are optional and when they are obligatory. Exchanges could be analysed using Sinclair and Coulthard's model, but it would make no distinction between those which seem acceptable ('well-formed') and those which seem odd ('ill-formed'). This is not in keeping with the notion of a discourse grammar.

A: Where is Paris?

For example, the above Initiation move requires a Response element, yet according to Sinclair and Coulthard's model, the Response move is not obligatory in this, or any other, situation. Coulthard and Brazil (1981) describe a model which has five possible elements to an exchange\(^4\), but as the Initiate move is still the only obligatory one, it suffers from exactly the same problems. Berry describes a model for elicit exchanges, based on whether the questioner knows the answer already, which is more predictive of the structure of the rest of the exchange (Berry, 1981a,b,c). But this is only one small aspect of exchange structure.

Other additions have been made to the basic model of exchange structure, such as 'challenge' exchanges (embedded exchanges) (Burton, 1981), and moves which have dual functions (Response-Initiate moves, Stubbs 1983), but although these have certainly broadened the coverage of exchange structure, they have not addressed its main failing. Exchange structure does not seem generalisable to casual conversation: moves can roughly be described as Initiate, Response and Follow-up (suggested by Coulthard and Brazil (1981) as an alternative to 'Feedback'), but they do not always occur in the order which Discourse Analysts suggest. It is also extremely difficult to locate exchange boundaries - a similar problem to that encountered by Grosz and Sidner (1986) when trying to define discourse segment boundaries. Constraining casual conversation (or even task-oriented dialogue) within these bounds means compromising on the labelling of moves. In this study, although we use similar definitions of moves to the IRF model, we do not attempt to impose a higher rank structure.

Another area of importance to our work is the issue of choice. Systemic linguistics is based in the idea of syntagmatic and, more particularly, paradigmatic choices. At the level of syntax, it uses system networks to enumerate the options open to a speaker at any given choice point - for example, the mood of a sentence, or whether to employ a modal verb. From the sociolinguistic/pragmatic point of view, the

---

\(^4\) I (R) (F) (F) (F)
importance of choice is surely the effect it has, yet Discourse Analysts do not seem to have considered this issue. If a Speaker chooses to use act X rather than act Y, then this will affect the course of the conversation (even if only slightly). Fawcett, van der Mije, and van Wissen (1988) and Davies (1991) have developed an initial system network (system of choices) for the level of dialogue, but again, they were trying to codify the structure, rather than consider the implication of particular decisions. This question of choice would seem to be an interesting one, although it is an avenue which has only been pursued in terms of its social implications (see the discussion of Conversation Analysis which follows), rather than its impact on the process of information transferral with which we are more concerned here. The dialogue coding system which we describe in Chapter Five attempts to address the question of dialogue choices, and their implications for the dialogue as a whole.

2.3.2 Conversation Analysis

Conversation Analysis has its basis more in sociology (ethnomethodology) than linguistics, and therefore has a quite different approach to Discourse Analysis. Conversation Analysts collect naturally occurring data, and try to find patterns within that data. They do not claim to be seeking an overall analysis, unlike Discourse Analysis, because they do not believe that conversation has such a quantifiable structure. Conversation Analysts are opposed to the idea of codifying a grammatical structure for dialogue. Their belief is that language is structured according to social principles rather than grammatical ones, therefore they consider the search for recurring patterns in data a more suitable approach (Levinson, 1983). Their interest is focused more on why speakers make certain choices about utterances rather than the discourse moves which are being used. The importance of any patterns of utterances is the impact which it has on the social role of the speaker - not its effect on the process of information transferral.

Conversation Analysis has identified several different patterns which occur in speech, the most well-known of which is that of adjacency pairs. These occur in the sequence first part - second part, which is roughly equivalent to the notion of an Initiation-Response sequence in the terms of Discourse Analysis. There are various types of adjacency pair, such as question-response, greeting-greeting, summons-answer etc, which occur in a wide range of conversation types, for unlike Discourse Analysis, Conversation Analysis is not usually genre-specific.
The importance of negotiation is also emphasised. For example, Sack et al.’s description of the highly-organised mechanism of turn-taking and how it is managed by both interactants, and the claimed importance of carefully negotiated closing sequences to the interpersonal success of a conversation (Schegloff and Sacks, 1973; Button, 1987). Conversation Analysis, then, concentrates more on how the dialogue is achieved. It views conversation as being negotiated by all of its participants: they work together to produce the discourse.

2.3.2.1 Turn-taking

Arguably, one of the most important contributions of early Conversation Analysis was the consideration of turn-taking. In conversation, there is a degree of overlapping speech, but Sacks, Schegloff, and Jefferson (1974) suggest that this overlap is brief, and accounts for less than 5% of the speech. In general, therefore, conversation is divided into turns where the interactants rotate the role of Speaker. The current Speaker is described as having ‘the floor’.

One of the important issues around turn-taking is how the changing of turns is negotiated. Turns do not have a pre-determined length, neither is the next speaker necessarily fixed. Sacks, Schegloff, and Jefferson (1974) propose two structures which we utilise to manage this process: Turn Constructional Units (TCUs) and Transition Relevance Points (TRPs). The turns themselves consist of one or more TCUs which are grammatically complete units (e.g. sentence, clause or phrase). The end of each TCU is termed a TRP, and it is considered to be a point of potential speaker change. The selection of the next Speaker is either fixed by the current Speaker (for example, a question addressed to a particular person), or another interactant can self-select. This process of speaker-change has to be negotiated at every TCU - either a change of speaker will occur, or the current speaker must make it clear that they do not intend to surrender the floor. Any change of speaker must be settled at that TRP; it cannot be fixed in advance - or, at least not in casual conversation.

2.3.2.2 Adjacency Pairs and Sequencing

Adjacency pairs are an important notion because they are highly predictive of what type of utterance the respondent will utilise, yet they do not proscribe any combination. Often, a first part could have several potential second parts, some of which occur more frequently than others. Atkinson and Drew (1979) introduce the concept of ‘preference organisation’ which handles the idea of there being
several potential ‘responses’. If an ‘offer’ is made, the preferred response would be an ‘accept’, the dispreferred response would be a ‘refusal’. There may also be other options, which might be rated somewhere on the cline between a preferred response and a dispreferred one, depending on the nature of the utterance. So adjacency pairs emphasise both the existence of choice and its importance.

It should also be noted that the reasoning behind preferred vs dispreferred refers to the interpersonal effect on the original speaker, not to either structural aspects of the dialogue nor the process of information transferral. Brown and Levinson (1987) consider these interpersonal issues in greater detail in their study of politeness and face which is developed, in part, from Goffman (1967). In our work, the emphasis on the interpersonal and transactional aspects of the dialogue is more equal.

Along with adjacency pairs, Conversation Analysts have also identified sequences. This is to account for there being a series of utterances which are related. The adjacency pair structure was seen to be too limiting as it only accounted for two-part structures. Sequences which have been suggested are:

**insertion sequences** (Schegloff, 1972). These are essentially what Discourse Analysts would term embedded exchanges: often a query, followed by an answer, after which the predicted second part of the original adjacency pair would occur.

**side sequences** (Jefferson, 1972). These are longer sequences embedded between two parts of an adjacency pair, which are essentially ‘off-topic': they are are an *aside*

**closing sequence** (Schegloff and Sacks, 1973). This is the section of a conversation which negotiates its closing.

**repair or clarification sequences** (Schegloff, Jefferson, and Sacks, 1977). These are longer sequences of clarification or repair, and do not generally occur between the first and second part of an adjacency pair.

Indeed, sequence is an important notion in Conversation Analysis. Sequential relevance is described in Sacks, Schegloff, and Jefferson (1974) as a general principle underlying the organisation of conversation. Conversations make sense because

\footnote{These are not pursued further here, for as Leech (1983) suggests, issues of politeness rarely occur within the context of a task-oriented dialogue.}
we interpret them in sequence. Eggins and Slade (1997) summarise this in the following way:

"... the speaker’s current turn will be interpreted as implicating some action by the responder in the immediate next turn. Similarly, the respondent’s subsequent talk will, where possible, be interpreted as related to the immediately prior turn."

Eggins and Slade (1997:29)

This idea of a turn being interpreted in relation to its predecessor, however tenuous the connection may be, can be seen as a replacement for Grice’s Cooperative Principle. Indirect utterances can be understood in this way: we look for an interpretation because as the two utterances are adjacent, we assume they must be related. In the discussion of our dialogue analysis method (Chapter Five), we take up this notion of implication as representing a speaker’s duties towards their fellow interactant.

2.3.2.3 Conversation Analysis and the Quantitative Approach

In terms of empirical analysis, Eggins and Slade (1997) argue that Conversation Analysis has some drawbacks. It lacks systematicity: there are no exhaustive accounts of adjacency pairs, nor any description of how you might recognise them. Neither do they define exactly what constitutes a TCU, and how you might locate such a boundary. This is a problem if you wish to undertake quantitative research, and Eggins and Slade (1997) argue that in order to describe what is happening in conversation, this is precisely the approach that you need. The same type of problem is caused by the fragmentary focus of Conversation Analysis. This approach tends to focus on the micro-structure in which it is interested, and ignore the rest. Again, this is not very useful when trying to analyse the larger structures in dialogue. However, it is possible to apply some of the insights from Conversation Analysis to a more quantitative approach, which is what we have attempted to do here.

In this study of language, we are interested in motivation, but we are concerned with the issue of task success too. Our analysis of dialogue concentrates as much on issues of information transferral as interpersonal dynamics (although these two are intertwined). Conversation Analysis is concerned solely with interactional genre, whereas our interests also extend to the transactional (Brown and Yule,
Conversation Analysis focuses more on the social role of conversation than how we achieve goals through language. Therefore, we have applied some of the notions from this social approach, such as sequence (implication), choice and preference organisation, to more transactional accounts of language (see Chapter Five).

2.4 Psycholinguistic Approaches to Dialogue

In this Section we will consider the research of Anderson et al (Anderson, Clark, and Mullin, 1991b; Anderson and Boyle, 1994; Anderson, Clark, and Mullin, 1994), Clark and his co-workers (Clark and Wilkes-Gibbs, 1986; Clark and Schaeffer, 1987a,b, 1989; Schober and Clark, 1989; Clark and Brennan, 1991; Wilkes-Gibbs, 1986; Schober, 1995) and Garrod et al. (Garrod and Anderson, 1987; Garrod and Doherty-Sneddon, 1994). Anderson’s work is important to the research presented here because it addresses the question of task success in task-oriented dialogues, and considers what aspects of a dialogue might affect task success. Clark’s work is concerned with the issue of how speakers produce language and achieve things with language. Garrod focuses on the concept of semantic coordination, which relates to Clark’s claims for both the Principles of Coordination and Collaboration.

2.4.1 Task Success in the Map Task Corpus

Anderson and her co-workers are primarily interested in what aspects of dialogue contribute towards effective communication. In other words, how do speakers achieve task success - what types of talk correlate with a good task result. Their interest in this aspect of dialogue research is also coupled with an interest in child acquisition of such communicative skills. As this part of their work is less relevant to our concerns, we will concentrate on their investigation of task success and communicative competence.

Anderson et al. undertook two studies, one involving children, the other involving adults (Anderson, Clark, and Mullin, 1994; Anderson and Boyle, 1994). Both studies used the Map Task to elicit task-oriented dialogue from both pairs of children and pairs of adults 6 (Brown et al., 1984). This task basically involves two participants: a Giver and a Follower. The Giver has a route drawn on her

6The Map Task Corpus, which is also the source of data for this study
map, the Follower does not. The Giver describes her route to the Follower, and the Follower attempts to replicate it on his map. The maps differ in the landmarks (little pictures) that they have: about half occur on both maps, the remainder are unshared. The Speakers are told that there may be differences in the maps. The Map Task is described more fully in Section 4.2; Figures 4.1 & 4.2 each show an example of a pair of maps.

The dataset of child dialogues showed there to be differences in task success according to age, and the use of certain strategies in the dialogue. There were three age groups: 7, 9 and 13 year olds (exact ages in Anderson, Clark, and Mullin (1994)). The 7 year olds produced statistically worse maps than the 9 and 13 year olds (there was no difference between these two age groups). But, there was more difference within the age groups than across age groups.

Anderson found that a number of aspects of the dialogue correlate with task success:

**Introduction of Features** Dialogue pairs who introduce features into the conversation in the form of a question (question introductions) in order to check whether they are known to the other speaker tend to produce better drawn routes.

**Asking and Answering Questions** The more questions that are asked in the dialogue, the better the task result. The fewer questions which go unanswered, the better the task result.

**Discovering Problem Points** The more differences that are discovered in the dialogue - e.g. when features aren’t shared - the better the task result.

These were all tested *individually* against task success, rather than as a composite ‘score’ of some kind. However, when a stepwise multiple regression was performed in order to assess the relative importance of each of these strategies, it was found that the discovery of problem points and the proportion of questions answered had the most impact on task success. The use of question introductions did not independently add significantly to the predictions of the equation. This is probably due to the overlap between problem points and feature introductions: the features are the differences which they need to discover.

Anderson and Boyle (1994) took a subset of dialogues from the HCRC Map Task Corpus (Anderson et al., 1991a). This involved first year undergraduates from the University of Glasgow; the basic task was otherwise the same. Here, again, the
use of question introductions was correlated with task success. The importance of responses to questions was also demonstrated. Anderson and Boyle (1994) showed that the quality of the Addressee’s response was also correlated with task success. Responses following the introduction of unknown features were categorised as INFORMATIVE, INADEQUATELY INFORMATIVE or UNINFORMATIVE. Not only was the use of INFORMATIVE responses correlated with task success, but also the use of question introductions was more likely to elicit INFORMATIVE responses from the Addressee. Anderson, Clark, and Mullin (1991b) found a similar effect with child participants in the Map Task.

These results are particularly interesting because not only do they investigate the idea of speaker choice and task impact, but also because they find an interaction between their two contributing factors. The Speaker can encourage an effective response from the Addressee by choosing their method of feature introduction appropriately. Equally, they can also choose what might seem to be less effective approaches.

The research we describe in this thesis takes a similar approach. We too consider the impact of particular choices on task success. However, we do not just concentrate on the introduction of new features into the dialogue. Our categorisation of choices attempts to produce a profile of the whole dialogue. We consider the relationship of Anderson et al.’s categories to our own in Sections 5.4.2, 5.5.5 & 5.6.3.

2.4.2 Clark’s Collaborative Model

Clark and his co-workers have developed a model of dialogue which draws on some of the structures in conversation analysis. His theory of contributions is rooted in Schegloff, Jefferson and Sacks’ (1977) analysis of repairs (Clark and Schaefer, 1987a). However, Clark’s research focuses more on the issue of mutual knowledge, the importance of what he terms grounding. His model considers how interactants negotiate information, and how they know when that information is mutually understood7.

Clark concentrates on language as joint action: dialogue is not simply a product of the Speaker’s contribution and the Hearer’s contribution, it is a sum of more than its constituent parts. In joint actions such as language, interactants use participatory acts rather than autonomous ones. Participatory acts only make

7See Clark (1996) for a comprehensive review of his theory of language as joint action.
sense in the context of the joint action of which they are part, they are not reducible to a chain of autonomous acts.

The metaphor of a piano duet is used to explain this difference. If A plays their part on their own (for practice, say), then although the actions are physically identical to when that part is played with the other pianist, at some level of description they are in fact different acts. They have different intentions: the intention to play a piece versus the intention to play the same piece as part of a larger whole.

The empirical evidence for this claim is based on Clark’s investigation of contributions to common ground, and the importance of collaboration to common ground. Clark argues that common ground is the basis of conversation - we can only base a conversation on what we share, and the goal of any conversation is to extend what we share, and thus add to our common ground. Therefore, we must tailor our utterances to the (perceived) common ground of our fellow interactants. Clark and Schaefer (1987b) demonstrate this accommodation by setting up a task where the interactants are asked to conceal the information they are exchanging from an overhearer. The interactants were pairs of friends, the overhearer a stranger, and the information they were attempting to exchange was the identification of a number of well-known Stanford (and therefore local) landmarks. The interactants attempt to conceal the information they are exchanging by appealing to private common ground. For example, the Director might refer to a particular occasion or event associated with this location that only the Matcher (and a few others, perhaps) would have been party to. In the control condition (where the interactants were not aware of the overhearer, and thus did not attempt concealment), the Director and Matcher did not bother to use private common ground, but rather relied upon descriptions which any Stanford resident was likely to recognise. Clark argues that this adjustment of strategy shows that speakers are aware of common ground, and are able to accommodate their use of it appropriately.

If we are so sensitive to common ground, Clark argues, surely we must also be equally aware of the pitfalls in attempting to extend that shared knowledge - merely to assume that an utterance has automatically been heard and understood would be foolhardy. Clark and Schaefer (1987a) claim that common ground is only added to through the collaborative process of contributing, rather than the Hearer individually decoding the utterance and interpreting it. A contribution

---

8What might also be termed 'mutual knowledge'.

24
has two phases: a presentation phase and an acceptance phase. A presents an utterance for B to consider, on the assumption that if B provides sufficient evidence of understanding, then A will believe that B understands the utterance. B accepts A’s utterance by giving sufficient evidence of understanding, under the assumption that A will then believe that B understands A’s utterance. This process is iterative, and will continue until sufficient evidence has been given in the acceptance phase for the original speaker to consider understanding to have been achieved. The concept of ‘sufficient evidence’ is defined in terms of the strength of evidence principle, which governs how much evidence is necessary for accepting a presentation. This is defined as:

**Strength of evidence principle:**
The participants expect that, if evidence $e_0$ is needed for accepting presentation $u_0$, and $e_1$ for accepting the presentation of $e_0$, then $e_1$ will be weaker than $e_0$.

Clark and Schaefer (1989:268)

Once a presentation has been accepted, it is considered to have been grounded: added to the interactants’ common ground9 (Clark and Schaefer, 1989).

Clark and Schaefer (1989:262) use the following example form the London-Lund corpus:

A: F . six two
B: F six two
A: yes
B: thanks very much

B accepts A’s contribution of a book number by repeating it verbatim. A then accepts B’s contribution by the weaker evidence provided by the utterance “yes”. This is in turn accepted by B’s second turn, which proceeds to the next contribution. Clark argues that this is weaker evidence because it is a new contribution: the previous contribution is accepted by the current speaker intimating that there is no more that needs to be said.

---

9*Grounding criterion:* The contributor and the partners mutually believe that the partners have understood what the contributor meant to a criterion sufficient for current purposes. (Clark and Schaefer 1989:262).
This element of Clark’s approach to Contributions does seem to be problematic. We would suggest that even in the example that Clark himself cites, the concept of the strength of evidence principle is not immediately obvious. However, speakers do manage to complete these ‘sections’ of talk. The mechanism, we would suggest, is not quite so clear.

For a further example of this phenomenon, we use an extract from a Map Task dialogue, where the Follower is a normal speaker\textsuperscript{10}, and the Giver is an aphasic speaker.

\begin{quote}
Capital letters denote loud speech  
Underscores (_ ) denote quiet speech 
\end{quote}

\begin{verbatim}
1 Follower: So I just stop next to the mountains  
next to the lake?
2 Giver: Mm!
3 Follower: Mm. ... That’s it ...
4 Giver: And er rocks and swamp
5 Follower: Mm
6 Giver: Noth- no really nothing about it
7 Follower: No we don’t go there
8 Giver: NO
9 Follower: _No_
10 Giver: _No_
11 Follower _No_
12 Giver: And er tree dead really nothing as well
13 Follower: Mm
\end{verbatim}

Merrison (1998:458)

The contribution initiated in turn 4 and continued in turn 6 is followed by a series of turns which form the acceptance process (turns 7 to 11). The first of these is strong evidence of acceptance, the Follower paraphrases the Giver’s utterance. The second, strong but less so - the capitals denote loudness. The turns after that all demonstrate weaker evidence, by getting progressively quieter. Again,\textsuperscript{10}

\textsuperscript{10}By this, we mean she has no language disorder.
this does seem to follow Clark’s model, but it is difficult to explain when the acceptance process has demonstrated enough evidence of understanding for it to terminate. Why does it complete in turn 11? Why didn’t it complete in turn 9? Turns 10 & 11 do not seem to add much to the acceptance process.

There do not seem to be satisfactory answers to these questions in Clark et al.’s work, and as this area is outside the focus of this thesis we will not consider it further at this point. However, these theoretical issues are explored in more detail in Merrison (1998), and we direct the interested reader to this discussion. We will return briefly to the question of acceptance termination when we consider Heeman and Hirst’s (1995) operationalisation of the refashioning process later in this section.

The other important aspect of the presentation/acceptance process is the role of the Addressee: Clark argues that the Addressee has an active role to play, and is not just the receiver of a message which needs to be simply decoded. Schober and Clark (1989) show the importance of the Addressee’s role by demonstrating a difference in understanding between a dialogue participant (Addressee) and a non-participant (Overhearer). In a tangram\textsuperscript{11} matching task, a Director described a figure to a Matcher, who had to identify the figure correctly in order that he could place his tangram figures in the same order as the Director’s (see Clark and Wilkes-Gibbs (1986) for a full description of this task). The Overhearer was seated the same distance from the Director as the Matcher, and had access to the same information. The presence of the Overhearer was explained in terms designed to avoid his presence affecting the linguistic behaviour of the other pair: he was not to be considered part of the conversation, but irrelevant to it\textsuperscript{12}. It was found that Matchers did significantly better than Overhearers, even though they had access to precisely the same linguistic information\textsuperscript{13}. The only difference between them, Clark claims, is the Matcher’s ability to control the presentation/acceptance process. The contribution is grounded when the Matcher says so, not when the Overhearer is necessarily ready. Therefore, this demonstrates that the process of information transferral is not simply a question of utterance

\textsuperscript{11}Tangrams are composite shapes, which consist of a number of triangle-shapes put together to create an overall figure which could be perceived as different things by different people. It is up to the participants to generate a referring expression which will disambiguate a particular figure for both of them. Clark and Wilkes-Gibbs (1986) reproduce the set of tangram figures which they used.

\textsuperscript{12}For convenience, Directors are referred to as being female, Matchers, male and Overhearers, male.

\textsuperscript{13}Matchers improved from 95% to 100% correct over 6 trials, Overhearers improved from 78% to 89% over the same period.
interpretation, otherwise the Overhearer should have been as successful as the Matcher. Rather, the Addressee (Matcher) has an important, active, role to play in managing the acceptance process.

In the context of referring expressions, this presentation/acceptance process can be instantiated in terms of 'refashioning': participants trade referring expressions in an attempt to identify a particular tangram figure out of a group (Clark and Wilkes-Gibbs, 1986; Clark and Schaefer, 1989). The trading of new suggestions, or alterations to the offered expression will continue until the participants are convinced that they agree on which figure is being referred to. The example below is taken from Clark and Wilkes-Gibbs (1986:22):

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.

Traum (1994) points out that this description of the presentation/acceptance process is problematic for two reasons. Firstly, Clark’s system suffers from the same problem of all the other analysis systems we have discussed so far: it is very difficult to define when an utterance fits into a particular category. Secondly, if, as Clark and Schaefer (1989) suggest, each utterance is a presentation which needs to be accepted (including acceptances), then how can this process ever ‘bottom out’ and finish? Clark and Schaefer (1989) later claim that there are some acceptances don’t need accepting themselves, but they do not clarify how one is supposed to distinguish these two types.

In their computational model of collaborating, Heeman and Hirst (1995) attempt to solve the second of these two problems. They describe the presentation/acceptance process as having three moves: present, judge, refashion. Here, the refashioning element is a coded part of the structure, rather than being a descriptor as used by Clark and Wilkes-Gibbs. The idea of acceptance is changed from being a move in the process to being a boolean value which controls the process:
This would seem to be a useful extension to Clark's terminology, as it makes it clear how the acceptance process should terminate. As Heeman and Hirst (1995) point out, it also emphasises the fact that both participants have access to the judging and refashioning process. In Clark and Wilkes-Gibbs' description, it could seem that the Director judges and accepts, whilst the Matcher refashions, whereas both speakers have access to these two roles. Heeman and Hirst's algorithm separates the process from the particular participants, thus focusing on the joint responsibility to produce an effective referring expression.

This refashioning behaviour can be seen in Map Task Corpus data (which is the data we consider in this thesis), where interactants trade descriptions of features or route-parts. See Section 4.2 for a full description of this data.

2.4.2.1 Coordination

Clark (1985) claims that we continually coordinate our language use with our talking partners. For example, when we try to get past someone in a narrow space, we have to work out (coordinate) who will go first. Similarly, if two people wish to arrange a meeting, they have to coordinate (agree on) where and when they will meet. In language, we manage the Contribution process by coordinating the state of knowledge with our partner's knowledge (common ground). We coordinate who speaks when (turn-taking), and we use coordination in order to interpret other speaker's utterances. Clark, in essence, states that all levels of our language use are guided by the Principle of Coordination.

Drawing on Lewis (1969), Clark distinguishes between three different types of coordination: convention, salience and precedence. He explains the difference between these three types of coordination by analysing the problem of two people meeting (from Clark 1985:182). If the two speakers are rational, then they should go where they mutually expect the other person to go.
explicit agreement Ann and Bob agree to go to the library, therefore they have good grounds for mutually expecting the other to go to the library.

salience Ann visits Bob’s home town, and they both know she has only been to Bob’s house in that town before. Therefore, this is the most obvious choice for mutual expectation.

precedence The last time Ann and Bob met, they went to the City Hall, so that is the most likely place to meet this time. (This is a special case of salience).

convention Ann and Bob always meet at the coffee shop, so they mutually expect the other will assume that is to be the meeting point.

Lewis (1969) argues that conventions are what evolves from coordination problems that recur. For example, in Britain, it is the convention to drive on the left. In America, the reverse is true. What is important about this behaviour is that:

- Everyone keeps to the convention.
- Everyone expects everyone else to keep to the convention.
- Everyone would have kept to an alternative convention, if a different regularity had evolved.

In terms of the driving metaphor, what is important is that we all agree on (or coordinate) which side to drive on. The side chosen doesn’t matter.

Language can be seen as a system of conventions: the arbitrariness of the sign is the perfect example of this. There is no reason why the word cat should connote the meaning ‘cat’. The English language has evolved this convention, the Welsh language has evolved a slightly different one: cath for ‘cat’. The precise word used does not matter, it is the fact that we agree that there is a relationship between the sign and the signified which makes that relationship exist.

However, Clark argues that the other aspects of coordination (salience and precedence) are also found in language use. How else do we interpret utterance which could have many meanings? For example, the phrase The World Cup final is on this evening could be taken to mean:

- Neither of us like football, lets go out instead
• Neither of us like football, so let's get a video instead
• If there's a big match on, it's not a good idea to go to the pub
• You like football, so I assume you will be watching the match
• I like football, so you can assume I'll be watching the match
• We both like football, so we'll stay in and watch it
• We both like football, so we'll go to the pub and watch it

The meaning which an Addressee will extract from this utterance will depend on the common ground which they share with the Speaker - its salience. Precedence can be used to interpret, for example, some forms of ellipsis. Take the pair of utterances:

1. *The World Cup match is on this evening.*
2. *How soon?*

The elided element can be re-instated by appealing to what had precedence in that exchange. The topic of this exchange is the football match, this topic therefore has precedence, and thus the original Speaker can interpret what the elided element should be. The concept of precedence can also be applied to the type of indirect responses which Grice explains via the Cooperative Principle:

1. *I've run out of petrol.*
2. *There's a garage just around the corner.*

The precedence here is that the response should relate to the topic of the first utterance. It is in this way that the utterance can be interpreted.

### 2.4.2.2 Collaboration

If language is a joint production, managed by all participants, then, Clark argues, there should be something special about being recognised as a participant, rather than just being exposed to the information arising from the interaction. Mutual knowledge requires the process of acceptance, not just the broadcasting of information. This argued importance of being a ratified participant in order to be
considered as sharing common ground is investigated in Wilkes-Gibbs and Clark (1992). A tangram task was set up, where there were overhearers. The overhearers were of three types: side-participants, omniscient bystanders and simple bystanders. Side-participants were sat at the side of the Matcher, omniscient bystanders watched the dialogue over a video screen and simple bystanders were in the same room as the conversation, but could not see the diagrams. The overhearers observed six trials, before becoming a Matcher for a further six, with the Director whom they had been observing. As a control, a naive matcher (not exposed to this task previously) also undertook the second six trials, so that a direct comparison could be made between their performance and the overhearers'. In Clark and Wilkes-Gibbs (1986), they had shown that in such tangram tasks, the referring expressions used by the participants, and the length of the dialogues decreases over the trials as the dialogue pair increases their common ground. The issue here was the effect of being an overhearer on this process: would the overhearers already share the common ground, and thus immediately use the shorter referring expressions and shorter contributions, or would they behave like a 'new' participant (naive matcher) and show the same gradual decrease in referring expression length and contributions?

The three types of overhearers showed different results. Side-participants showed the effect of their previous experience (shorter, faster dialogues), the omniscient bystanders were slower, but the simple bystanders were slowest of all: there was no difference between their performance and the naive matchers. Therefore, there is a difference in the pair's perception of mutual knowledge according to whether the Matcher was actually present at the previous conversations - this is the only difference in the position between the side participant and the omniscient bystander. Wilkes-Gibbs and Clark (1992) argue that this result demonstrates the joint production of language: the achievement of mutual knowledge is only possible through an acceptance process engaged in by all participants (even if that engagement is non-verbal). In Clark's terms, the side-participant was a Ratified Participant\textsuperscript{14}. The Director recognises the side participant as being part of the previous conversations, and therefore party to the Common Ground which the Director and previous Matcher built up. The omniscient bystander was not physically present, so is not considered by the Director to have been part of the

\textsuperscript{14}It should be noted that it is not co-presence which is important, but being accepted as part of the conversation. Overhearers were co-present in the experiment described in Schober and Clark (1989), but neither the Director or Matcher considered them to be part of the conversation, so they were not party to the acceptance process: the maintenance of their common ground was not important.
acceptance process. Therefore, the Director assumes less Common Ground when she has a former omniscient bystander as a Matcher, than if she has a former side-participant.

Therefore, having access to the information transferred is not enough in itself to achieve mutual knowledge. However, it should also be pointed out that access to all this relevant information would appear to be an important factor, as the simple bystanders were slower than the omniscient bystanders even though they were physically present. Neither type of bystander was a Ratified Participant, but the omniscient bystanders seem to benefit from having access to the experience of the previous trials. Therefore, both this experiment and Schober and Clark (1989) (involving co-present Overhearers) show that non-Ratified Participants can understand a fair amount even without being part of the acceptance process. This does not necessarily weaken the claim for the existence of Collaboration between participants, but it does perhaps question its relative importance in the communication process.

The collaborative theory also makes the claim that speakers are parsimonious in the effort that they jointly invest. Clark and Wilkes-Gibbs (1986) found that in their tangram task, referring expressions shortened, and the number of turns needed to identify a figure also decreased; participants had increased common ground, thus they needed to say less in order to identify a figure, thus they did say less. Schober (1995) argues similarly, this time for the use of locative expressions in a task where the Director and Matcher have a different point of view. Most speakers gravitated towards using locative expressions which were both-centred (i.e. were interpretable regardless of point of view) rather than expressions which were either Speaker-centred or Addressee-centred. He argues that Speaker-centred locatives would be low effort for the Speaker, but high effort for the Addressee; Addressee-centred locatives would be low effort for the Addressee, but high effort for the Speaker. Both-centred ones, however, should be low effort for both. This behaviour is explained by the Principle of Least Collaborative Effort, which states that participants will use the least joint effort possible to achieve their goal (Clark and Wilkes-Gibbs, 1986; Clark and Schaefer, 1989; Clark and Brennan, 1991; Schober, 1993, 1995; Wilkes-Gibbs, 1986).

**Principle of Least Collaborative Effort:**

"Participants in a contribution try to minimise the total effort spent on that contribution - in both the presentation and acceptance phases."

Clark and Schaefer (1989:269)
Wilkes-Gibbs (1986) demonstrates this importance of effort in relation to the Grounding Criterion: that the Addressee has understood what the Speaker meant to “… a criterion sufficient for current purposes.” (Clark and Schaefer 1989:262). She set up a task in some ways similar to the Map Task Method used to elicit our data. Her participants each had a map of the same city centre. The map was divided into squares, 18 in total, 9 of which were blanked out on each map. Therefore, the participants had to pool their available knowledge to achieve their set task, much as in the Map Task. The participants were given one of two tasks:

**High Criterion (HC)** Be able to describe the route from A to B to someone intended to drive between those two points

**Low Criterion (LC)** Be able to estimate how long it would take to drive the route between A and B at 1pm

The pairs of interactants were either matched pairs (both HC or both LC) or mixed pairs (one HC interactant, one LC interactant). The participants were not told of these two conditions, and presumably assumed that their partner had the same remit as their own.

Wilkes-Gibbs’ aim was to test whether there was a difference in degree of collaboration between the HC and LC matched pairs, and also to see what effect the contrasting goals of the mixed pairs would have on their collaborative behaviour. This was tested in two ways. Firstly, after the pairs had completed this task on three city centre maps, they were asked (as individuals) to replicate the route A-B on a full version of each of the maps. These maps were graded according to their accuracy. It was found that there was no significant difference between the performance of the LC and HC pairs. However, these matched pairs had achieved a significantly better result than the mixed criterion pairs (HC participants in mixed pairs did particularly badly). Secondly, the amount of talk which the pairs used to complete the task was compared. HC pairs talked significantly longer than LC pairs: 11.5 minutes on average, versus 7.5 minutes on average. Mixed pairs used about the same amount of talk as LC pairs.

It was concluded from these results that the degree of collaboration does indeed vary according to the current purposes: HC pairs invested more time and talk to ensure understanding (although the degree of understanding achieved did not appear to be greater). Also that the Principle of Least Collaborative Effort was adhered to because the mixed pairs had oriented towards the requirements of the
lower effort goal (LC) rather than the higher effort one of the HC speaker. However, we would argue that this interpretation is simplistic: there are alternative (and more convincing) explanations of the participants’ behaviour. We consider this study further, and evaluate Wilkes-Gibbs’ interpretation of these results more fully in Sections 3.2.3.2 & 3.2.3.3.

Whether Clark and his co-workers’ claims for Coordination, Collaboration and for the Principle of Least Collaborative Effort are either philosophically or empirically useful in the examination of notions of Cooperation will be considered in Chapters Three and Six, respectively. We will evaluate the contributions of the Collaborative theory in more depth there. However, what should be noted at this point is the way in which this research has moved investigation into dialogue beyond the simple structures of Discourse Analysis and Conversation Analysis, and into a sphere which demands a more coherent explanation of the whole process, rather than just the mechanics.

2.4.3 Output-Input Coordination Model

The Output-Input Coordination model is a mechanism for describing the way in which speakers coordinate on the use of a semantic framework (Garrod and Anderson, 1987; Garrod and Doherty-Sneddon, 1994). It was developed through the analysis of dialogues produced by the Maze Task. This is a computer-based game, where the two participants are seated in different rooms, and can only communicate via a 2-way audio link. Each can see a ‘maze’ displayed on their monitor; this consists of a series of small boxes with links between them, which the players can travel between. The maze on each screen is identical, except in three points. Firstly, each has a marker showing the player’s current position, secondly, a goal which the player must move towards, and finally, a series of switch points and gates (which bar a route). The significance of the switch point is that when a switch is triggered (by the other speaker positioning their marker in that box on their screen), all the gates are opened, and all previously open routes are ‘gated’.

The participants’ aim is for both players to reach their respective goal points: at this stage the game is completed. This task is collaborative because the participants must both coordinate on the method of locating points in the maze, and gain the assistance of their partner in opening the necessary routes to the goal. Therefore, the maze dialogues contain many exchanges concerning the descriptions of current positions, as the participants attempt to agree on the position of
current and target locations.

These descriptions of locations were categorised into four different types. The separate categories represented alternative semantic frames of reference, for example using coordinates (MATRIX) versus using some salient feature of the maze as an identifier (FIGURAL). Garrod and Anderson (1987) found that pairs of participants would coordinate on a particular semantic framework. What is of particular interest is not so much that the speakers achieve this coordination, but how they achieve it.

Coordination is not an instant process, the participants tended to slowly gravitate towards their favoured scheme. For example, the most popular process in the data was to move from more concrete schemes (e.g. FIGURAL) to more abstract schemes (e.g. MATRIX). However, this process did not tend to be actively negotiated. Active negotiation tended only to occur when the pair were already experiencing coordination problems. And even where it did occur, the speakers tended not to stick with that mode of reference.

The favoured method of coordination was simply to use the mode of reference which your partner most recently used. For example, if one speaker used a FIGURAL frame of reference to describe a location, then, providing that the description was successful, their partner would then use a FIGURAL description in the formation of their next utterance. Therefore, the speaker formulates their output according to the same semantic framework as that used by the most recent relevant input (i.e. last utterance from their partner). The reasoning for this is that the current speaker has set up a successful interpreter for one semantic framework, so it is much more straightforward to ‘reverse’ that interpreter, and generate a location description using the same semantic framework (Garrod and Doherty 1994:209). In terms of Lewis (1969), this type of coordination is an example of precedence. The semantic frame of reference used by your partner is the most salient one to use, and is therefore the one most likely to be understood.

Garrod and Anderson (1987) and Garrod and Doherty-Sneddon (1994) argue that the Principle of Output-Input Coordination is in keeping with Clark’s Principle.

---

15Garrod and Anderson (1987) define this as being where terms such as ‘left’ and ‘row’ were being defined, or where the speakers were trying to characterise the whole maze, e.g. “imagine it as a grid” (Garrod and Anderson 1987:206).

16It predicted the form of only 50% of subsequent descriptions.

17Obviously, this is a generalisation of what happens, otherwise the interlocutors would never progress beyond the first semantic frame which they tried! Garrod and Anderson (1987) describe a ‘Master/Slave’ situation, where typically one participant tended to shift semantic frame, and the other would follow. The reader is directed to their discussion of this topic, as it lies outside the focus of this research.

36
of Least Collaborative Effort (see Section 2.4.2 above). Using the same semantic frame of reference shows a conservation of collaborative effort, as speakers are maintaining the same interpretation system, rather than making the effort of changing systems. We evaluate this argument in Section 3.2.3.3 below.

2.5 Computational Approaches to Dialogue

In this Section, we will describe some aspects of dialogue modelling in more computational approaches to linguistics. Many of the dialogue systems currently available draw on linguistic input from the linguistic areas which we have discussed above. Here, therefore, we will concentrate on approaches which have been developed within the Computational/Artificial Intelligence paradigm. We will concentrate on the ideas behind the linguistic analysis rather than the systems themselves, as the computational implementation is not of much relevance to the work we present here.

2.5.1 Dialogue Games

Dialogue games describe a similar structure to Discourse Analysis’ exchanges, although the motivation behind the description is different.

Power (Power, 1974, 1979) suggested that human talk is a way of achieving goals: if you want a window opened, you can either open it yourself, or you can ask someone to open it for you. This emphasis on the effect of language is developed from Austin (1962) and Searle’s (1969) work on Speech Act Theory, which is discussed above (Section 2.2.2). Power’s agents in his dialogue system only communicated with each other when it was necessary for a goal to be achieved. Otherwise, the agents performed their own actions independently, and didn’t communicate.

Their talk was modelled in the form of ‘games’, which were goal-directed structures. For example, there were games dedicated to attracting attention, requesting information, or making a statement about the state of their world. Therefore, unlike Exchange Structure, Power was developing task-specific structures, not general ones.

The two agents, John and Mary, lived in a world where there were doors and bolts. Their goals mainly revolved around getting to the other side of doors, which often required the other agent to lift a bolt. Power (1987) contrasts the way his agents might have achieved a particular goal with a similar human conversation. At this
point, John is OUT, and Mary is IN; both doors (FRONT and BACK) are CLOSED.

Example 1: Human conversation

JOHN: ‘‘Help me get in!’’
MARY: ‘‘Okay, I’ll come to the front door.’’

Mary walks to front door.
Mary pulls front door open.
John walks inside.

Example 2: Power’s system

JOHN: ‘‘Mary!’’
MARY: ‘‘Yes?’’
JOHN: ‘‘May I request something?’’
MARY: ‘‘Okay.’’
JOHN: ‘‘Will you help me get in?’’
MARY: ‘‘Okay. May I ask you something?’’
JOHN: ‘‘Okay.’’
MARY: ‘‘Is the front door open?’’
JOHN: ‘‘No.’’
MARY: ‘‘May I suggest a plan?’’
JOHN: ‘‘Okay.’’
MARY: ‘‘Let’s get the front door open.’’

Etc...

Power (1987:85)

In Power’s system, the relationship between utterances and outcomes is very explicit. In comparison, the human dialogue hides much of the planning (or, collaboration) process. Power likens this to an iceberg, where only small peaks may be visible to the observer, but there is much beneath the surface. His interest
here is primarily in the issue of joint plans, and the difference between agreeing on a plan before embarking on an action (AGREEMENT), or one speaker deciding on a plan, and simply initiating that action (PRESUMPTION\textsuperscript{18}). The interests of our work are more focused on the reasons why speakers might choose agreement (for example, a less extreme version of the conversation postulated by Power's system above) or presumption, and what the resulting effect might be.

Power's work was developed by Houghton (Houghton, 1986; Houghton and Isard, 1987), who introduced the concept of interaction frames, which are similar to Power's notion of 'games'. These frames were employed in a computational model of dialogue, where the agents interacted in order to open doors, get to the other side of doors, and move blocks around, much as in Power's system. The term 'interaction frame' makes a more explicit connection to the Artificial Intelligence problem-solving/planning work of the time (e.g., Fikes and Nilsson, 1971). Indeed, both Power and Houghton's computational structures, with preconditions and changes in the knowledge state had much in common with this work. However, Houghton's work made a more explicit connection between the act of speaking and the goal structure of the planning system. Frames consist of preconditions, actions and results which are connected to a particular goal. Fred and Doris (Houghton's agents) talked as a result of the relevant interaction frame initiating the ACTION of speaking in order to achieve a goal. In other words, the action of communication was part of the planning structure (interaction frame). In Power's work, the planning process was separate to the action of communicating. Game structure only includes the action of communicating, it does not include the planning which initiates the communication. Therefore, one of the main differences between dialogue games and interaction frames is that the latter includes both the dialogue structures (the games) plus all the other associated planning mechanisms necessary to achieve a particular goal.

In addition, Houghton's agents also generated their own 'speech' (in a form appropriate as input to a speech synthesiser). Power's agents used pre-stored strings, whereas Houghton's generated their own utterances.

In this Work, we will use Power's term 'games' to refer to this concept of goal-based dialogue structures, as this is the term which has been used in later work (e.g., Kowtko, Isard, and Doherty, 1992).

We briefly mentioned the structure of dialogue games above. Here, we will consider it in a little more detail. Games, like exchanges, consist of a series of moves,

\textsuperscript{18}Power notes that the term PRESUMPTION is not intended to have any negative connotations.
or embedded exchanges, but their structure is defined according to the context, rather than being fixed: they are genre-specific. We have suggested previously that the rigidity of the IRF model (in whichever definition) is problematic, because moves do not always follow this set pattern. Dialogue games avoid this by firstly making the games more context-specific, and also not imposing a rigid order. The context-specificity is goal-based: different types of goals employ different types of games, which have different structures, and employ different move types. For example, Houghton distinguishes between a MAKEKNOWN game, in which information is offered, a FINDOUT game, in which information is elicited, a GETDONE game, in which a favour is achieved and a GETATTENTION game in which the other speaker's attention is sought. All these games define the preconditions for the game's use, procedures for carrying out any non-verbal activities involved, and the replies that would be expected of the addressee. However, there is no fixed length for a game (it can be any number of moves), and the moves do not have to be in a particular order.

One of the most important developments of this model from Austin and Searle's work is the emphasis on the role of the utterance. Houghton and Isard (1987) point out that identical sentences can have different meanings according to the context.

For example:

**Example 1**

S1. Hey mac, how come all the bars're shut?
S2: It's three fifteen.

**Example 2**

S1. It's three fifteen, Dave.
S2. Thanks.

Houghton and Isard (1987:252)

Both these utterances are assertions, and therefore in Searle's analysis, they would have the same preparatory, sincerity and essential conditions. However, it is quite clear that these two utterances perform quite different roles. In the first example, S2's utterance is an indirect answer to a question. Giving the time here is not only
a provision of information, but also a recognition of S1’s goals, and an agreement to take part. Houghton argued that if S1 accepts S2’s contribution, then S1’s goals should have been achieved, and this phase of the conversation should be ended. In the second example, however, S1 is a presentation of unsolicited information, and as such, it needs some degree of acceptance from S2 before the goal of the conversation can be said to be achieved\textsuperscript{19}.

The importance of role to utterance meaning seems to lead naturally to an emphasis on utterance role itself. It is this aspect of Power and Houghton’s work we have developed, by considering the efficacy of particular utterances in particular roles.

This model was developed to apply to natural language dialogue, rather than just being used in a constrained computational model (Kowtko, Isard, and Doherty, 1992). The dialogues involved were from the HCRC Map Task Corpus, which is described briefly in Chapter One and in more detail in Section 4.2. Kowtko, Isard, and Doherty (1992) identified a series of 6 types of game, and 12 different moves\textsuperscript{20}, some of which are specific to the task involved, others less so. Their view is that human interactants have a pool of such games which they can draw on, some of which are genre-specific, others which are more general. Interactants can also learn new games.

From the point of view of our work, dialogue games offers a more goal-directed view of conversation, but it is still a labelling system based on what actually happened\textsuperscript{21}, rather than on what might have happened. Neither does it consider the impact of the choices that interactants make. These two points are essential to the study of task-oriented dialogue which is presented here.

\subsection*{2.5.2 Communicative Postures}

Shadbolt (1984) introduces the concepts of ‘risk’, ‘effort’ and the risk-effort trade-

\textsuperscript{19}Clark would probably argue that both these assertions require some sort of acceptance. The difference between them would be expressed by the presentation-acceptance structure. S2 in Example 1 is both an acceptance of S1, and a new presentation, whereas S1 in Example 1 is simply a presentation. See Clark and Schaefer (1989) for a more detailed discussion of presentation-acceptance chains.

\textsuperscript{20}The 6 games are: INSTRUCT, CHECK, QUERY-YN, QUERY-W, EXPLAIN, ALIGN. The remaining 6 moves (the game types are classified by the initiate move) are: CLARIFY, REPLY-Y, REPLY-N, REPLY-W, ACKNOWLEDGE, READY. See Kowtko, Isard, and Doherty (1992) for explanations of these.

\textsuperscript{21}This comment is made from the point of view of dialogue analysis rather than dialogue generation.
off. He analysed dialogues similar to those in the Map Task Corpus\textsuperscript{22}, and noted that the speakers varied in the strategies they used, according to the situation. The strategies he identified were primarily concerned with how much information the participants transferred explicitly, and how much was left for the other speaker to infer. These strategies, which were mostly concerned with issues of knowledge, he calls parameters. The settings which an agent uses for these parameters are called communicative postures, and these are described in terms of risk. In Shadbolt's system, there are, in fact, only two settings, low or high risk. The same is true of Carletta's (1992, 1996) extension of his system.

In terms of the Map Task, an example of low versus high risk might be in the choice speakers make when they introduce the landmarks which the route on the map circumnavigates. A speaker can either introduce a landmark before using it in a route description\textsuperscript{23}:

\textit{Do you have a burnt cottage?}
\textit{Go to the left past the burnt cottage.}

Or they can directly incorporate it into an instruction:

\textit{Go to the left past the burnt cottage.}

In the first example, the Speaker ensures that the Addressee understands the reference before proceeding with the instruction. Therefore, the instruction is more likely to work as planned (low risk). In the second example, the Speaker is relying on the fact that the Addressee will have this feature on his map too - otherwise some potentially costly repair work may result (high risk).

Therefore, high risk settings mean that the Hearer has to infer information, and this risks the possibility of miscommunication, and thus plan failure. Low risk settings are more explicit, and are more likely to work as intended. However, high risk settings require less effort than low risk ones and Shadbolt suggests that these are more favoured by human agents. He argues that speakers operate according to the Principle of Parsimony.

\footnote{They were taken from the pilot run of the Map Task, which used slightly different maps.}

\footnote{This is the structure described by Anderson \textit{et al.} (Section 2.4.1) as a \textit{question introduction}.}
Principle of Parsimony:
"...a behavioural principle which instructs processors to do no more processing than is necessary to achieve a goal."

Shadbolt (1984:342)

Basically, this states that interactants should do no more work than is necessary to achieve a goal: they should be maximally efficient. Speakers choose that setting on the parameter which involves the most risk, the least effort, but still achieves the goal. This inverse relationship between effort and risk is termed the ‘Risk-Effort Trade-Off’, and is discussed further in Section 3.2.4.

The type of parameters which Shadbolt defined are concerned with general strategies, rather than being a move-based account. We summarise these parameters below. Each one is phrased in the form of a question which could have the answer ‘maximally’ or ‘minimally’. ‘Maximally’ implies a low risk posture, ‘minimally’ implies a high risk posture:

**Difference** Do you assume that your partner may have different information about the map features as yourself?

**Specification** How unambiguously do you specify the current Discourse Object?

**Ontology** Do you introduce new objects into the discourse?

**Focus** Do you assume that your partner may not share your area of focus

**Decentering** Do you construct a view of your partner’s knowledge state?

**Feedback** Do you provide feedback about your partner’s discourse contributions?

Adapted from Shadbolt (1984:346-7)

The interactants’ overall choices of strategy (parameter) would affect their choice of move, but it is not this level of language with which Shadbolt is concerned. However, it is precisely this level of language with which we are concerned. Therefore, Shadbolt’s approach is useful in terms of the concepts of effort and risk, but the detail of his work is less relevant. The importance of this work from our perspective is the focus on the questions of dialogue choice and dialogue outcome,
even if the relationship between the two is postulated, rather than empirically tested.

Carletta (1992, 1996) does develop the interaction between parameters and the choice of moves in the dialogue game. Her system, JAM, generates a simple conversation between two participants in the Map Task domain. The conversation is varied according to the degree of risk each participant is prepared to take. The parameters in the system are developed from Shadbolt's and are concerned with task planning, discourse planning, utterance realisation and the speaker's commitment to the current plan. However, as her system also only has two settings (low or high risk), this interaction is very limited. The point of Carletta's work is to show the possibilities of effort and risk in a computational linguistic framework. Our concerns are more centred in the linguistic domain, and therefore we would want to examine further the dialogue possibilities of effort and risk.

For our purpose, we need an analysis system which has some of the concerns of Shadbolt/Carletta (effort and risk), but is based at the level of the move24, and has a greater consideration for the effect of interactant's choices, both in terms of the dialogue, and for its eventual task outcome. The relationship between the concepts of Risk and Effort and the Typology of Move Attributes is detailed in Chapter Five.

It should also be pointed out that both the Principle of Parsimony and the Risk-Effort Trade-Off are not based on a quantitative evaluation but on informal observation. These concepts were put forward by Shadbolt, and developed by Carletta, on the basis of their studying dialogue data in an informal way. Part of our intention is to investigate whether the concepts of Parsimony, Effort and Risk are upheld in a more formal empirical study.

We consider the broader issues surrounding the Principle of Parsimony and the Risk-Effort Trade-Off in the next Chapter (Sections 3.2.4 & 3.2.5).

### 2.5.3 Discourse Obligations

Traum (1994) used the notion of discourse obligation which is intended to explain why speakers answer questions or queries. He argues that the intentional account of discourse structure (e.g., Grosz and Sidner, 1986) leads to a strong assumption of cooperativeness. In the case of the question-answer pair, the fol-

---

24 We chose the level of the move because we wish to analyse human dialogue. We cannot easily analyse speakers' higher beliefs - moves are the most concrete evidence we have.
lowing set of actions would be generated:

- Agent A asks question
- Agent B recognises Agent A’s goal to find out the answer
- Agent B adopts a goal to tell Agent A the answer in order to be cooperative
- Agent B plans to achieve the goal, thereby generating the answer

Adapted from Traum and Allen (1994:1)

In this model of the process, Agent B must take on Agent A’s goals. This works reasonably well in a straightforward example, but does not explain why we answer questions either when we don’t particularly want to, or when we don’t know the answer. If, however, we view giving an answer to a question as part of a social code, then it is easier to understand why questions are answered. Take the example of a politician responding to a journalist’s question. The answer given evades the question, but it is still an answer. This behaviour can be better explained by discourse obligations than by intentions:

Politicians answer journalists’ questions because they are socially obligated to do so, not because they wish to take on the journalists’ goals. If the intentions model was accurate, then surely the politician would provide the type of answer which the journalist sought rather than evading the question.

Traum’s model relates particular obligations to particular conversation acts. Carberry (1990) has used the term expectation in this way, but this does not express the compulsory nature of obligations. Discourse obligations have to be carried out in order for the conversation to proceed normally, whereas expectations may or may not be fulfilled. The following sample discourse obligation rules are an adaptation from Traum and Allen (1994:3):

<table>
<thead>
<tr>
<th>source of obligation</th>
<th>obliged action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker1 Accept or Promise A</td>
<td>Speaker1 achieve A</td>
</tr>
<tr>
<td>Speaker1 Request A</td>
<td>Speaker2 address Request: accept A or reject A</td>
</tr>
<tr>
<td>Speaker1 YNQ whether P</td>
<td>Speaker2 Answer-if P</td>
</tr>
<tr>
<td>Speaker1 WHQ P(x)</td>
<td>Speaker2 Inform-ref (x)</td>
</tr>
<tr>
<td>utterance not understood or incorrect</td>
<td>repair utterance</td>
</tr>
</tbody>
</table>

25See Traum and Hinkelman (1992) for an explanation of the difference between conversation acts and speech acts.
These obligations are derived in part from the notion of the adjacency pair in Conversation Analysis (see Section 2.3.2.2). These too have the same idea of the expected relationship between the first part and second part of an utterance pair, although the relationship is not codified in such an explicit fashion.

This idea of obligation does seem to motivate speaker behaviour in terms of why speakers open their mouths and start speaking. However, it does not seem to have as much to contribute towards the debate on how speakers decide what they will say, once they have made that decision. It is this latter question in which we are more interested here, and therefore we will not be investigating the concept of discourse obligation itself in more detail. However, Traum's approach to speaker motivation is relevant to the Typology of Dialogue Attributes, and we will return to this briefly in Chapter Five.

2.6 Summary

The study of dialogue has been undertaken in many different disciplines. Here, we have summarised what we consider to be the major contributions from philosophy, linguistics, sociolinguistics, psycholinguistics and computational linguistics. It is these particular research areas which have formed the basis for the research described in the remainder of this thesis.

Our aim has been to locate each approach in terms of their roots and their intentions, in addition to proving a basic description. This provided a basis for signposting their contribution to the development of either the mode of analysis presented in Chapter Five, or the examination of 'Cooperation' in Chapter Three.

The discussions of Gricean Cooperation, Clark's Collaborative Theory and the Principle of Parsimony will be pursued in the next Chapter. Here, we try to move from theoretical discussions to empirically-testable hypotheses.

The introduction of the Typology of Move Attributes in Chapter Five will draw more on both the structural analysis approaches which have been discussed here (Discourse Analysis, Conversation Analysis, Dialogue Games) and those which operate at higher levels of planning (Communicative Postures and Discourse Obligations). In this Chapter we will also explain the rationale behind using the structural aspects of dialogue analysis to investigate the claims of the Dialogue Principles discussed in Chapter Three.
Chapter 3

Cooperation as a Structuring Principle

3.1 Introduction

The aim of this Chapter is to present a discussion of ‘cooperation’ and related notions.

The initial part of the Chapter will consider the concept of ‘cooperation’ through its philosophical beginnings (Grice, 1975), its current usage, and its relationship to the plethora of other principles which have also been advocated (e.g. Coordination (Clark, 1996), Collaboration (Clark and Wilkes-Gibbs, 1986) and Parsimony (Shadbolt, 1984; Carletta, 1992)).

Using this discussion as a basis, we will then give an operational definition for the term ‘cooperation’ as it is used in this research.

The final part of the Chapter is concerned with the empirical application of cooperation, and the other structuring principles. We discuss what effect we would expect from each of these principles if they were operating in a task-oriented dialogue. These expectations are later investigated through the Typology of Move Attributes (outlined in Chapter Five), and the results are presented and discussed in Chapter Six.

3.1.1 Competing Principles for Dialogue Structure

When we are considering the various Principles in this - and other - chapters, we will be treating them as competing theories of conversation rather than individual contributors to some other overall theory. There are two main reasons
for this approach. Most importantly, the originators of these Principles and their proponents treat them as overarching Principles and it is their ideas which we are interested in analysing and testing. From a more pragmatic point of view, it is easier to empirically test more extreme positions. For example, empirically showing the differences between extreme right wing and left wing politicians should be easier than showing the differences between the moderates of any political party.

Of course, we are aware of the unlikelihood that it is only one of these Principles which governs how we operate in dialogue. It is far more probable that there is more than one factor affecting interactants' behaviour. However, we would suggest that in any given context, it is likely that one or other of these Principles will be the major factor in dialogue structure. In this work we hope that the empirical evidence will enable us to identify those Principles whose predictions are largely met by the data, and those whose predictions are generally not fulfilled. This should provide a first step towards identifying those major factors in dialogue structure.

### 3.2 Structuring Principles for Conversation

In this Section we will examine these competing conversational principles which have been offered as explanations for aspects of the structure of dialogue: how do we, as speakers, decide what to say, and, as listeners, how to interpret it?

We will concentrate mainly on notions of Cooperation (Grice, 1975; Stenström, 1994; Brown, 1995), the principles described by Clark and his co-workers: Coordination and Collaboration (Clark and Schaefer, 1987a,b, 1989; Clark and Wilkes-Gibbs, 1986; Schober, 1995; Schober and Clark, 1989) and Shadbolt's (1984) Risk-Effort Trade-Off. Essentially we will be considering the differing ways in which these various groups of researchers have characterised the same phenomenon: conversation.
3.2.1 The Cooperative Principle

Grice's Cooperative Principle:

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

(Grice 1975:45)

Grice was mainly concerned with the interpretation of indirect speech acts - how listeners negotiate the semantic leap between the literal and intended meaning. The question of why speakers should choose to use such implicatures rather than keeping to more straightforward structures is not an issue addressed by Grice, although it is now an important part of what is often termed Gricean Pragmatics (Thomas, 1995).

The emphasis of Grice (1975) is on the wide margins within which a Speaker can construct their utterance: the Speaker makes the choice, it is up to the Hearer to work out the meaning. The Cooperative Principle is not about making the task of the Hearer straightforward, quite the reverse! In fact, it allows the Speaker to make their utterance harder, rather than easier, to interpret: we can omit information or present a non-literal utterance, and expect the Hearer to do the extra work necessary to interpret it.

This approach is similar to that taken by Sperber and Wilson (1986) in their analysis of Grice. They claim that we negotiate the gap between literal and intended meaning by assuming that the utterance is relevant, and then looking for an appropriate interpretation. The concept of relevance here subsumes the Cooperative Principle and all the maxims: Sperber and Wilson’s argument is that all of Grice’s implicatures can be reduced to this. Like the Cooperative Principle, relevance demands that the Hearer works hard to extract the intended meaning.

The Cooperative Principle therefore works on the basis that all interactants operate according to its assumptions. If Hearers could not assume that Speakers were using the Cooperative Principle, then they would have no basis on which to make implicatures. Conversely, if Speakers could not assume that Hearers were using the Cooperative Principle, then they would not be able to use non-literal utterances, as they could not guarantee that the utterance could be interpreted.
3.2.1.1 Problems with the Gricean Approach

As it stands, the Cooperative Principle and the attendant maxims have certain limitations. Firstly, Grice’s work is primarily part of the philosophical paradigm rather than the linguistic one. Like Austin’s description of performative speech acts, Grice’s ideas are based on introspection and invented examples. Current linguistic work in dialogue and discourse expects a more empirical approach, and the evaluation of real language data. This does not invalidate Grice’s work, nor the importance of the Cooperative Principle and maxims as a conceptual framework, but this difference in methodological approach does make the application of Grice’s ideas to dialogue data more problematic. In essence, we are attempting to apply concepts developed in a theoretical atmosphere to a very practical project. If we compare the attempt to apply, say, Clark’s various Principles to dialogue data, then we can see the contrast. Clark’s Principles were developed through experimentation with language in context, therefore we would be using the same type of ‘test environment’ for the development and the evaluation. However, this is not true of any attempt to apply Grice’s framework: the current researcher has to make the ‘leap’ from theoretical concept to practical application.

Secondly, the Cooperative Principle relies on shared knowledge\(^1\). Implicatures rely on the Hearer having access to the knowledge required in order to interpret them. Grice does not consider cases where implicature fails: in his world, conversation works without redress to repair sequences. He has no term for a failed implicature. Speech Act Theory (Austin, 1962; Searle, 1969) can talk of *infelicitous* acts, when utterances misfire, or of a mismatch between the illocutionary act (intention) and the perlocutionary act (result), but even these are more concerned with either some essential condition not being fulfilled, or the deliberate misinterpretation of an utterance. Addressees do not simply fail to understand. In task-oriented dialogues such as those produced by the Map Task (see Section 4.2), where the status of information has to be determined, and the participants do not always understand each other, such an approach can have limited applicability. Much of the conversational work in these dialogues consists of the negotiation of uncertain information, an area which Grice’s work, or the philosophical approach in general, does not attempt to explain.

Thirdly, Grice’s paper principally deals with the issue of indirect utterances, particularly indirect answers to questions. It addresses one small aspect of language,

---

\(^{1}\)This is a vexed area in itself, and not one which we intend to discuss here. For the purposes of this thesis, we assume that there are such things as shared and mutual knowledge.
largely out of context. It tells us little about how we use other aspects of language: why we use indirect speech acts, how we formulate initiating utterances, how we decide what to say next, or even how we plan the next part of our dialogue (particularly important for information-based dialogues). So what use is it in terms of explaining how interactants behave? We will attempt to address this question in the following subsection, when we consider interpretations of Grice, and how his approach to language could be generalised to broader language use.

3.2.1.2 Interpretations of Grice

Grice’s Cooperative Principle and maxims could be interpreted in several ways. We will consider two main strands here: Cooperative Principle as rational action and Cooperative Principle as efficiency.

1. The Rationalist Interpretation

Although the process of implicatures can be described in fairly simple terms, their place in Grice’s overall theory of language use is less clear. Grice himself chooses to emphasise the importance of language as rational behaviour:

“... one of my avowed aims is to see talking as a special case or variety of purposive, indeed rational behaviour.”

(Grice 1975:47)

His explanation of the implicature process is based on the principle of rationality, rather than just his observations as a competent language user (Grice 1975: 47-48). Obviously, his work is based on some degree of observation, and is therefore empirical by nature, but Grice chooses to focus on the rational explanation of language as a higher order principle. In Grice (1975), he addresses the question of why we should assume that speakers will follow the Cooperative Principle and Maxims, and answers himself thus:

“A dull but, no doubt at a certain level, adequate answer is that it is just a well-recognized 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, emphasis in original)
Therefore, Grice is attempting to analyse implicatures in such a way that highlights their rational aspect. This is not to say that this is the sole, or maybe even the most appropriate, interpretation of Grice’s work, but it is obviously an aspect which he wished to emphasise.

If one follows this aspect of Grice’s work, then implicatures could be characterised thus:

> Hearers assume that an utterance addressed to them is intended to be meaningful, therefore if the utterance doesn’t have an appropriate literal meaning, they will look for a more useful interpretation. As far as the Hearer is concerned, the Speaker providing an uninterpretable (meaningless) utterance would be pointless, and therefore irrational.

Grice believes language to be based on rational principles, in the same way as he would argue that all human action is rational. As far as Grice is concerned, the Cooperative Principle demonstrates how we use language rationally, and it is this principle of rationality which could be generalised to broader language use, rather than the Cooperative Principle itself. We will return to this idea in Section 3.4.1.

### 2. The Efficiency Interpretation

An alternative way to view the Cooperative Principle and maxims would be to see this as a characterisation of a language system which speakers can exploit.

We can use the CP and maxims to justify the use of briefer utterances. Grice (1978) uses the following example:

> Assume that it is mutual knowledge that New York and Boston were blacked out last night. Assume also that C was in New York at that time. If A asks B whether C saw a particular television programme the previous evening, B could reasonably make one of the following responses:

1. No, he was in a blacked out city
2. No, he was in New York
3. No, He was in New York, which was blacked out

Grice argues that either utterance one or two would be most appropriate, as the other information could be inferred from the mutual knowledge of the lack of television reception in New York and Boston the previous evening. He states of
utterance 3 that in providing the extra information "...the gain would have been insufficient to justify the additional conversational effort." (Grice 1978:114).

The Cooperative Principle and maxims instruct the Hearer to 'fill in the gaps' in the Speaker's utterance by using mutual knowledge, and the expectation that the Speaker is following the Cooperative Principle. In this case, one could argue that the possibilities outlined by the CP lead more efficient language, as less needs to be said in order for the message to be conveyed.

However, the exploitation of the CP does not always lead to shorter utterances. Take the following example adapted from Grice (1975:55):

*Failure to be brief or succinct*

- (a) Miss X sang 'Home sweet home'
- (b) Miss X sang 'Home sweet home' badly
- (c) Miss X produced a series of sounds that corresponded closely with the score of 'Home sweet home'.

In (b), the Speaker states that Miss X sang badly. In (c), the Speaker *implicates* that this is the case by using an overly complex utterance, flouting the maxim of manner. In this case, the implicature also generates irony. One could argue that this exploitation of the CP is still efficient because it enables the Speaker to fold in both the intended meaning and the humour. Leech (1983:80) argues that such exploitation of the CP can also allow us to follow the Principle of Politeness:

A: *We'll all miss Bill and Agatha*

B: *Well, we'll all miss Bill.*

Here, the Maxim of Quantity is flouted: to give the whole truth Speaker B should also refer to Agatha. Therefore, Speaker B is being polite by not being explicitly negative about Agatha. However, one could also argue that the generation of *humour* here is far more important than the issue of politeness. Even if this were the case, the efficiency of the utterance would not be in question, as Speaker B's utterance would still be combining two functions: the message content, and the generation of irony.
Therefore, the CP and maxims could be characterised as **efficiency** in language use. We will consider how this concept could be applied to wider language use in Section 3.4.1.

### 3.2.2 Dialogue as Cooperation

In the previous Section, we examined Grice’s Cooperative Principle. Here, we consider more general views of ‘cooperation’ which are current in the linguistic literature.

> “...speakers cooperate.... When studying transcripts of genuine conversation one is struck by the general atmosphere of cooperativeness and harmony”

Stenström (1994: 1)

Dialogue is seen to be ‘cooperative’, although what is meant by that term is rarely defined. This is demonstrated by Stenström (1994:1) herself who says of ‘cooperation’:

> “This does not mean, of course, that the listener always waits for the speaker to finish before taking over. Nor does it mean that speakers never disagree. Nor does it mean that speakers never disagree, object or contradict each other.”

Yet, like many other commentators[^2] in this area she never actually states throughout her discussion what was meant by this term. This problem is compounded by the difficulty of separating the technical notion of ‘cooperation’ from the everyday understanding of the word. There is a tendency for readers to assume the latter interpretation, unless an alternative clear definition is offered. For example in the opening Chapter of his textbook on Pragmatics, Levinson (1983:50) refers to an implicature relying on “some very general expectation of interactional cooperation [sic]” which could be interpreted in a Gricean sense, but could equally be taken to refer to the everyday meaning. It is not entirely clear from the context what meaning was intended, but it certainly demonstrates the difficulty in the term’s use[^3].

[^2]: For example, Clark and his co-workers use the term ‘cooperation’ but never define it explicitly (Clark and Wilkes-Gibbs, 1986; Wilkes-Gibbs, 1995, 1997).

[^3]: See Thomas (1995), Chapter Three, for further discussion of this issue.
In most linguistic descriptions, dialogue is seen to be successful: human interactants are considered to be 'good' at dialogue. We manage turn-taking with only tiny gaps, rarely talk over one another (overlap), and are able to recover from miscommunication. Much research in this area has been devoted to the analysis of unproblematic, smooth dialogues: much of the data presented in Conversation Analysis (e.g., Sacks, Schegloff, and Jefferson, 1974; Schegloff, Jefferson, and Sacks, 1977; Schegloff and Sacks, 1973), Discourse Analysis (e.g., Coulthard, 1994; Coulthard and Montgomery, 1981), and Pragmatics (e.g., Thomas, 1995; Grundy, 1995) is from unproblematic talk exchanges. Coupland et al. (1991) call this the 'Pollyanna Perspective' - concentrating on the successful communication that we can observe, and ignoring the problematic encounters.

When dialogues are described as being 'cooperative', it is rarely clear what they are being compared against. What elements of a dialogue are 'cooperative'? What would an 'uncooperative' dialogue look like? There is no benchmark of comparison: the term 'cooperative' is being used in a vacuum, as there are no alternative descriptions by which we can define its meaning. We have only the binary opposition of 'cooperation' and 'non-cooperation', which is problematic in two respects. Firstly, as we will argue later in Section 3.3, it is questionable that a concept dealing with interaction can only have two values (binary rather than gradeable), and secondly as there are few explicit definitions of 'cooperation', how can we make even that binary division meaningful? If 'cooperation' is difficult to define, then 'non-cooperation' is even more awkward.

Given the lack of explicit discussion of the term 'cooperation', alongside the type of data which is often presented in linguistic texts, it is perhaps unsurprising that there is some tension in the term's meaning. There is a conflict between the approach favoured by Grice, and the more general usage of the term 'cooperation' which is endemic in the literature. Therefore, if we describe a set of dialogues as being 'cooperative', it is unclear in two respects: what we intend to mean by that statement, and what the reader takes as the meaning.

We discuss the meaning of 'cooperation' which we will use in this thesis in Section 3.3. This will distinguish between what we will term 'Gricean Cooperation' and 'Cooperation'. Firstly, we will consider some alternative structuring principles which have been offered as alternatives to Grice.
3.2.3 Dialogue as Joint Action

In this Section we will examine the contribution of Clark and his co-workers to this research area. We will look at the Principle of Coordination and Collaborative theory.

3.2.3.1 Principle of Coordination

Clark and his co-workers have developed a theory of language use which is dependent on the notion of language as joint action (Clark, 1996; Clark and Brennan, 1991; Clark and Schaefer, 1987a,b, 1989; Clark and Wilkes-Gibbs, 1986; Schober and Clark, 1989; Schober, 1995). Grice's approach assumes a packaged utterance which is sent to the hearer, who then interprets it, and sends an appropriate message back: it is modelled as a series of individual actions rather than a joint action made up of individual participatory actions (Clark's terminology). Clark likens language to a musical duet, where we coordinate our playing to our partner: we slow down, play louder, or softer at the same time as the other person. If we do not accommodate in this fashion, then we cannot really be said to be playing that piece of music together. In such a Coordination problem, each person’s actions are dependent on the actions of the other.

We coordinate on many levels: who speaks and who listens (turn-taking), the content of the utterance (e.g. are the references interpretable), the common ground between the two speakers⁴, or how we refer to objects or spatial coordinates (Semantic Coordination: Clark and Wilkes-Gibbs 1986; Garrod and Anderson 1987). Clark (1985) uses the categories of coordination described by Lewis (1969): explicit agreement, convention, salience and precedence (see Section 2.4.2.1). However, these categories seem to be more descriptive than predictive: they do not motivate interactants’ behaviour. Convention defines a particular behaviour as fixed for no apparent reason: another behaviour could be used as successfully. Salience and Precedence describe some of our actions in dialogue (using common ground to interpret an indirect utterance; using the same mode of description as our partner), but it does not explain why we make these choices. Indeed, the concepts of salience and precedence could as easily be accounted for by Communicative Accommodation Theory (Giles and Coupland, 1991), although the lack of motivation is apparent here too.

⁴Defined as “...the sum of their mutual, common, or joint knowledge, beliefs and suppositions.” Clark (1996:93)
Also, all of Clark's examples are of successful communication, where interactants understand their partners and manage their turn-taking successfully. This is important in two respects, firstly it again supports this idea that humans are 'good' at language (The Pollyanna Principle), and secondly, it misses the opportunity to examine what important points cause us to miscommunicate. Maybe considering this area would illuminate the how and why of language a little more.

At the level of language we are interested in, the concept of convention seems particularly unhelpful. This should not be taken to imply that convention is not applicable to conversation per se, on the contrary, there is plenty of evidence for the existence of conventions in language use. Turn-taking behaviour, and the use of conventional implicatures are two examples that come to mind. However, we are interested in what causes the range of task success results which we see in our corpus of task-oriented dialogues. If these dialogues were structured according to convention then firstly, there shouldn't be much variation in behaviour (conventions are fixed), and secondly, there shouldn't be much variation in result. This is because a convention is defined as a behaviour which everyone adheres to, and is also a behaviour which has evolved by chance: another convention could have evolved instead, and the outcome should be equivalent. However, in the type of task-oriented dialogues with which we are dealing, we do see variation in both behaviour and results. Interactants make decisions about their behaviour which affect their task result: this should not be possible if the behaviour was purely conventional.

The concepts of salience and precedence are more relevant, and we relate them to our practical description of Coordination in Section 3.4.2.

### 3.2.3.2 Dialogue as Collaboration

The concept of 'collaboration' emphasises again the notion of language as joint action: something co-produced by the participants. Interactants work together to ensure that new contributions\(^5\) presented by the speaker are understood by the Hearer. This process is defined in terms of the Principle of Mutual Responsibility:

\(^5\)It is difficult to find a good definition of 'contribution'. One explanation would be to term it as the current idea/piece of information which the speaker is trying to convey.
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 and Wilkes-Gibbs (1986:33)

Collaboration occurs between ratified participants: only those who are perceived as taking part in the dialogue have control over the Collaboration process. Overhearers (subjects who observe a conversation) do not have the same understanding of information transferred as participants, even if the participant takes no verbal part in the interaction⁶ (Wilkes-Gibbs and Clark, 1992). This evidence is used to demonstrate that there is some level of activity in progress between members of a dialogue which is not accessible to those not included. Speakers pay particular attention to the needs of their interlocutors, but not the needs of those outside the dialogue. This demonstrates that there is an extra level of understanding achieved by ratified participants, but it does not demonstrate how it is achieved.

In Wilkes-Gibbs (1986), the degree of Collaboration is variously described as the number of turns, and number of words used by the participants in order to complete one segment of a dialogue, or the mean number of words per turn. Collaboration is a gradable concept: speakers can be more or less Collaborative. The most Collaborative dialogue has a high number of turns per segment, and either a high number of words or a low mean number of words per turn ⁷. Although this is a limited way of 'measuring' collaboration (discussed below), it marks an important sea-change. Principles now have the possibility of being gradable, not always absolute. 'Cooperation' as used in the linguistic literature is a binary concept - speakers are either 'cooperative' or 'not cooperative'. This is a narrow view, and one which is challenged by Wilkes-Gibbs' measures of Collaboration.

Wilkes-Gibbs' operational definition of low mean words per turn is meant to best represent a large amount of talk, with the maximum of back-and-forth exchanges. This is justified by arguing that more talk equals a greater degree of work, and

⁶Omniscent bystanders versus side-participants. See Section 2.4.2 for a fuller description of this experiment.
⁷Wilkes-Gibbs (1986) uses both of the latter and neither distinguishes between them nor comments on the relationship between them. We note that it would be entirely possible for a dialogue (or dialogue segment) to have a large number of turns and words (thus fulfilling one criterion for collaboration), but having a high mean number of words per turn (failing the other criterion).
shorter turns represents a sharing of that effort, and thus Collaboration. Although this does have some merit as a non-subjective measure, it is also open to criticism. Short conversations and a quick resolution can be due to participants working together particularly well: producing the relevant information in a concise form, which is easy for their partner to understand - this requires the minimum of dialogue for understanding to be achieved. Conversely, long dialogues can be due to inefficient, or not very capable speakers struggling to formulate the relevant information in an understandable form, whilst their partner tries to extract the information that they need. As a measure, their definition of Collaboration is very basic, and it is not entirely clear what aspect of the dialogue they are trying to focus on.

Long, inefficient dialogues might be considered to have a high degree of Collaboration as they do involve a high degree of back and forth discussion, but it is difficult to tell whether Clark and/or Wilkes-Gibbs intend Collaboration as either a purely quantitative measure - which is what that seems to imply - or whether there is any qualitative element. If the former is true, then the question of what this quantitative measure tells us about language has to be asked. Being able to say that (in their terms) a dialogue is more, or less, Collaborative is meaningless unless there are some predictions about the differences between these dialogues, either in content or outcome. The only prediction that we can see is that of the Principle of Least Collaborative Effort, which we discuss in the next Section.

Indeed, even within Clark and his co-workers research, there seems to be some conflict here. The idea of Collaboration was originally developed from the observation of speakers negotiating referring expressions for tangram figures. The dialogue pairs traded referring expressions until both speakers were satisfied that mutual understanding had been achieved. The referring expressions might be new suggestions, or slightly modified versions of the previous speaker’s suggestion. This process is referred to as contribution theory, and the speakers were described as ‘collaborating’ on referring expressions. This is a far more fine-grained, qualitative analysis, rather than the counting of words and turns. We would suggest that some measure based on this qualitative approach might produce more interesting results. Brown (1995) uses this operational definition of Collaboration in her analysis of Map Task dialogues. She argues that:
"It is certainly the case that those pairs who behaved in ... a collaborative manner in undertaking the Map task were much more likely to achieve a map drawn by B which in most respects resembled A's original map, than those pairs who did not collaborate in this way..."8

Brown (1995:207)

This is the opposite to Wilkes-Gibbs' (1986) findings: extra Collaboration had little impact on task success9. It is therefore unclear what we can assume to be the relationship between increased Collaboration and task success.

Neither does it assist a Clarkian definition of Collaboration. Nowhere in Clark et al.'s papers is there a definitive descriptive definition of this principle. On one hand, Collaboration seems to be the higher level of understanding which participants in a conversation build through the addition to their Common Ground, on the other, it is a simple numerical calculation of the number of words and turns involved, or the mean number of words per turn.

This dichotomy between what might be termed the qualitative and quantitative descriptions of collaboration is nicely summarised in Schober's (1995) explanation of Clark and Wilkes-Gibbs (1986).

"According to Clark and Wilkes-Gibbs (1986), the collaborative effort that conversational partners try to minimize includes two types of effort. There is the (a) effort involved in the individual acts of producing and comprehending, and the (b) effort involved in the collective act of agreeing that a reference has been understood. A speaker might minimize collaborative effort by putting in additional individual effort to select just the right noun phrase, thus reducing the addressee's individual comprehension effort and also reducing the collective effort (the number of words and turns that the pair takes to agree the noun phrase was understood). Alternatively, a speaker might minimize collaborative effort by using a less than ideal noun phrase that is easy to produce, judging that the additional individual cost (extra comprehension effort for the addressee) and the additional collective cost (the risk of needing extra words and turns to agree the noun phrase was understood) are worth it."

Schober (1995:229)

8'A' is the Giver, and 'B' is the Follower, in our terminology.
9In Wilkes-Gibbs (1986), more turns and more words gained the High Criterion matched pairs a slight improvement in task result, but the difference between their results and the Low Criterion matched pairs, or Low Criterion speakers in a mixed pair, was not reported as significant.
Again, we have moved from qualitative aims to a purely quantitative measurement. To be fair, Schober (1995) does try to address this problem, which we discuss in relation to the issue of Least Collaborative Effort (later in this Section).

Of these two approaches to Collaboration, qualitative and quantitative, it is the latter which we will have to carry forward to the empirical work described in the following Chapters. We would agree that this is not an ideal measure, but as the available work on Collaboration uses no other non-task-specific explicit definitions\(^{10}\), then neither can we.

**Operational Definition of Collaboration:**

These are adapted from Wilkes-Gibbs (1986)

1. **Number of words and turns in a dialogue**
   This measures the amount of effort which the speakers are investing: more words and turns = more effort

2. **Mean number of words per turn**
   A low mean number of w.p.t. should indicate many changes of speaker, and a sharing of the workload

There is also the issue of understanding "...to a criterion sufficient for ... current purposes." Clark and Schaefer (1989:262). Wilkes-Gibbs describes an experiment where there are two types of participant: low criterion and high criterion (Wilkes-Gibbs, 1986, 1997). Low criterion (LC) speakers needed to know less information than high criterion (HC) speakers in order to meet their goal\(^{11}\) There were three conditions for this experiment: matched LC pairs, matched HC pairs and mixed pairs. The speakers were unaware of the differences between their stated goals. HC pairs and LC pairs performed similarly on a route-replication test (HC pairs performed slightly better, but the difference wasn’t reported by Wilkes-Gibbs as being significant). The mixed pairs produced the worst result: HC Speakers in the mixed pairs did particularly badly. In terms of Collaboration, the HC pairs used more words and turns than either of the other conditions, but this didn’t

---

\(^{10}\)Schober (1995) provides a test situation where the Addressee and Matchers’ point of view has been manipulated, in order that references can be classed as more or less collaborative. Neither this, nor the Tangram referring expressions are generalisable to other types of conversation.

\(^{11}\)Low criterion speakers needed to be able to estimate how long it would take to drive a route. High criterion speakers needed to be able to describe the route in enough detail that another person could drive the route from their instructions. See Section 2.4.2 for a more detailed discussion of this experiment.
seem to gain them anything particularly. All the other pairs were considered to have used about the same amount of Collaboration - yet they produced differing task results. So what does measuring Collaboration tell us about dialogue?

In terms of Collaboration, all this study seems to say is that more talk does not necessarily get a better task result. Individual goals (criterion) are more important indicators of task result - participants with matched goals do better than those with unmatched ones. Wilkes-Gibbs explains this in terms of the LC speaker in the mixed pair determining the criterion necessary for understanding, rather than the HC speaker imposing their higher needs. This result is interesting in that it does demonstrate the fact that dialogue is certainly a joint product, but it does not really explain why the conflict of interest leads to the lower Collaborative criterion being chosen: what precisely determines whose purposes become the “current purposes”. Indeed, Clark’s Principle of Coordination would also seem to provide evidence against this choice, as the LC participant would not be coordinating on their partner’s beliefs about the conversation. In terms of the other approaches we have discussed here, it would seem to fit easier with Grice’s rational or efficient approach12 than the more folklinguistic notion of ‘cooperation’, where one would expect the highest criterion to be taken as the standard. We discuss this issue further in the evaluation of Least Collaborative Effort (Section 3.2.3.3).

3.2.3.3 Dialogue as Least Collaborative Effort

As an adjunct to the Principle of Collaboration, Clark and his co-workers also proposed the idea that participants achieve the successful addition of a new contribution to their common ground through the least possible Collaborative Effort. The Principle (of this name) states that participants will achieve the ‘grounding’13 of a contribution through the least joint Effort.

Principle of Least Collaborative Effort:
“Participants in a contribution try to minimise the total effort spent on that contribution - in both the presentation and acceptance phases.”

Clark and Schaefer (1989:269)

This could be generalised to something approximating the following:

12Although even Grice describes talk adhering to the Cooperative Principle as having “at least a mutually accepted direction...” Grice (1975:45)

13Clark’s term for the addition of a contribution to the participants’ common ground. This is explained further in Section 2.4.2.
**Principle of Least Collaborative Effort:**

“Participants in a conversation try to minimise the total effort spent in that interactional encounter”

This Principle is designed to account for the process whereby the participants negotiate joint understanding: this may well include repairs and refashioning. In the context of referring expressions, Clark argues that this is a minimisation of both participants’ Effort, rather than the classical view of Least Effort\(^\text{14}\) where the Addressee’s effort is minimised by the Speaker presenting the shortest unambiguous noun phrase for the context. This would seem to be a false opposition. Firstly, the effort involved in producing this kind of referring expression would be extremely effort-intensive for the Speaker — an NP-hard task in computational terms (Garey and Johnson, 1979 in Dale and Reiter, 1995). Secondly, it has long since been demonstrated that humans do not produce such referring expressions even in simple naming tasks, let alone the type tangram experiments described here (see Levelt, 1989, pp.129-134 for a summary of this work). Clark’s comparison is at best meaningless, and at worst, is guilty of an attempt to lead the reader astray. Such an argument should have its basis in the current accepted theories of the process, not in an approach which has long since been demonstrated to be incorrect.

Indeed, Clark and Brennan (1991) make a similar claim for whole utterances. Grice (1975) exhorts us (via the maxims of *Quantity* and *Manner*) to produce utterances which are appropriately informative and brief. Clark and Brennan (1991:148) quite justifiably argue that we rarely produce such “proper utterances”, pressures of time lead to errors or otherwise inadequate utterances. However, they argue further that Grice’s maxims and the idea of the “proper utterance” are a representation of traditional Least Effort. Again, this is an odd comparison because it is generally accepted (e.g., Carletta, Caley, and Isard, 1995) that to produce such a flawless utterance would involve a great deal of effort, and would be very time-consuming. For Clark and Brennan (1991) to argue that their Principle of Least Collaborative Effort involves less effort than this purported classical principle is meaningless. It has long been accepted both that language does not conform to this level of accuracy, and that for it to do so would require a great deal of processing power. Again, this is an ill-founded argument.

Even if one ignores this problem, then the claim that there is a decrease in Collaborative Effort is also problematic. Certainly, the behaviour Clark is describing

\(^{14}\text{Clark and Wilkes-Gibbs (1986) cite Brown (1958); Brown and Lenneberg (1954); Krauss and Glucksberg (1977); Olson (1970) and Zipf (1935) in support of this argument.}
involves some redistribution of the effort involved in the task: the Speaker uses less effort, and the Addressee, more. However, as Clark does not put forward any arithmetic measure of effort, there seems little conclusive evidence to suggest that as an entire process it is either less effortful than the classical principle discussed above, or the least joint effort (the sum of individual efforts) which could be involved in the operation. One could as easily argue that the Speaker is trying to minimise their effort at the expense of the Addressee. We will refer to this idea as the Principle of Least Individual Effort, in order to distinguish it from Clark et al.'s Principle of Least Collaborative Effort.

Principle of Least Individual Effort:
“Participants in a conversation try to minimise the individual effort which they invest in an interactional encounter.”

The evidence which Clark and Wilkes-Gibbs use to argue for this point is also questionable. Clark and Wilkes-Gibbs found that the amount of dialogue (‘collaboration’) used by their participants decreased each time they repeated the task. This is hardly surprising as the tangram figures which the speakers were talking about were the same each time, and the same pairs were used. Therefore, the participants were increasing their common ground in each repetition, and thus the talk needed to ground each contribution decreased each time. It is true that the speakers recognised that less talk, ‘collaboration’, was required for them to achieve their goal. However, this does not seem to be enough evidence to say that language in general makes use of this suggested Principle. The repetition of a tangram task is a very narrow use of language, and to argue that the language used in this context is representative of all language use is open to debate. Indeed, as we shall argue later, the linguistic behaviour which Clark and Wilkes-Gibbs describe here could as easily be explained by the Principle of Least Individual Effort.

Wilkes-Gibbs (1986) also makes the claim for the Principle of Least Collaborative Effort in relation to the route finding task we discussed in the previous Section. Low Criterion (LC) pairs used a lower degree of Collaboration than High Criterion (HC) pairs because as their criterion was lower they needed less exact information, and could therefore spend less time (effort) on the task. As we have pointed out before, although the effort decreases, there is no particular reason to say that the participants are choosing least joint effort rather than just minimising their own effort. In fact, the argument for Least Individual Effort fits rather better than Wilkes-Gibbs’ when it comes to the mixed pairs. HC Directors ac-
commodated downwards (less Collaboration) to LC Addressees; LC Directors do not accommodate upwards to HC Addressees. Wilkes-Gibbs’ explanation is that HC speakers are more willing to accommodate to their partner’s needs. What could equally be argued is that both speakers, individually, wish to minimise their effort. HC Directors will gladly take the opportunity to decrease their effort level down to the LC Addressee’s level because it benefits them (meets with the Principle of Least Individual Effort). However, LC Directors won’t adjust their effort because it would involve an increase in effort, which they do not see to be absolutely necessary: it would go against their Principle of Least Individual Effort. It should also be remembered that HC participants in mixed pairs did significantly worse on the test than any other condition, so their partner’s minimisation of effort adversely affected their understanding and task result. How can this be explained through a Principle that emphasises the jointness of language? Indeed, Heeman and Hirst (1995) interpret Clark and Wilkes-Gibbs’ (1986) concept of collaboration as:

“... [allowing] the agents to interact so that neither assumes control of the dialog, thus allowing both to contribute to the best of their ability without being controlled or impeded by the other.”

Heeman and Hirst (1995:367)

This is not what we see here. Wilkes-Gibbs’ LC interactants can be seen to both control the dialogue and impede their partner in their aims. Surely this behaviour could be better explained by the joint production of language being reliant on the lowest criterion speaker, regardless of the efforts of the other speaker? Language may be a joint production, but it is reliant on the foibles of individuals.

Schober’s continuation of this work (Schober, 1995) is more convincing. He ran an experiment to examine what perspective participants used when describing the location of an object in a circular sheet of cardboard (the material used looks rather like a pizza with odd toppings (Schober’s comment)). Each of the pieces of ‘topping’ were numbered, and the Director had to describe the location of each number, so that the Matcher could number their ‘topping’ appropriately. There was also an added dimension to this task, there were three position conditions: the Matchers’ boards were either unchanged, rotated 90°, or rotated 180°15. Schober argues that one can see whose effort is being minimised by seeing whose viewpoint

---

15The participants were made aware of these offsets before the task began.
the Director takes. This is particularly an issue where the Matcher’s board is offset from the Director’s. Schober’s argument is presented below:\footnote{This is simplified in order to concentrate on the issues of relevance here. See Schober (1995) for the full experimental details.}

- Director’s viewpoint = minimum effort for Director, maximum for Matcher
- Matcher’s viewpoint = minimum effort for Matcher, maximum for Director
- Neutral viewpoint = minimum joint effort

By ‘neutral location descriptions’ Schober means references which are not dependent on either view: they are true whatever the speakers’ frames of reference. He gives the following as examples: near the lamp, between the car and the house, and in the middle of the room (taken from Schober (1995:222)). Schober claims that as the neutral viewpoint is neither maximum or minimum effort for either speaker, then it must therefore constitute the minimum joint effort. As we have discussed earlier in this Section, such an assumption is not necessarily true. It may well involve a redistribution of effort, but we have no meaningful way of measuring individual efforts before or after such a redistribution. Neutral viewpoints may be less than maximum effort, but it does not follow that (2 times neutral viewpoint) ≤ (1 maximum effort + 1 minimum effort), the reverse could as easily be true.

Each pair of interactants did the task three times. Schober found that the Directors tended to use more Matcher-centred descriptions initially, but they also gravitated towards using more and more neutral location descriptions over time, whatever the position condition (regardless of the degree of offset). He therefore argued that if the participants had only wanted to minimise effort for the Matcher, then they could have just used Matcher-centred descriptions. As they ended up using more neutral descriptions, then we should conclude that the participants wished to minimise Collaborative Effort.

The change in strategy over time from Matcher-centred to neutral descriptions does indeed suggest a decrease in effort on the part of the Director, but does it necessarily support the concept of Least Collaborative Effort? Such a change may have made relatively more work for the Matcher than the decrease in the Director’s work would have justified - but this is extremely hard to measure. It may be even that the Director realised that the Matcher would allow them to use a lower effort strategy, and therefore the Director changed to the lower effort
approach with little thought of the impact on their Matcher\textsuperscript{17}. Maybe Schober’s results merely show a decrease in effort on the part of the Director: the seeming equalisation of effort could be a side-effect. We would argue that such a change in approach is equally well explained by the Risk-Effort Trade-Off, a concept which we will discuss in the next Section.

Garrod and Clark (1993) also point out the difficulty in distinguishing between Least Collaborative Effort and Least Individual Effort. They argue that their Principle of Output-Input Coordination (Garrod and Anderson, 1987) conforms to the Principle of Least Collaborative Effort. Their Principle states that:

**Principle of Output-Input Coordination:**

"output/input coordination, ... may be simply stated as one of formulating your output (i.e. utterances) according to the same principles of interpretation (i.e. model and semantic rules) as those needed to interpret the most recent relevant input (i.e. utterance from the interlocutor)."

Garrod and Anderson (1987:27)

In terms of Semantic Coordination (which is the locus of their interest) this means that a speaker should base their next description on the last one that they have interpreted successfully. That is, they should stick to a semantic model that has been shown to be understood. Garrod and Anderson showed that this approach helps speakers to establish a common ground for how they refer to entities, and thus it minimises Collaborative Effort. But equally, it also minimises Individual Effort because the speaker simply uses the same frame of reference for production as they have used for comprehension. Again, it is impossible to distinguish between these two Principles on the current evidence. We will investigate this relationship further in Chapter Six.

### 3.2.3.4 The Role of the Listener

Although Clark and his co-workers emphasise language being joint action, and the importance of the Collaborative process, the Addressee is still seen as the passive participant. It is the Director of these tasks that invites the addressee’s

\textsuperscript{17}It may be argued that if Matchers were unhappy with this change, then they would revert to Matcher-centred descriptions when they are given the chance to be a Director - this at least appears to be Schober’s implicit argument. This, however, assumes a very unselfish Director, and, particularly, one who is averse to the idea of revenge!
participation, and decides what approach should be used (Clark and Wilkes-Gibbs, 1986; Schober, 1995). Yet in Wilkes-Gibbs’ (1986) research, it is the Low Criterion participant who seems to determine the course of the conversation, whether they are currently in the role of Director or Addressee.

We would argue that Wilkes-Gibbs’ results are more representative of the reality of language use: the results of a conversation are dependent on the input and attitude of both Director and Addressee. We will come back to this issue in Section 3.4.3.

3.2.4 Dialogue as Risk-Effort Trade-off

When taking part in a dialogue, a speaker has to decide how they intend to achieve their current goal. Typically, they will have several options open to them, which will take more or less Effort for them to formulate, and, conversely, imply more or less Risk that the intended goal will be achieved first time. How will the interactant decide which approach to take? According to Shadbolt (1984), we decide according to the Principle of Parsimony, which is

**Principle of Parsimony:**

"... a behavioural principle which instructs processors to do no more processing than is necessary to achieve a goal."

Shadbolt (1984:342)

In other words, a participant will choose the approach which will be the least effortful (most risky) that will succeed.

What is meant by the terms ‘effort’ and ‘risk’? Effort is the amount of work that you put into the construction of an utterance: an effortful utterance would consider such things as shared knowledge, making sure references are interpretable and ensuring that the utterance is interpretable given the current focus. It is, essentially, trying to make the utterance as easy to understand as possible. An utterance constructed with little effort would have little concern for the Hearer’s ability to understand it: it would simply be the least effortful way of tackling the next goal. In terms of real dialogue, it might be the difference between directing a tourist to Waverley station via North Bridge (low effort), or asking to check
whether they knew where North Bridge was, before formulating the direction (high effort)\(^{18}\).

Risk is concerned with the likelihood of the Speaker achieving their current goal without redress to a repair sequence. A low risk utterance is more likely to work as intended, conversely a high risk utterance may work as intended, but may also fail, and lead to the need for a repair. In terms of the examples we gave above, the low effort approach is high in risk: the Speaker has not checked whether North Bridge is shared knowledge, so the plan will fail if that piece of information is not shared. The high effort approach is low in risk because the Speaker is taking account of possible differences in knowledge, so their plan should work first time, regardless of any differences in shared knowledge.

So why should participants choose the more risky options? If, in our example, the tourist does know where North Bridge is, then asking them whether they know about it before embarking on the directions is wasted effort - the directions via North Bridge would work whichever approach was taken. So, high risk Speakers take a chance that their assumption of shared knowledge is correct: if they’re right, then they use less effort and ‘win’. If they’re wrong, then they may well end up using more effort to repair the initial sequence than would have been needed in the original low risk option. It’s a gamble, but one, Shadbolt argues, that we take all the time.

This decision that we make is called the ‘Risk-Effort Trade-Off’, as we try to find the point of least effort - and thus most risk - with which we can still achieve the original goal.

As it stands, this approach to language does have some problems. Firstly, like many of the principles we have examined here, it is based on a small aspect of language: particularly whether we ask about the existence of landmarks before using them in a route description (Shadbolt, 1984; Carletta, 1992, 1996b). The question must arise whether this Principle is generalisable to broader language use, and if so, how it should be characterised. We will address this more closely in Section 3.4.4. Secondly, and perhaps more importantly, the Risk-Effort Trade-Off as presented here has a major fault: it assumes that any failure due to a high risk choice will be both discovered and repaired. This, in human dialogue, is not a safe assumption. Communication failures can easily continue undiscovered, and the participants will never know that they made a mistake. Even the discovery of

\(^{18}\)The concept of Effort, and its importance in the empirical work described in this thesis, is discussed in detail in Section 5.3
such problems does not guarantee their resolution. Participants sometimes decide (for whatever reason) to ignore the problem, and do not try to resolve it. Or, they may try to redo that part of the route, and fail again. We cannot always guarantee to communicate successfully, however hard we try. So, for the purposes of this research, we will use the following definition of Risk.

**Risk:**

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

It is likely that these two limitations of the Principle of Parsimony and the Risk-Effort Trade-Off derive from the way in which it was developed. As we have described previously (Section 2.5.2), these two concepts arose from informal investigation rather than rigorous empirical work - the linguistic aspect of this work was not the focus of the projects in question, and therefore they have not been empirically tested. Part of the intention of our work was to investigate the usefulness of the Principle of Parsimony and the Risk-Effort Trade-Off in practice.

### 3.2.5 The Principle of Parsimony and the Principle of Least Individual Effort

In the discussion of the Principle of Collaboration, we have introduced the Principle of Least Individual Effort (Section 3.2.3.3). We have argued that this Principle provides a far more convincing explanation of the behaviours considered to by Clark and Wilkes-Gibbs (1986) and Wilkes-Gibbs (1986) to be representative of their Collaborative Principle.\(^{19}\) This is because Least Individual Effort is more reliant on the idea of *individuals* contributing to a dialogue rather than language being jointly produced: the latter seems to have less selfish connotations which we do not believe to be upheld by their data (particularly Wilkes-Gibbs, 1986). But why is this relevant to the Principle of Parsimony?

We would argue that the Principle of Parsimony is equivalent to the Principle of Least Individual Effort. Parsimony demands that a speaker uses "...no more processing than is necessary to achieve a goal." (Shadbolt 1984:342), which could be seen as a more computational way of expressing the concept of Least Individual Effort. The way in which the Principle of Parsimony is demonstrated in terms of the Risk-Effort Trade-Off also leads to this conclusion. Speakers seek

---

\(^{19}\) The behaviours in question are: (1) The importance of matched commitment to task success and (2) The lack of relationship between high effort and task success.
to use language which needs the least effort required to achieve their goal. As Effort is inversely related with Risk, speakers must work out what level of risk is acceptable, and therefore what constitutes the lowest effort they can manage.

Thus, the Principle of Parsimony and the Risk-Effort Trade-Off are reducible to the Principle of Least Individual Effort. The Risk-Effort Trade-Off provides a way of both explaining how the Principle of Least Individual Effort could work, and how to investigate its behaviour empirically. Support for the Risk-Effort Trade-Off would imply support for the Principle of Least Individual Effort. For the remainder of this thesis, reference to the Risk-Effort Trade-Off will imply reference to the Principle of Least Individual Effort and the Principle of Parsimony.

3.3 Definitions of Cooperation

The aims of this Section are twofold. Firstly, we will discuss the notion of ‘cooperation’ and the difficulties involved in defining a useful term. Secondly, we shall suggest that the concept of ‘cooperation’ can usefully be based on the concept of Effort (discussed above).

In Section 3.2.2 we discussed the Pollyanna Principle, and the tendency of linguists to concentrate on communication successes rather than communication breakdowns. We are seen as being ‘good’ at dialogue - phenomena such as managing turn-taking using only tiny gaps between speakers, the lack of overlaps (interruptions) in conversation, and our ability to repair miscommunications are often cited to support such an argument. Being ‘good’ is then often seen as presuming being ‘cooperative’. This is an oversimplification in several respects.

Firstly, there is a difficulty in deciding what the term ‘good’ refers to here: this could relate to either the interpersonal or task level of the dialogue, or both. The interpersonal level is certainly important, and these phenomena may be relevant, but is extremely hard to quantify. However, if you are more interested in the outcome of the dialogue, then the dialogue phenomena described above may not have a straightforward relationship to the efficacy of the dialogue. For example, Thompson’s (1996) work on the HCRC Map Task Corpus suggests that interruptions and overlaps may well contribute to task success20.

In this study, we will concentrate on the task success element of a ‘good’ dialogue. This does not mean to say that we will ignore the interpersonal aspect - the two

\[20\text{Spearman's Rank correlation between the deviation of the drawn route and the amount of overlaps is } p \leq 0.05.\]
aspects are in fact inextricably bound together - rather that we will consider how the dialogue relates to the task rather than, say, issues of face (Goffman, 1967) or politeness (Brown and Levinson, 1987; Leech, 1983).21

Secondly, ‘good’ compared to what? Dialogues vary in how successful they are: little exposure to a dialogue corpora will soon demonstrate that speakers go about tasks in different ways. What makes one dialogue more successful (in task terms) than another? This is a question that Clark and his co-workers do not seem to address, partly because their dialogues seem to be presented to the reader as a homogeneous mass, rather than having some variation. Brown’s (1995) qualitative study addresses this issue more, but she does not make any explicit numerical comparison between dialogue variation and task success.

Thirdly, there is also a danger in assuming that the terms ‘good’ and ‘cooperative’ are necessarily synonymous. Depending on how we define ‘cooperation’, it may be that an effective dialogue is different to what we might want to term a ‘cooperative’ dialogue. It is partly this interaction that we wish to study: are more ‘cooperative’ dialogues more successful? Can you have a successful dialogue that isn’t ‘cooperative’? Or is there, perhaps, some sort of trade-off? We will return to this issue later in this Section, and also in the discussion of our empirical results (Section 6.4).

So, if we intend to define a ‘good’ dialogue as one which successfully achieves its goal, how are we to define a ‘cooperative’ one? The first issue we need to address here is that of ‘cooperation’ (however you define it) as being a gradeable concept. There is no distinct dividing line between dialogue which can be classed as ‘cooperative’ and those which can’t: cooperation is a cline. We will use a contrasting example to demonstrate this. Geographical borders are absolute dividing lines. When I cross the border at Gretna Green, I go from being in the part of Scotland near England, to being in the part of England near Scotland. There is no such point of change between ‘cooperative’ and ‘uncooperative’. Our definition of ‘cooperation’ will be similarly gradeable - we will only be able to describe a dialogue as being more, or less, cooperative.

In general usage, ‘cooperation’ means “working together”. How can we translate this into a metric useful for language? We propose that the concept of Effort discussed earlier (Section 3.2.4) should provide the basis. Effort could be described as the linguistic evidence for cooperation, whatever the underlying reason behind

21In fact, Leech (1983) claims that his Politeness Principle is of less relevance in task-oriented dialogue, because the focus on the task outweighs many of the politeness considerations we may have in casual conversation.
To work together, you invest effort in order to achieve your own, your partner's, and your joint goals. The more effort you invest, the more you could be considered to be trying to achieve those goals. This is how we would like to envisage cooperation, as the effort you put in to achieve not only your own but also your partner’s and your joint goals. It should be noted that this exhortation should apply equally to both Speaker and Hearer. Brown’s (1995) definition of “cooperative conversation” comes closest to what we intend:

**Cooperation in Dialogue:**

“...communication does not consist of a fail-safe exchange of the same thought, but is, rather, 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.”

Brown (1995:16)

So, Cooperation is to be measured by the amount of Effort you invest in a creating a useful utterance. We make the assumption that interactants will ensure that their effort is directed appropriately: otherwise it is not worth their while to invest the effort in the first place.

**Operational Definition of Cooperation:**

*The more effort invested by the speaker, the more cooperative that utterance is perceived to be.*

Therefore, the harder you work at a dialogue, the more cooperative you are considered to be. This, of course, begs the question of how one measures Effort. We outline our plan for this in the discussion of the Typology of Move Attributes.

### 3.3.1 Gricean Cooperation versus Cooperation

Now that we have evaluated the concept of Cooperation, it would seem appropriate to re-iterate its differences from Grice’s Cooperative Principle. As we will see in the empirical definitions of these Principles, there is a large contrast in the underlying motivations of these two Principles.
No interpretation of Grice’s work should equate to dialogue being ‘helpful’. Grice is more concerned with how we exploit language conventions, and thus minimise speaker effort (often at the expense of the addressee). Addressee’s are made to work harder - this is scarcely ‘helpful’ to them.

On the other hand, Cooperation, as we have defined it above, is entirely concerned with ‘helpfulness’. Cooperative speakers work hard to help each other, investing effort in the dialogue in order to work at a good task result.

These differences will be apparent in the definitions below. Grice may be interpreted in many ways, but his broad approach is always going to be different to that of ‘Cooperation’.

3.4 Cooperation in an Empirical Context

In this Section, we will take each of the structuring principles we have examined in turn, and suggest how their view of communication could be instantiated in task-oriented dialogues. These suggestions will be presented in the form of numbered hypotheses, each of which will be justified in relation to the Principle they represent.

3.4.1 Gricean Cooperation

Formulating empirical hypotheses based on Grice’s views is somewhat more awkward than for the other principles we have discussed. This is because Grice’s work was not based in a empirical framework. The suggestions we make, therefore, can only be a limited representation of Grice’s views. What is important here is to demonstrate the similarities and differences from the alternative approaches.

Firstly, we need a definition of Gricean Cooperation which is generalisable to dialogue. In Section 3.2.1 we suggested two possible interpretations of Grice: dialogue as a rational activity or dialogue as an efficient activity. These two explanations are in fact closely related, and can just be seen as two angles on the same or similar behaviour. Indeed, to behave efficiently is entirely rational.

i. Speakers will avoid unnecessary effort

Although speakers should be committed to doing the work necessary to the task, they are not expected to do any more than that. The Cooperative Principle makes it possible for speakers to decrease their effort, and thus meet this ideal.
ii. Speakers will improve at tasks
Agents should have the ability to learn. In terms of task-oriented dialogues, we would expect agents to produce better task results over time. The motivation to learn is based on the desire to decrease effort: the agent learns the minimum that is required to do the task.

iii. Speaker effort will decrease
This hypothesis is linked to hypotheses i and ii. 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.

A redistribution of effort would also support this hypothesis. This would be evidence of speakers learning where effort can be minimised, and where it should be increased, in order to gain a good task result.

3.4.2 Coordination
We will concentrate here mainly on the ideas of salience and precedence, as these are the areas most appropriate to our study: we have not looked at turn taking behaviour, for example, which might be explained in terms of convention. This, too, is the area which is focused on by both Clark and his co-workers (Clark and Wilkes-Gibbs, 1986), and Garrod (Semantic Coordination) (Garrod and Anderson, 1987; Garrod and Clark, 1993; Garrod and Doherty-Sneddon, 1994).

iv. Task result will improve as speakers converge on a convention
Speakers will coordinate on an approach to a task. This Coordination should improve task result as the speakers are less likely to have conflicting goals, or to be talking at cross purposes. One would also expect speakers to converge on a useful approach, which should also contribute towards task success.

Why this should happen is unclear. It could be suggested that speakers just keep trying different behaviours until they find one that’s successful, but this explanation is just as problematic. Firstly, why should speakers converge? There must be many ways to complete the Map Task successfully, so why do speakers all seem to gravitate towards a similar approach? The Principle of Coordination does not motivate this hypothesis - this and the following hypothesis seem to be entirely descriptive, rather than considering why this behaviour occurs.

Notice that this hypothesis is effectively the same as hypothesis ii above.
v. Speakers will converge, and therefore their behaviour will change
As speakers converge, their behaviour should change, as it is unlikely that they will automatically choose the same strategy as their partner. The Principle does not predict or motivate what type of behaviours speakers are likely to converge on.
This hypothesis is reliant on hypothesis iv being upheld.

3.4.3 Collaboration and the Principle of Least Collaborative Effort

In this Section we will consider what hypotheses would represent both the Principle of 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. We will use Clark and Wilkes-Gibbs (1986) and Wilkes-Gibbs (1986) numerical (quantitative) definition of Collaboration, rather than the more qualitative view used by Brown (1995) or the descriptions of Collaborative behaviour also found in Clark and Wilkes-Gibbs (1986).

vi. Both speakers contribute to the completion of the task
This is not a particularly interesting hypothesis; however, it does represent the definition of Collaboration given by Clark and Wilkes-Gibbs (1986) and Wilkes-Gibbs (1986). Collaboration means that participants talk to each other, contributing words and turns. More talk and turn-taking means more Collaboration. The fact that speakers obviously do talk to each other in order to complete the Map Task is sufficient evidence to support this hypothesis, and we will not test it any further.

vii. 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 Principle of Least Collaborative Effort is difficult for us to measure in Schober’s (1995) terms because participants in the dialogues which form our dataset had the same perspective - there were no offset conditions. We will rather concentrate on the explanation of Wilkes-Gibbs (1986), where evidence for Least Collaborative
Effort is taken from the decrease in the number of turns taken to finish the task.

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

This is related to the previous hypothesis. Participants need less Collaboration (Effort), therefore they use less. In Clark and Wilkes-Gibbs (1986) this is demonstrated by the decrease in dialogue length when pairs repeated the tangram task. This was due to an increase in common ground. However, the Map Task is a repetition of form rather than content and therefore doesn't exhibit such a marked increase in common ground. Therefore, a measure of average effort was considered more appropriate than a measure of absolute effort.

ix. 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 empirical study, high criterion will be considered to be equal to high effort, and, similarly, low criterion will be considered equivalent to low effort.

x. There is no correlation 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 (1986) gives three indicators of high collaboration:

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

In Wilkes-Gibbs (1986), having the same criterion as your partner was far more relevant to task success. The Collaborative view would be weakly supported by the acceptance of the null hypothesis²²

²²Null hypotheses are not statistical proof. They can only state that there is insufficient evidence to support the experimental hypothesis, not that the reverse of the experimental hypothesis is necessarily true. However, one can argue that the direction of the data is consistent with the reverse of the experimental hypothesis.
High effort and equal commitment These two hypotheses are, in a sense, complementary. In Wilkes-Gibbs’ work, the fact that high effort in itself does not have an impact on task success strengthens the significance of the importance of equal commitment. The matched criteria of individuals is more important than the absolute amount of effort that they invest. Both these hypotheses being supported in our own work would uphold the importance of individual input in the same way.

Least Individual Effort Note that the hypotheses relating to Least Collaborative Effort apply equally well to the Principle of Least Individual Effort, as was argued in Section 3.2.3.3. In fact, we would argue that, in fact, these hypotheses apply better to the Principle of Least Individual Effort, as there is no mechanism to measure Collaborative Effort, but shifts in effort are easier to detect. This issue was discussed in Section 3.2.3.3, and we will return to it in the discussion of the results relating to these hypotheses in Chapter Six.

3.4.4 Principle of Parsimony

In this Section we assume that the speakers would be trying to find the most efficient point in the Risk-Effort Trade-Off.

xi. 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.

xii. 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.

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

The last hypothesis leads directly to this task-level hypothesis. If speakers work out the best point in 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 hypotheses ii & iv, although the motivation
xiv. Behaviour would modify as speakers try out different Risk-Effort combinations, and eventually settle on a set of useful combinations

The Risk-Effort Trade-Off would lead speakers to try out various strategies, until they found one which satisfied their constraints. This is similar to the predictions of hypothesis v (Coordination), which states that behaviour will converge, although unlike this Principle, the change has no explicit motivation. Hypothesis iii (Gricean Cooperation) also predicts behaviour modification, but unlike the Gricean hypothesis, this one does not predict the direction of the modification: speakers are equally likely to start from either a high or low risk posture, and thus may expect to adjust down, or up, respectively.

The Principle of Parsimony and the Principle of Least Individual Effort

As we argued earlier (Section 3.2.5), any hypotheses relating to the Principle of Parsimony and the Risk-Effort Trade-Off should be considered to be supporting evidence for the Principle of Least Individual Effort. The relationship between Risk and Effort is a way of formalising how the Principle of Least Individual Effort could work in language.

Motivation for Change

There are two elements to this series of hypotheses, descriptive ones, and motivating ones. Changes in task success and dialogue behaviour are descriptive: we can postulate why they might occur, but essentially we are viewing the available phenomena. However, if the hypotheses which relate directly to Risk and Effort (hypotheses xi & xii) are supported, then we would also have evidence for the motivation behind the changes we predict in task success and dialogue behaviour. We would show that speakers take risks, but that they learn what constitutes risky behaviour in this context, thus decreasing the level of risk, and producing the changes that we can measure elsewhere. None of the other principles discussed here can motivate as strongly how behaviour might change in order to produce a better task result. Gricean Cooperation23, Coordination and Cooperation predict a learning effect, but it is not clear how this would be evidenced in the dialogues themselves.

23We do predict a decrease in effort for Gricean Cooperation, which could be argued to motivate learning and behaviour change, but this hypothesis is not part of the supportive network of hypotheses we see here. If all of them are shown to be supported by the data, then the overall support for the motivation behind Risk-Effort would be that much more.
3.4.5 Cooperation

This assumes the definition of greater Cooperation meaning more effort, as described above.

xv. 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.

xvi. Low effort would be associated with low task success
Not being helpful - not investing effort - should have a detrimental effect on task success.

xvii. 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. Less 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 is upheld).

xviii. 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 iii, v & xiv.)

xix. 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.

It should be noted here that although the reason for improved task success is considered to be a learning effect for both the Principles of Cooperation and Parsimony, the motivation for learning is quite different. The Principle of Parsimony suggests the factors of Risk and Effort. As interactants gain experience of a task, they work out a better trade-off between Risk and Effort, and are therefore less likely to make mistakes which lead to worse task performance. Cooperation, on the other hand, suggests that interactants should orient towards useful strategies (and thus better task results) in order to help their partner.

There is an important difference here: speakers following the Principle of Cooperation learn in order to improve the joint result, whereas speakers following
the Principle of Parsimony learn in order to decrease their effort. For the latter, the improved task result may well be a side-effect rather than the focus of the learning process. This may also be true of the speaker orienting towards Gricean Principles, who also learns in order to minimise effort.

(Similar to hypotheses ii, iv & xiii.)

High effort and low effort  It should be obvious that hypotheses xv and xvi are strongly related: they could be seen as the opposite sides of the same prediction. If both are upheld, then the interpretation is simple. However, if one is not proven, then the interpretation of the other’s result will be affected.

If hypothesis xv is supported and hypothesis xvi isn’t, then the relationship between effort and task success would be unclear. High effort would improve task result, but low effort would not affect it. This seems somewhat contradictory, and such a strange result would have to undermine the support given to the Principle of Cooperation as a whole. However, such a result would not obviously provide support for any of the alternative Principles presented here.

What happens if hypothesis xvi is supported by the data, but hypothesis xv isn’t? In itself, the empirical finding supports the view that low effort would lead to a worse task result, but we would argue that you cannot then make the step which says this is the case because this behaviour is not helpful, and dialogue should be helpful. If this were to be the case, then we would have to view the result as not supporting the motivation behind the original hypothesis, even though the empirical result could be said to support it.

In this situation, we would argue that this result would provide more evidence towards the Principle of Parsimony and the Risk-Effort Trade-Off. Low effort would presume a high degree of risk, and a high degree of risk is likely to lead to task errors, and thus a poor task result. This explanation would provide a suitable motivation for the above result. If this combination of results should occur, then we will assume that it will provide support for the Principle of Parsimony rather than the Principle of Cooperation.
3.4.6 Summary of Hypotheses

Here, we will briefly recap the hypotheses in the form of a series of empirically testable hypotheses so that it is clear both where more than one structuring principle predicts the same outcome, and where conflicting outcomes are predicted.

1. **High effort would be associated with task success**  
   Hypothesis xv (Cooperation) - effort *would* correlate with success  
   Hypothesis x (Collaboration) - effort/collaboration *wouldn’t* be associated with success (*null hypothesis*)  
   **Note conflict between hypotheses**

2. **Speakers with equal commitment (whether high or low) should be associated with more task success**  
   Hypothesis ix (Least Collaborative Effort)  
   (by implication Hypothesis ix suggests Least Individual Effort)

3. **Risks would be taken**  
   Hypothesis xi (Principle of Parsimony) - risks *would* be taken  
   Hypothesis xvii (Cooperation) - risks are *unlikely* to be taken  
   (*null hypothesis*)  
   **Note conflict between hypotheses**

4. **Low Effort would be associated with low task success**  
   Hypothesis xvi (Cooperation)

5. **Dialogue Strategies will change over time**  
   Hypothesis iii (Gricean Cooperation)  
   Hypothesis v (Coordination)  
   Hypothesis xiv (Principle of Parsimony)  
   Hypothesis xviii (Cooperation)

6. **Risks would decrease over time**  
   Hypothesis xii (Principle of Parsimony)

7. **Effort will decrease - effort minimised**  
   Hypotheses i & iii (Gricean Cooperation)  
   Hypotheses vii & viii (Collaboration / Least Collaborative Effort)  
   (by implication vii & viii suggest Least Individual Effort)
8. Task success will improve over time  
   Hypothesis ii (Gricean Cooperation)  
   Hypothesis iv (Coordination)  
   Hypothesis xiii (Principle of Parsimony)  
   Hypothesis xix (Cooperation)   

Relationship between task success and low effort/risk  Hypotheses 4, 6 and 8 are closely related. If low effort correlates with task success, and risk\textsuperscript{24} decreases over time, then it follows that we would also expect task success to improve over time. If this were not the case, then the relationship between task success and low effort/risk would be questioned. If all three are supported, then it would imply a strong relationship between low effort/risk and task success. Of course, hypothesis 5 is also dependent on hypothesis 3 being supported: risks cannot decrease, if they do not happen in the first place.   

Changes in Behaviour  Hypotheses 6 and 7 are reliant on Hypothesis 5 being supported. If no changes in behaviour (dialogue strategies) are found, then there would be no point in seeking the more specific changes in either the level of risk (Hypothesis 6) or the level of effort (Hypothesis 7).   

3.5 Summary   

In this Chapter we have attempted to operationalise the concepts of Gricean Cooperation, Collaboration, Coordination, Folklinguistic Cooperation and the Risk-Effort Trade-Off. This has involved a careful analysis of the literature in each of these areas, in order to generate useful empirical definitions.   

Two important theoretical areas of discussion have also arisen. These we will outline briefly below. Firstly, the question of the use of the term 'Cooperation'. We have highlighted the need to distinguish between the intent of Grice's writings about the Cooperative Principle, and the more general use of the term 'Cooperation' in the linguistic literature. As we have discussed earlier, to interpret absolutely Grice's intent is not straightforward, but we can differentiate between some concept of effort minimisation and efficiency on the part of the speaker (a Gricean interpretation) versus some notion of speaker effort and 'helpfulness' as described in Brown (1995). We also pointed out the need to view all these  

\textsuperscript{24}Which is effectively low effort
concepts as having gradable rather than binary values: a distinction between 'cooperative' and 'non-cooperative' is a difficult one to make, whereas 'more cooperative' versus 'less cooperative' may be more useful.

Secondly, we have identified some problems within the Clarkian model of 'Language as Joint Action'. The concept of Collaboration suffers from a confusion of qualitative and quantitative definitions, and a lack of a concrete relationship between Collaboration and dialogue outcome: a Principle which appears to go nowhere. More importantly, we question the argumentation which is used to support the Principle of Least Collaborative Effort - a foundation stone of the Clarkian model. Experimental evidence which they cite in support of this Principle seems better explained by a Principle of Least Individual Effort, derived from the Principle of Parsimony and the Risk-Effort Trade-Off. We argue the view that Least Individual Effort should be taken as a replacement for the Principle of Least Collaborative Effort in the remainder of this thesis.

The last part of this Chapter sets out how we intend to translate empirical definitions of the concepts above into testable hypotheses. The background to this experimental work is explained in the following two Chapters, and the results are presented in Chapter Six.
Chapter 4

The Empirical Background

4.1 Introduction

In this Chapter we will describe the basic materials of our research. This will involve a description of the data, how it was collected and the empirical tools used.

We will explain the structure of the Map Task Corpus (Anderson et al., 1991a), and how its design makes it particularly appropriate to studies concerned with communicative success.

4.2 The Map Task Corpus

This is a corpus of task-oriented dialogues, involving 64 participants who produced 128 dialogues (16 hours of speech). This corpus was designed to investigate a number of interesting research variables, some of which we will describe below ¹. Our data set comprises a subset of this corpus.

4.2.1 The Subjects

The participants in the task were first year undergraduate students from the University of Glasgow. They were asked to volunteer with a friend, in order that the variable of familiarity could be manipulated within the task design.

¹A full description of the design and methodology is given in Anderson et al. (1991a)
4.2.2 The Task

Each speaker in a dialogue is designated as either ‘Giver’ or ‘Follower’. The Giver has a route drawn on the map in front of her, the Follower has a similar map, but no route (see Figure 4.1). Their joint task is to enable the Follower to replicate the route as accurately as possible on his map. The route which the Giver has on her map is based around a number of small pictures (known as features or landmarks), but unfortunately for the participants, not all of these features are on both the Giver’s and Follower’s maps (shared features). Out of the 12 features (approximately) which occur on each map, about 8 will be shared and the remainder unshared (occur on only the Giver or the Follower’s map). It is this problem of unshared knowledge which provides the difficulty of this task. As some of the essential route-pivots for the route will be omitted from either map, the participants have to become involved in an exchange of information in order that the task be completed successfully. No one participant has all the information required.

The two participants were seated opposite one another, and were unable to see the other person’s map. They were either separated by a board which obscured the other speaker completely (no eye-contact), or the board came up to shoulder level so that the speakers could still see each other’s faces (eye-contact). The participants who were involved in the task under the condition of eye-contact were instructed not to use non-verbal gestures during the conversation: the half-height screen was intended partly to shield any such gestures from view.

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 route’ around the obstacles, 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 of difference, 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.

4.2.3 The Design

The corpus was collected in a series of “quads”. A quad is a set of eight dialogues, produced by a group of four speakers. Within each quad, the data collection was
Figure 4.1: Example Map from the HCRC Map Task Corpus
This map was used in dialogues EAQ4C1, EAQ4C7, EAQ8C1 & EAQ8C7 in our subset of the Corpus.
based on a Latin square design. This means that each participant would interact with two of the three other speakers in the quad: their original partner, and a member of the other partnership. Each participant was involved in four dialogues, two with their familiar partner, and two with the unfamiliar one. Each speaker took the role of Giver and Follower twice. They ‘follow’ a map from both their familiar and unfamiliar partner. As a Giver, the participant ‘gives’ the same map twice, albeit to different Followers. This part of the design enables us to study the differences (if any) between first and second givings.

The whole Map Task Corpus consists of 16 such quads. These are split equally between the conditions of eye-contact and no-eye-contact.

### 4.2.4 Materials

Within the corpus as a whole, 4 different sets of 4 maps were used (16 in all). Therefore, each set of maps was used four times, twice in in eye-contact conditions, and twice in no eye-contact conditions. The repetition of the maps within each of the eye-contact conditions was to balance for the effect of familiar and unfamiliar dialogues: in the first running of a set of maps the first two conversations would be unfamiliar pairs, and in the second running they would be familiar pairs, and so on. This would clarify whether any effect occurring over the duration of a quad (e.g. task success) could be attributed to familiarity or some other factor.

Each set of maps was also internally balanced. There are four different route shapes used in the corpus, each of which has 4 instantiations, which are topologically the same. An example of two such topologically similar maps are given in Figures 4.1 & 4.2. They vary only according to the particular feature which occurs in a particular location, and which of the features are unshared Giver features or unshared Follower features. In the example here, pyramid has been replaced by slate mountain, banana tree has been replaced by tribal settlement and is now an unshared Giver feature, rather than an unshared Follower one, conversely, poisoned stream has changed to machete and become an unshared Follower feature. Each set of 4 maps includes one of each route shape, and the ordering of the route shapes is cycled across the sets: i.e. route shape 1 will be the first map tackled in one quad, the second in another, and so forth. Again, this is to control the effect of route shape in the corpus as a whole. In the example given here, the crest falls map is the fourth map to be used in quad three, but the extinct volcano map is used first in quad four.

88
Figure 4.2: Example Map from the HCRC Map Task Corpus
This map was used in dialogues EAQ3C4, EAQ3C6, EAQ7C4 & EAQ7C6 in our subset of the Corpus.
There is one other important attribute of the maps which should be mentioned here. Each of the maps has a **duplicate feature**: a particular landmark which occurs twice on the map. One of these features is necessary to the route, the other is irrelevant. The Giver has both these features, whereas the Follower only has the irrelevant one; unsurprisingly, this can cause some referential difficulties. The proximity of these features on the map varies, and this can provide some insights into what constitutes a relevant focus space (what size area of the map is salient) when attempting such a task. The duplicate features on the example maps here are *poisoned stream* in the *crest falls* map, and *golden beach* in the other map. In terms of proximity, the two *golden beaches* are closer, and tended to cause more problems.

### 4.3 Advantages of the Design

We now explain why the elements of this design are particularly appropriate to studying the communicative process in general, and communicative success in particular.

Firstly, the dialogues - although very natural sounding - are constrained in comparison with casual conversation because of their task-oriented nature. The advantage of this is the greater certainty with which we can posit the goals of any given utterance. This makes it possible to study the approach interactants take to these goals, as the dialogues tend to be far more straightforward than casual speech. Therefore, we can make reasonable predictions of what choices are appropriate at a particular point, given that the speaker wishes to achieve a certain goal.

The potential problems of studying speaker intention and mutual knowledge are lessened further by the limiting of communication channels. The board between the speakers effectively barred non-verbal gestures, so that information content had to be carried by the verbal channel alone. In the *eye-contact* condition, a certain amount of communication could potentially be achieved through gaze, but this is mostly limited to minimal responses, which is not the locus of our interest.

The main advantage of the Map Task Corpus over most other task-oriented corpora (e.g., Grosz and Sidner, 1986) 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 between being the expert and the novice). In this corpus, the speakers have to negotiate
the transferral of information if they are to succeed. It is not sufficient merely for the Giver to describe the route to the Follower: without the knowledge of the unshared features, it is likely that the route drawn by the Follower will negotiate some aspects of the route incorrectly. This is important because the dialogue is therefore 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. In expert-novice dialogues, the responsibility is more with the expert than the novice. It is precisely how speakers overcome the problem of negotiating knowledge that we are interested in: how do speakers attempt to transfer that information? What is effective? Do interactants alter their methods?

The lack of mutual knowledge, and the need for negotiation obviously provides interesting material for research into speaker behaviour, but the structure of the corpus provides even more scope for investigation. The design of the quads allows us to track the development of the way in which people tackle the problem of information negotiation. We can compare the differences between the first and second givings of a map: do the participants’ strategies change as they gain experience of the task? Any such changes can tell us more about what underlying principles speakers orient towards.

In terms of a study into communicative success, the Map Task Corpus has one other plus: the route drawn by the Follower. This can be used as the metric of task success that such a study requires. It is independent of the dialogue, and can be judged according to appropriate criteria (this is discussed further in Section 4.6.5). This enables us to rank Follower maps according to their degree of success, and thus compare this ranking with variation in particular aspects of the dialogue.

### 4.4 The Data Set

For this study, we used a subset of the Map Task Corpus: four quads collected in the *eye-contact* condition. This comprises 32 dialogues, involves 16 speakers, and is approximately four hours of speech.

The data is basically one quarter of the corpus: the variable of *eye-contact* vs *no-eye-contact* is excluded, and only half the map sets were used. The design is matched for all other variables: e.g. the ordering of familiar (F) and unfamiliar (UF) pairings. We considered it more important to preserve the famil-
iar/unfamiliar distinction, rather than retaining the cycling of route shapes by using all four sets of maps. It was thought that familiarity was more likely to affect communicative behaviour than route shapes, and duplicating the use of map sets also has the advantage of being able to compare directly the behaviour of different participants on identical maps across quads. For each map, we have four dialogues, which should give sufficient speaker variety to avoid individual speakers affecting the results. In terms of task success, limiting the variety of maps also makes comparison of performance more realistic.

As previously mentioned, the quads in this study are entirely from the eye-contact condition. The main reason for restricting the data to eye-contact only was firstly not to introduce this condition as a variable, and secondly, to avoid the greater number of interruptions and simultaneous speech found in the no-eye condition (Boyle, Anderson, and Newlands, 1994). In a study where the status of knowledge is of paramount importance, we need to be sure that a speaker could have heard and interpreted a particular utterance, otherwise we are unable to make an appropriate coding judgement. Therefore, we chose the condition which would provide the smaller number of these potentially problematic utterances.

4.5 Approach to Data

Now that we have described the experimental design which underlies the data we are using, we will describe the tools used in the analysis of our data set. In order to study communicative behaviour and task success, we needed to develop ways of assessing the dialogue and maps.

- A coding system to highlight the potential areas of interest: The Typology of Move Attributes
- A metric of task success appropriate to the task both in terms of ease of application, and the criteria used

We describe these, and associated tools, below.

4.5.1 Typology Development

Our original aim in this study was to investigate the process of information transfer: what elements of utterances appeared to lead to success or failure. When information appears to be transferred successfully, it is difficult to pinpoint what
exactly about the utterances leads to this conclusion. One can only hypothe-
sise about which aspects are essential, and which extraneous. However, when
miscommunications occur, the problem in the information transfer is more open
to identification and categorisation. Therefore we felt it appropriate to take in-
stances of miscommunication as our starting point for pinpointing the elements
of dialogue which are most important to the information transferral process.

The categories in the Typology of Move Attributes were identified through a two
stage process. As we have already stated, we took points of miscommunication -
places where the speaker was forced to query their partner in order to show their
misunderstanding, or directly elicit some missing information - as our starting
point. These questions are characterised by +QUERY or +OBJECTION attributes,
see Sections 5.6.1 and 5.6.2 respectively). Secondly, we then identified the aspects
of the previous utterance(s) which caused the need for clarificatory behaviour, and
attempted to categorise them into rough groups. This analysis was undertaken
on one quad of dialogues, and produced an initial list of elements which it was
potentially problematic (in terms of the dialogue, not necessarily task success) to
omit. These groupings were refined through application to more dialogues, the
development of which led to more rigorous definitions for each category.

This process produced a list of attributes, which when not used appropriately
could cause a communicative problem. Such negative attributes as not allowing
for differences in knowledge, not answering questions appropriately, not giving
sufficiently precise instructions were shown potentially to lead to clarificatory
behaviour on the part of the other speaker.

4.5.1.1 Limitations of this Method

This method produced a useful typology of categories, but there were certain
aspects of the dialogues which it didn’t address. So far, we had considered solely
the negative, rather than the positive aspects of dialogue management.

All of the negative attributes identified have a positive form by default: they
were identified by the absence of that positive form. However, it was our hypothe-
sis that there are elements of dialogues which may not cause problems by their
absence (or, at least not such obvious ones), but may well speed the informa-
tion transferral process along by their presence, and these would not have been
identified by the method described above.

A number of further attribute types were identified by studying the other types
of utterances that speakers used. These attribute types are associated with the introduction of new information (either offered, or an attempt to elicit) into the dialogue: NEW-SUGGESTION, NEW-QUESTION, RELEVANT-INFO and INFO-INTEG. Section 5.7 in Chapter Five explains in more detail why these categories rarely appear in the negative form, and Chapter Six discusses their importance in relation to task success.

4.5.2 High and Low Risk

We should point out here that although our attribute types were identified via their being absent and by their causing a communicative problem, this does not imply that their absence will always lead to a miscommunication of some sort. Absence is equivalent to a degree of risk: if I talk about a new feature in a route description without checking you know about that feature (an absence), then I risk this causing a knowledge mismatch. However, if we share the feature, then it is likely that no problem will occur: a risk which pays off.

The concept of risk and, more particularly, the Risk-Effort Trade-Off is discussed more fully in Sections 2.5.2, 3.2.4 & 3.2.5.

4.5.3 The Typology - A Summary

The result of the method described above was to produce a taxonomy of attribute types which represent certain aspects of the choices available to dialogue participants. This Typology is described in detail in Chapter Five, but here we will summarise briefly the interests it characterises:

- The paradigmatic choices made at the level of dialogue structure, e.g. what type of INITIATE is it?
- Was this an appropriate choice of move type, e.g. was there a knowledge mismatch which needed to be addressed?
- How much effort was used to formulate the utterance, e.g. is it a high or low level effort utterance attribute?

The coding system was designed to investigate what impact our dialogue choices have on the effectiveness of that dialogue. It particularly focuses on the idea of utterance effort, which we have argued in Chapter Three to be similar to
Cooperation: high effort is equivalent to a high degree of Cooperation. By looking at the effort used in dialogues, we can make an empirical investigation into the structuring principles which we discussed in Chapter Three.

4.6 Using the Typology to Investigate Dialogues

In this Section we will describe how we have used the Typology to study our areas of interest. The results of this work are reported and discussed in Chapter Six. Here, we outline briefly the principles of coding, the need for replicability and a non-linguistic method of comparing the dialogues.

4.6.1 Dialogue Coding

Each utterance in the 32 dialogues of the data set was coded for all the appropriate Typology categories, both positive and negative. This provided a large corpus of data for investigation, and many possible lines of study.

Our research concentrates on comparing whole dialogues with one another, via both a non-linguistic measure (discussed below) and a profile of the dialogue provided by the extensive coding.

4.6.2 Levels of Effort

We have previously stated that our main area of interest is the application of the discussed structuring principles to an empirical study (See Section 4.5.3 in this Chapter, and discussions in Chapter Three and Chapter Six). These principles are mostly defined in terms of either effort, or effect on effort. To this end, the move attribute types are categorised according to the effort they are perceived to require from the speaker. There are four levels in all. This information, in conjunction with the data provided by the basic typology, gives a more detailed representation of how the dialogues work.

Statistical tests were used to make meaningful comparisons between the coding profiles of the dialogues, and in order to support their usage, it was necessary to ensure the coding was replicable (see Section 4.6.4).

---

2See Section 5.2.2 in Chapter Five for a further explanation of these terms, and for a discussion of the coding process.

3The categories are explained and justified in Section 5.3
4.6.3 Dialogue Segmentation Coding

The Typology of Move Attributes is concerned with the utterance level of dialogues. Another form of coding was developed in order to be able to compare how speakers dealt with particular parts of the task. This is called 'dialogue segmentation coding'. The dialogue was subdivided according to which part of the map the speakers were discussing. As our dataset provided us with four conversations about any one map, we could make direct comparisons of the four attempts at particular route segments using our utterance level coding.

Our segmentation is based on the route rather than the dialogue, therefore, there is a segmentation pattern associated with a map. For the map given in Figure 4.1 it would be:

1. start -> extinct volcano
2. extinct volcano - tribal settlement
3. tribal settlement -> rope bridge
4. rope bridge -> <bump> [machete]
5. [machete] -> collapsed shelter
6. collapsed shelter -> turn towards middle [crevasse]
7. [crevasse] -> saxon barn
8. saxon barn -> pelicans
9. pelicans -> white mountain
10. white mountain -> golden beach LHS
11. golden beach LHS -> slate mountain
12. slate mountain -> point where turn downwards [submerged rocks]
13. [submerged rocks] -> secret valley
14. secret valley -> finish

Note on example
This segmentation is based on the Giver's map, as it is the Giver who makes most of the decisions about how the route will be described. The features in square brackets are features which appear only on the Follower's map - hence here they are referred to as changes in the route shape.

The segmentation is task-based. It was developed on the general algorithm that
participants appear to use for breaking up the task into manageable chunks\(^4\). A typical participant will delimit a route section by:

- The occurrence of a landmark (strongest influence)
- The occurrence of a significant bend in the route

The route segmentation is high resolution, it was designed so that segments could be collapsed (e.g. if a speaker chose to describe more than one route part at a time), rather than subdivisions occurring. For example, a speaker could choose to describe segments 9 and 10 in one dialogue part.

This idea of route segments is used in the Typology (e.g. for checking routines, see Section 5.5.1) and for examining the subparts of dialogues in a structured way.

### 4.6.4 Replication Study

This Section briefly describes and reports the results of the study which we carried out to test the reliability of the Typology of Move Attributes. A full description and justification can be found in Appendix A.

#### 4.6.4.1 Introduction

Recently, more interest has been shown in the problem of reliability of subjective coding schemes. Such schemes have been open to the criticism of being ill-defined and not replicable by other coders. We undertook a reliability study in order to demonstrate that our scheme did not suffer from these problems.

#### 4.6.4.2 Subjects

Two naive coders and one expert coder were used in this study. The two naive coders were Linguistics PhD students\(^5\). The expert coder was myself, re-coding

---

\(^4\)An original part of this study involved an attempt to segment the dialogue, according to the type of principles put forward by Grosz and Sidner (1986). This investigation showed the typical points at which segment breaks occurred. Our findings were also supported by Isard and Carletta (1995) who found similar segments using a system based on the dialogues rather than the maps.

\(^5\)They are termed 'naive' because they had not used this system before. One coder was familiar with dialogue game coding (Kowtko, Isard, and Doherty, 1992) and Map Task data, the other was familiar with the Map Task methodology, but not with any dialogue analysis methods.
the data after an 18 month break.

4.6.4.3 Materials

The following data were used:

- Two practice dialogues, with accompanying Giver and Follower maps
- Two test dialogues, with accompanying Giver and Follower maps

The subjects were provided with the following information about the Typology:

- A full guide to the Typology, equivalent to the description given in Chapter Five
- A set of instructions, which included:
  - A guide to the coding process
  - A summary of the Typology (reproduced in Appendix B)
- A tabular summary of the coding system (reproduced in Appendix C)
- A tree diagram of the coding system

The dialogues were divided into moves, and had the move supertype (INITIATE, INSTRUCT, RESPONSE, FOLLOW-UP) marked. The coders were asked to add the appropriate paradigmatic codes (e.g. the type of INITIATE move) and other utterance attribute codes to the dialogue. An example of the coding and layout can be found in Appendix A.

4.6.4.4 Method

The experiment was in four stages:

- Introduction to the system
- Practice coding of two dialogues
- Feedback on practice coding

\footnote{See Chapter Five for an explanation of this terminology.}
• Coding of two test dialogues

The first stage involved describing the system and how it worked to the participants, and answering any immediate queries. They were shown a sample of coded dialogue, and the process of coding was described. This was an informal session. At the end of this session, they were given two practice dialogues to code; this constituted the second stage.

After the participants had completed the practice coding, another meeting was organised. In this session, the participants compared their coding with that of the 'expert', and any discrepancies and misunderstandings were discussed. This, again, was an informal session, and the participants were encouraged to air any uncertainties that they had.

As the fourth stage, the participants were asked to code two test dialogues. They were told to work on their own, and not to collaborate with another coder.

4.6.4.5 Results

The data was analysed using either the kappa statistic or pairwise analysis, depending on the appropriateness for that particular test. These two measures are discussed in Sections A.2.4.1 and A.2.4.2 respectively.

Results from the Kappa Statistic

<table>
<thead>
<tr>
<th>Test</th>
<th>Code tested</th>
<th>Dataset (N)</th>
<th>Data-points</th>
<th>Kappa value</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>INITIATES</td>
<td>153</td>
<td>n/a</td>
<td>0.84</td>
<td>Reliable</td>
</tr>
<tr>
<td>2</td>
<td>RESPONSES</td>
<td>109</td>
<td>n/a</td>
<td>0.85</td>
<td>Reliable</td>
</tr>
<tr>
<td>3</td>
<td>FOLLOW-UPS</td>
<td>71</td>
<td>n/a</td>
<td>0.76</td>
<td>Tentative</td>
</tr>
<tr>
<td>4</td>
<td>+INFO-INTEG</td>
<td>49</td>
<td>7</td>
<td>0.69</td>
<td>Tentative</td>
</tr>
<tr>
<td>5</td>
<td>+KNOWLEDGE-MISMATCH</td>
<td>30</td>
<td>3</td>
<td>0.76</td>
<td>Tentative</td>
</tr>
<tr>
<td>6</td>
<td>-REPLY-FULL</td>
<td>109</td>
<td>19</td>
<td>0.86</td>
<td>Reliable</td>
</tr>
<tr>
<td>7</td>
<td>-ACK-FULL</td>
<td>71</td>
<td>4</td>
<td>0.72</td>
<td>Tentative</td>
</tr>
</tbody>
</table>

Interpretation of the Kappa Statistic

This table summarises the degree of reliability that a particular Kappa value suggests:

<table>
<thead>
<tr>
<th>Value of K</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>K &lt; 0.67</td>
<td>No reliable agreement</td>
</tr>
<tr>
<td>0.67 &lt; K &lt; 0.8</td>
<td>Tentative agreement</td>
</tr>
<tr>
<td>K &gt; 0.8</td>
<td>Reliable agreement</td>
</tr>
</tbody>
</table>
Note: The difference between dataset and datapoint is explained fully in Section A.3, but can be summarised as:

1. *Dataset*: the number of places where the coding could potentially have occurred in the dialogue

2. *Datapoints*: the number of times the named coding *did* occur in the dataset

### Results from Pairwise Analysis

<table>
<thead>
<tr>
<th>Test</th>
<th>Code tested</th>
<th>Datapoints</th>
<th>Agreement</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>+FEATURE-INTRO,</td>
<td>23</td>
<td>85.9%</td>
<td>Good</td>
</tr>
<tr>
<td></td>
<td>-FEATURE-INTRO</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>-RELEVANT-INFO</td>
<td>10</td>
<td>80%</td>
<td>Good</td>
</tr>
<tr>
<td>10</td>
<td>-CHECK</td>
<td>8</td>
<td>83.3%</td>
<td>Good</td>
</tr>
<tr>
<td>11</td>
<td>-REPLY-MIN</td>
<td>5</td>
<td>66.7%</td>
<td>Poor</td>
</tr>
<tr>
<td>12</td>
<td>-ACK-SHORT</td>
<td>7</td>
<td>81%</td>
<td>Good</td>
</tr>
<tr>
<td>13</td>
<td>-FEATURE-LOC</td>
<td>11</td>
<td>81.8%</td>
<td>Good</td>
</tr>
<tr>
<td>14</td>
<td>+FEATURE-UNIQUE</td>
<td>1</td>
<td>100%</td>
<td>?</td>
</tr>
<tr>
<td>15</td>
<td>-QUERY</td>
<td>1</td>
<td>100%</td>
<td>?</td>
</tr>
<tr>
<td>16</td>
<td>-KNOWLEDGE-MISMATCH</td>
<td>1</td>
<td>100%</td>
<td>?</td>
</tr>
</tbody>
</table>

Note: Anything less than 80% agreement is considered to demonstrate poor reliability. This is due to the difficulty in interpreting the results of Pairwise Analysis. See Section A.2.4.2 for a discussion of this.

Although not all of our categories reached agreement values which would guarantee reliability, we would argue that the majority reached an acceptable level, given the type of data involved. Those which didn’t can mainly be explained by the difficulty in achieving reliability on very small datasets.

A fuller discussion of these results can be found in Section A.3.

### 4.6.4.6 Discussion

We would argue that these results, although not always *strictly* reliable, constitute support for the reliability of the Typology of Move Attributes. The degree of reliability found is always close to the acceptability level, even when it does not quite reach that level. This is especially true when one considers the constraints on this study.

We were unable to collect either the amount of data which we needed, or train our coders sufficiently. Ideally, we would have introduced another training stage.
into the study, which would have remediated some errors, and the final test data would have consisted of four dialogues rather than two. This would have been a better test of reliability in terms of both the degree of training and the amount of data.

Given these limitations, the naive coders performed very well to reach these levels of reliability, and with more training we believe all the categories would have been shown to be reliable.

### 4.6.5 Measures of Success

In order to investigate the Map Task dialogues, we needed a non-linguistic method of comparing the efficacy of the conversations. The route drawn by the Follower was the obvious choice, as it is non-linguistic, and it can be viewed as an indicator of task success. However, it is not straightforward to develop an appropriate non-subjective way of rating Follower routes. In the next sections we will outline two different methods of measuring task success (deviation score and Incorrect Entity score), and discuss their suitability for our purposes.

Firstly, however, we will describe briefly a small experiment which compared seven naive raters' subjective judgements about the Followers’ drawn routes with the two methods given below (Davies, Merrison, and Sotillo, to appear). They were asked to subjectively rate maps by quad, and these results were then compared to the ratings produced by several measures of task success (including the two discussed here). The results showed that both measures of success were highly correlated with the naive coders subjective rating but there was no difference in the levels of significance between the two measures (All tests produced significance levels of \( p < 0.001 \))^7. This meant that our decision of which measure to use was based mainly on which one was considered to be best for our purposes. We needed a measure which was entirely based on the map (no reference to the dialogue), and which concentrated on the aspects of route-drawing in which we were interested. Therefore, we wanted a measure which concentrated on how the features were negotiated, rather than one which was based on the absolute accuracy of the map.

---

^7Using the Kendall Coefficient of Concordance \( W \), the following results were found. For deviation score, for quad 1 \( N = 8, W = 0.788, X^2 = 44.15, p \leq 0.001 \); for quad 2 \( N = 8, W = 0.791, X^2 = 44.31, p \leq 0.001 \). For IE score, for quad 1 \( N = 8, W = 0.792, X^2 = 44.35, p \leq 0.001 \); for quad 2 \( N = 8, W = 0.787, X^2 = 44.07, p \leq 0.001 \).
4.6.5.1 Deviation Score

Calculating a deviation score involves overlaying the Follower's map with a copy of the Giver's map, and working out the area between the Follower's route and the Giver's route in cm². The Giver's map has 1cm grid squares marked on it, to simplify this process.

Advantages This is a numerical calculation, not based on any subjective judgement of the quality of the route. It is relatively simple to teach to naive raters, and easy to use.

Disadvantages The concept of the deviation score has several problems. Firstly, it may seem fairly simple to calculate, but estimating portions of grid squares is not straightforward: it is inevitable that the section being calculated will not always contain whole grid squares. The process of counting grid squares is also time-consuming, and open to errors.

Secondly, there are some major disadvantages in terms of the aims of our study. Using deviation scores, Followers can be penalised quite heavily for drawing a slightly inaccurate route which negotiates all the landmarks correctly. Interactants vary in how they interpret the level of accuracy required: very few aim for absolute replication. If one is attempting to gauge effective communication, then it seems inappropriate to measure an aspect that the majority of participants were not even trying to achieve. Features are the pivots of the routes, and it does not seem sensible to use a system where a Follower drawing through the middle of a feature is penalised less than a Follower whose line navigates the feature correctly, but is a little wide.

4.6.5.2 Incorrect Entities Score

Incorrect Entities score (IE score) is based in the number of features which are not negotiated correctly: the higher the score, the more inaccurate the map. Features which were drawn through, drawn past on the wrong side or where the line was very wide of the feature, were deemed to be incorrect. The amount added to the score for one incorrect feature varies according to two planes: the type of feature, and whether the line drawn was considered to be a 'good' or 'bad'

---

8This method is used, and described more fully, in (Anderson, Clark, and Mullin, 1991b; Anderson and Boyle, 1994)
misses. Features can be shared (S), unshared giver (UG), or unshared follower (UF). The IE system is weighted so that more points are incurred if an S feature is incorrectly navigated than if a UF feature is incorrect: it should be easier to negotiate a feature you both know about. Similarly, more points are incurred if a UF feature is incorrect than if a UG feature is incorrect: it is easier to work out where the route goes when you have the missing feature (you have the feature, if not the route) than when trying to draw round a feature which you don’t have, as you have direct knowledge of neither feature nor route. The relative difficulty of these three types of feature has been shown statistically (Davies, Merrison, and Sotillo, to appear).

**Advantages** This measure of success produces a value which is dependent on aspects of task success in which we are interested. It is a system which can be taught, although it does perhaps involve more subjective aspects than deviation scores. It does have the advantage of being fairly quick and straightforward to calculate.

**Disadvantages** The IE score system is based on partially subjective decisions (i.e. what counts as a miss) which does introduce the question of reliability. However, the categories are very well defined, and we found a high degree of inter-coder reliability when we were developing the system. The system is unable to distinguish between an extremely accurate route, and a route which negotiates the features correctly but doesn’t stay as close to the Giver’s route. This is a disadvantage, but we felt that the features were a more important element of the task than the absolute accuracy.

**4.6.5.3 Conclusion**

We chose to use IE score rather than Deviation score because it relies more on how people negotiated the features, which we considered to be pivotal part of the task. The problem of subjectivity was minimised by thorough definitions, and by producing a ratified coding of the data set (each map was coded by three coders, and from this an agreed coding for that map was produced).
4.7 Summary

In this Chapter we have presented the data which we used for the empirical part of this research, and described the tools which we have employed to investigate the use of language in this corpus.

Firstly, this involved describing the structure of the HCRC Map Task Corpus, and demonstrating how the design was appropriate to our interests. The dataset which we used (a subset of the Corpus) was also described, and examples of the maps used can be seen in Figures 4.1 & 4.2.

The Typology of Move Attributes was also introduced, and the process of its development was outlined. This categorisation system is explained in detail in the following Chapter, as is the concept of levels of Effort. As this is a subjective coding scheme, a replication study was undertaken. A brief description of this and its results are included here. A full report can be found in Appendix A.

Finally, we discussed the method used to score the maps drawn by our pairs of interactants. As we are interested in the relationship between dialogues and task success, it is important that an appropriate method of scoring the maps (and thus measuring task success) was found. We described how the Incorrect Entities Score (IE Score) was calculated, and justified its use in terms of its appropriateness to our objectives.
Chapter 5

A Typology of Move Attributes

5.1 Introduction

The main aim of this Chapter is to describe the Typology of Dialogue Attributes which we have developed, illustrate each of the categories, and demonstrate how the coding works in practice.

We start with a description of the motivations behind the Typology, and explain the idea of dialogue coding. We then move on to defining each of the move attributes, describing their linguistic background, the reasoning behind that particular category, and illustrating their usage through examples. This part of the Chapter is organised according to the type of utterance the move attribute applies to: INITIATE, RESPONSE, FOLLOW-UP and attribute types not fixed to particular move types. A brief outline of this structure is given in Section 5.8.

More detail on the coding process is given in Appendix B. This contains further explanations, and some of the documentation which was given to participants in the replication study (Section 4.6.4 and Appendix A).

5.2 A Typology of Move Attributes

In this Section we wish to outline briefly the basis of the Typology of Move Attributes which we have developed. Its roots lie both in the process of analysing a corpus (from a particular point of view), and in some of the basic components of Discourse Analysis (e.g., Sinclair and Coulthard, 1975; Coulthard, 1994; Eggins and Slade, 1997), Conversation Analysis (e.g., Sacks, Schegloff, and Jefferson, 1974; Atkinson and Drew, 1979; Eggins and Slade, 1997) and dialogue games (e.g., Power, 1974; Houghton, 1986; Kowtko, Isard, and Doherty, 1992).
The starting point of this research was an interest in the process of information transferral, especially in relation to the concepts of ‘Cooperation’, ‘Effort’ and ‘Risk’ in dialogue. We wanted to investigate how real speakers behaved in task-oriented dialogues, and how this related to the theoretical discussions in this area.

In order to achieve this, we have used certain aspects of the three methodologies named above. At this point, we would like to explain what concepts have been used from each of the methodologies, and how they have contributed to the Typology which we present in this chapter.

**Birmingham School Discourse Analysis** We use the terms INITIATE, RESPONSE, FOLLOW-UP and the associated concept of the move in largely the same way as the Birmingham School. Neither the levels of acts nor exchanges were adopted. We are not particularly concerned with the internal structure of moves: the move-type and the status of the information in the move were considered to be more important. Similarly, the exchange level is not important to the issues in which we are interested. In addition, we found the IRF structure (or even I (R/I) (R) (F) (F) (F) ) too much of a constraint. For our work, the accurate labelling of the move as INITIATE, RESPONSE or FOLLOW-UP is more important than trying to ‘force’ the utterances into a particular move sequence.

The issue of choice also stems in some respects from this methodology. Birmingham School Discourse Analysis is derived from what is now termed Systemic Functional Grammar (SFG) (e.g., Halliday, 1994). One aspect of this theory is the use of system networks to represent the series of choices which a speaker makes in order to produce a particular syntactic structure. However, SFG does not consider the consequences of these choices, which is what we attempt to do here.

**Conversation Analysis** One of the main differences between Conversation Analysis (CA) and Birmingham School Discourse Analysis is that CA looks for patterns in talk, not for an exhaustive analysis system. Like CA, we are concentrating on those aspects of the dialogue which seem to be relevant to our interests, and do not attempt to account for all aspects of the dialogue. This enables us to focus on those elements, and ignore the rest.\(^1\)

\(^1\)Of course, this approach has its problems, as one could potentially be ignoring an aspect of dialogue which is, in statistical terms, a confounding variable. However, given that this is still a fairly new area of research, we feel that we are justified in narrowing our focus.
Again, CA does introduce the notion of choice. The adjacency pair structure consists of two parts, where a second part is highly predicted, and there are typically several possible second parts. From our point of view this has two important aspects. Firstly, the adjacency pair structure highlights the idea of speaker obligation. First pair parts require an appropriate second pair part. This is the basis of our concept of a speaker failing to undertake an utterance necessary to/predicted by the current state of the dialogue. Secondly, the concept of Preference Organisation introduces the idea of there being more and less appropriate utterances at particular points. Admittedly, in CA the choices are concerned with social face and politeness (e.g. preferring an acceptance rather than a refusal of an invitation) rather than our more information-based concerns, but we believe the principle is generalisable.

**Dialogue Games** The ideas which we have adopted from Dialogue Games are more conceptual than structural. Of the three approaches to dialogue which we have referred to here, Dialogue Games places the most emphasis on the intentions of the speaker, and the goals which they are trying to achieve. The system we describe here also emphasises these aspects. This is because the concepts that we are concerned with - Cooperation, Effort, Risk - operate more at the level of intentions and goals than at the lower, more structural, levels which Conversation Analysis and Birmingham School Discourse Analysis concentrate on.

### 5.2.1 Motivations

Our primary interest in this Typology is the amount of effort that a speaker invests over and above what is strictly necessary to keep the conversation going (see Sections 2.5.2 & 3.2.4 for a discussion of Risk and Effort). This is a major factor in the categorisation of the Typology in a set of effort levels (see Section 5.3 below). In the case of the Map Task, the basic task is for the Instruction Giver to describe the route to the Instruction Follower. This could consist purely of a series of inform-response exchanges (and in places does), yet most speakers choose to employ some of the other options open to them, such as question-asking, checking, providing new information, and so forth. What we wish to investigate is what these speakers gain by using (or not using) these alternate move attributes. Therefore, we do not attempt to classify the most basic structure which could achieve the telling of the route in our Typology: we take for granted that at least this will occur. We are concerned with what extra people do, or don’t do, and
what effect this has on task success. In terms of this Typology, these differences are mainly defined in respect of the interpersonal level of dialogue: whether the speaker is taking note of their partner’s goals, and to what extent this is the case.

As we have set out in Chapter Three (Section 3.3), Effort is considered synonymous with the concept of Cooperation: high effort utterances are highly cooperative, and vice versa. Effort is considered to entail taking into consideration such things as thinking about the knowledge state and/or the goals of the other participant when planning an utterance - the more aspects you consider, the more effort you invest. Both these terms will be referred to throughout this Chapter.

Secondly, our categories are defined in terms of speaker intent rather than any more functional or syntactic basis. We are interested in the type of information given, rather than the fact that the utterance is an INITIATE move, and similarly, we distinguish between types of FOLLOW-UP moves or types of questions (for example) which neither Coulthard (1994) nor Kowtko, Isard, and Doherty (1992) would differentiate. Therefore, we are interested in the effect of the variation within move types on both the dialogues themselves and task success. This variation in moves is considered in terms of a variation in speaker effort.

Why do we focus our attention on this level of dialogue? Research into dialogue has long illustrated that speaker behaviour tends to follow identifiable ‘rules’ for patterns of move sequences: question-answer pairs, Initiate-Response-Feedback and so forth. Deliberately violating these conventions is seen as marked behaviour: it occurs infrequently, and it may be perceived as being impolite, unhelpful, or at least odd in some respect. Therefore, there is not a great deal of variation in this plane. Certain types of move are expected at certain points, but the content of these moves is not so highly predicted. A follow-up move, for instance, could be a simple “Yeah” or “Aha”, or it could provide some more information as additional feedback. Therefore, the move level produces variation without being associated with non-typical dialogue.

Thirdly, both Discourse Analysis and Conversation Analysis are concerned with what happens in a dialogue, rather than what might have happened: any labelling is based on what moves or structures speakers used. Our system is equally concerned with what speakers don’t do. Dialogue is essentially a system of choices and we may not always make the right ones. Therefore, our coding system considers whether the dialogue move chosen was the most appropriate at that point. If another move would have been a better choice, then such errors are coded too. The errors coded are dialogue errors, not map errors: judgements are always
based on the dialogue, not the map. Therefore, the Typology is essentially an evaluative system, rather than just a descriptive one. The nature of this evaluation is based on the concept of speaker effort. How this is achieved is described in Section 5.3.

Finally, we are concerned primarily with one aspect of dialogue in a limited setting: the transfer of information in a task-oriented domain. Many of our categories would be generalisable to other types of conversation, but this is an additional feature rather than our intent. Sinclair and Coulthard were attempting to produce a more general analysis of dialogue, although it is arguable whether such an intention was reasonable given their (also) limited domain of classroom talk. Kowtco et al. ’s work is more general than ours in the sense that it isn’t focused on just the informational aspect of dialogue, although it is intended to function in the same domain as our current work.

5.2.2 Coding the Dialogues

Much mention will be made in this, and following Chapters, of coding. We should explain what we mean by this term. In the transcripts of our dialogue data, each utterance was assessed in terms of the move attributes defined in this Chapter. Obviously, particular utterances have certain sets of move attributes which are more applicable: RESPONSE moves will employ a different range of move attributes from INITIATE moves, for example. The move attributes are grouped according to which move-types they apply to, and a brief list is given in Section 5.8.

Positive Coding
Positive coding is based on the types of move attributes which speakers have used in an utterance. If an utterance meets the definition of a particular attribute, then it will be coded positively for that attribute. Each dialogue turn will typically be coded for one or more attributes, depending on the move-type involved.

Negative Coding
Negative coding is based on what speakers don’t do. Interactants make a series of choices when they speak. However, we would suggest that they do not always make the best choice, and that an analyst can interpret the utterance in the dialogue context. When speakers don’t make the best choice, they are considered to have made a dialogue error, or failure, and the utterance will be negatively coded for the attribute which the analyst believes would have been most appropriate. Such negative codings are not as common as positive codings, and will only apply
to a small subset of turns in a typical conversation.

5.3 Effort in the Typology of Move Attributes

The definitions of attributes in the Typology rely on the amount of Effort which a speaker is considered to have invested: the main difference between a +REPLY-YN and +REPLY-FULL coding for a RESPONSE move is the amount of useful information (and therefore effort in preparation) which is included in that move. We would argue that there is an implicit hierarchical structure for INITIATE, RESPONSE and FOLLOW-UP moves which is based on the concept of Effort.

Given our interest in Effort, and the way in which the Typology is structured to investigate this, we needed to represent the differences in Effort in some measurable way. We want to be able to compile a rating for each dialogue based on the amount and type of attributes employed, or not employed. In other words, we want a profile of the Effort pairs of speakers invest. Tallies of the number of positive and negative codings do not effectively achieve this. There would be no difference drawn between, say, the use of the relatively high effort +ACK-FULL, and the relatively low effort +ACK-SHORT - each would add another point to the overall tally. Our argument is that the points ‘awarded’ for the uses of different attributes should be differentiated like tries and penalties in rugby: the former is worth more points than the latter, because it is harder to achieve: effort again.

The obvious solution to this problem is to weight each of the attributes appropriately. However, this is not as straightforward as it might at first appear. This is because we must identify the types and extent of effort which a speaker invests in order to produce a particular utterance. Effort, in itself, is not easy to define.

5.3.1 A Definition of Effort

Earlier in this thesis we have equated Effort with ‘work’; here, we intend to set out what we consider to be the ‘work’ of utterance creation.

Effort could be considered in several different ways. For example, one could concentrate on the planning needed in order to produce a particular utterance, or in order to break a task down into manageable subtasks. Or, the Effort involved in maintaining an accurate belief model, especially as the Map Task was designed to create the potential for conflicting beliefs. There is also the issue of interpersonal relations. Much Effort can be spent maintaining social relationships, for example,
Goffman’s notion of ‘face-work’ (Goffman, 1967).

In this work, there is not a focus on one type of effort. Instead, we draw on all these aspects to a lesser or greater extent. This approach is largely due to the way in which the Typology was developed. The categories were not identified by the type or amount of Effort they used, but rather by their potential to aid or cause problems in the dialogue. Therefore, when Effort is considered in the context of this Typology, several strands can be identified.

We define Effort as follows:

Effort
The work invested by the Speaker in the following aspects of utterance planning:

1. The social needs of the dialogue
2. The responsibility of supplying the needs of your partner
3. The responsibility of maintaining correct mutual beliefs
4. The responsibility of initiating new subtasks

These aspects are listed from least to most effort. We will now consider each of these levels of effort in turn, and justify their position in the Effort hierarchy.

1. The social needs of the dialogue

In order for a dialogue to continue, we must at least provide some sort of minimal Responses and Follow-ups to our partner’s utterances. Without this social awareness of dialogue needs, the dialogue will cease as the other interactant becomes frustrated by the lack of response.

These brief utterances are classed as low effort because they do not require much planning, nor much consideration of the joint beliefs of the speakers, nor what the contribution of the utterance is to the overall dialogue and joint task.

In Shadbolt (1984) terms, this would entail a high risk approach to all the parameters: Difference, Specification, Ontology, Focus, Decentering, Feedback.

See Section 4.5.1 for an explanation of this.

See Section 2.5.2 for an explanation of the various parameters.
2. 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 consider more your partner’s intention in formulating their utterance. In other words, this requires the ‘work’ of evaluating the utterance in the discourse context, and planning a response which attempts to meet the intention of the original speaker.

This aspect of Effort is considered to involve more work because there is some concept of the contribution of the response to the task, and the goals of the original speaker.

In terms of Shadbolt (1984), this would mean a low risk approach to Feedback, but a higher risk approach to all the other parameters.

3. 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 accurate.

We make the assumption that going against the predicted move in an exchange structure (e.g. Initiate, Response, Follow-Up) 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. Also, the speaker should also consider the need for redressive action for such a potentially face-threatening act (Brown and Levinson, 1987).

In terms of the speaker producing utterances which do not make risky assumptions about mutual knowledge, the extra work involved is the reasoning about beliefs, and the potential longer utterances needed for the identification of ambiguous features, for example.

In terms of Shadbolt (1984), this aspect of Effort focuses on the parameters of Difference, Specification, Decentering and Focus. These would all have a fairly low risk approach.

4. The responsibility of initiating new subtasks

The previous levels of Effort which we have discussed all consider the actions of a speaker within the context of a particular subgoal. It assumes that one
speaker has set up the current subgoal, and its relationship to the overall task. At some point, however, an interactant has to take the responsibility of moving the dialogue on to the next subgoal.

This we perceive to be high effort because it requires reasoning about the approach to the task, and then planning and producing an appropriate utterance. This is in comparison to reacting to an utterance (and therefore precise discourse context) provided by the other speaker.

Shadbolt’s (1984) parameters do not seem to relate particularly well to this level. They are more concerned with the planning of individual utterances than the higher level planning of goals and subgoals. However, we consider these levels of Effort which we have considered to be part of an inheritance hierarchy. In other words, each level of Effort should ‘inherit’ the concerns of the lower effort levels. Therefore, where new subtasks are initiated, we would expect such an utterance to show an awareness of the appropriate Shadbolt (1984) parameters.

5.3.2 Effort in the Typology

In the previous Section we have discussed the notion of Effort, and set out what aspects of this we believe are important and require work on the part of the speaker. Here, we set out how this concept of levels of Effort is translated into the structure of the Typology.

We identified these four levels of effort above:

1. The social needs of the dialogue
2. The responsibility of supplying the needs of your partner
3. The responsibility of maintaining correct mutual beliefs
4. The responsibility of initiating new subtasks

These broad levels are used in the Typology, and are described thus:

1. [Minimum Effort]
   Which can also be described as the absolute minimum required to keep the conversation going
2. Moderate Effort
   Which can also be described as moderate behaviour, prompted by other speaker
3. Medium Effort
   Which can also be described as medium behaviour, more initiated by current speaker, but still prompted by other speaker.

4. High Effort
   Which can also be described as high effort, initiated by current speaker.

By associating each category in the Typology with one of these levels, we can then weight the contribution of each utterance in terms of the effort invested.

The relationship between each of the category types and its assigned Effort level is discussed in the introductory part of the Section appropriate to that Effort level (Minimum, Moderate, Medium, High).

Where the assignment of categories is not straightforward (e.g. +CHECK and +FEATURE-INTRO which are assigned lower levels of Effort than might be expected), we argue why we believe this to be appropriate (Section 5.5).

5.3.3 Using Effort Groupings

The effort groupings outlined above, and described in detail below, are intended to differentiate between the different types of attributes and their perceived effort ratings. But how are these employed empirically? These groupings are translated into a simple metric: instead of making a tally of positive codings, each instance of a coding is weighted according to its effort grouping. Thus for positive coding, a level 1 attribute has a weighting of one, level 2 has a weighting of two, and so forth. For negative coding, the weighting is reversed. Thus, failing to do the absolutely necessary (level 1) earns a weighting of 4, whereas failing to use high effort move attributes (level 4) where they might be appropriate incurs a much lesser penalty: a weighting of 1. The table below shows the weightings for positive and negative codings.

<table>
<thead>
<tr>
<th>Effort Level (Least first)</th>
<th>Positive Weighting</th>
<th>Negative Weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Level 2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Level 3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Level 4</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

The weightings that we have chosen are arbitrary in their value, but (obviously) not in their order. Neither would we claim that they are a true interval scale.
This is a first attempt at evaluating what speakers do in dialogue. What we are trying to model could be seen in terms of Traum's (1994) discourse obligations (Section 2.5.3). Traum does not use the concept of discourse obligation in a gradeable way, but this is in effect what we are attempting to do here. Our argument is that speakers are more obliged to engage in those move attributes which are in the lower levels than those that are in the higher levels. As an interactant in a speech situation, I am more obliged to make some sort of answer to a question (e.g. REPLY-MIN, REPLY-YN: level 1) than I am to make a fuller answer to that same question (e.g. REPLY-FULL: level 2). I am under even less obligation to initiate an exchange (e.g. QUERY: level 3), and even less to use a move type which is not immediately prompted by the discourse context (e.g. RELEVANT-INFO: level 4).

In terms of the weighting system outlined above, this means that the greater the obligation for you to undertake a dialogue utterance, the lower the positive weighting, and the higher the negative weighting. This reflects the impact on the dialogue. Carrying out a low level utterance will keep the conversation going (low positive impact), failing to meet your discourse obligation will potentially affect the smoothness of the conversation, and its final outcome (high negative impact). The reverse relationship holds for high level utterances: a high positive impact (high positive weighting), a low negative impact (low negative weighting).

Using this system enables us to provide a positive and negative score (sum of codings) for a dialogue, which attempt to account for the effort invested in some principled way. It is on these ratings that much of our empirical results are based.

5.3.4 A Note on Examples

Our account of each of the move attributes is aided by use of examples from the Map Task corpus. This Section gives a brief outline of the transcription conventions which we have adopted.

- Examples are presented in a different font

  in order to distinguish them from the main text.

- The label $Eaq\text{cl}$ (for example) at the top of the extract is simply the dialogue from which the extract is taken.
• #TA refers to the Giver’s utterance, #TB to the Follower’s turn. The numbers appended to these codes refer to the turn number in the dialogue.

• When referring to utterances in the dialogue, the Giver is always referred to as female, and the Follower as male. This is for ease of disambiguation.

• #G is the coding line for the Giver, #F the coding line for the Follower. They are given separate coding lines in order to make the process of data extraction easier.

• All coding apart from the one being discussed at that point has been deleted from the dialogue extracts. This is both to make the text more readable, and to clarify the point of each extract.

• Angled brackets (<,>) mark the beginning and end of an interrupted utterance. A forward slash (/) marks the point of the interruption (to the nearest word boundary.)

• ‘...’ at the beginning or end of a turn means that part of the turn has been omitted from this example. ‘...’ elsewhere means either a pause, or a reformulation (restart).

5.4 1 - Minimum Effort

• PARTNER-BELIEF

• REPLY-MIN

• REPLY-YN

• ACK-SHORT

These categories are classed as minimum effort because they are the least a speaker can do in order to maintain a conversation: a co-participant will quickly give up if not even short responses and follow-ups are forthcoming. These utterances fulfil, in Traum’s (1994) terms, the discourse obligations of an interactant (see Section 2.5.3). Such a conversation, however, would rely on the co-participant initiating all the conversational games, and effectively investing a disproportionate amount of effort in order to achieve the task goals.

In terms of the Effort levels discussed earlier, the three move-types (REPLY-MIN, REPLY-YN and ACK-SHORT) are considered to be Minimum Effort because they
do not require much effort in planning or production, nor much consideration of joint beliefs, nor do they contribute to sub-task planning. Therefore, in most possible definitions of Effort, they would be considered to need a low input of effort in order to be produced. However, they are essential to the smooth running of a dialogue\(^4\): they may not be sufficient in themselves, but they are the bedrock on which any dialogue is built.

However, the inclusion of the category PARTNER BELIEF here is for slightly different reasons. In itself, it could be argued that it is easier (less Effort) to assume that your partner is telling the truth. However, we do not have a category which states that it is more Effort not to do so! PARTNER BELIEF is included here because, like the other categories described here, it is necessary to the social requirements of the dialogue. If the co-participants do not believe that their partner is acting in good faith, then the conversation will rapidly break down. A speaker cannot act upon information given unless they believe that this information is a true representation of their partner’s beliefs. Therefore, the belief in your partner is an important basis to dialogue.

5.4.1 PARTNER-BELIEF - Belief in partner’s contribution

In a situation where the transferral of unshared information is the goal, it is vital that each speaker believes that the information being given to them is bona fide. Without that basis, it would be impossible to negotiate the mismatch of information in the Map Task (or most other dialogue). If we accept that this assumption holds as a basic tenet, then to question that assumption may be seen to be a violation of the code which operates between two (or more) speakers.

In general, speakers tend to be reluctant to accept differences in the maps, even though they are warned that such variations may exist. Brown (1995) found a similar reluctance in the participants in her Map Task data. In our data, the first unshared feature which a dialogue pair encounter often caused some difficulty because of this unwillingness to believe that there were differences. Brown (1995:66) suggests that participants who behave in this way must be “... attributing quite irrational and unco-operative behaviour to their interlocutors.”, because it involves believing that your partner has introduced false information into the dialogue. We would argue that such a reaction is more likely to be due to the speaker not wanting to reassess the dialogue situation: if they can’t make assumptions about the sharedness of features, then getting through this task is

\(^4\)Except REPLY-MIN, which is insufficient in itself

117
going to involve much more effort. In Shadbolt’s (1984) terms, they would have to choose a lower risk (and thus higher effort) setting on their difference parameter\(^5\). As we argue in Section 3.2.3.3, such a change means going against the Principle of Least Individual Effort - not a change a participant makes lightly.

It should be pointed out, however, that the type of behaviour we are referring to here is the repeated evidence of disbelief of a particular contribution, not a one-off query or check, even if an element of ‘unwillingness to believe’ may be involved. Queries are often seeking confirmation that the other speaker’s contribution has been understood properly. Or, at most a suggestion that there is some relevant (possibly new) information which hasn’t been taken into account by the original speaker. Its intention is not to question the quality of the information given, but rather to check that it has been interpreted correctly. As we have suggested above, if speakers are confronted with a difference they weren’t expecting, then they are going to make sure that this difference definitely exists before changing their overall strategy. This isn’t questioning the quality of their partner’s contribution, it is a pragmatic check to avoid wasted effort. However, when the notion of quality is questioned, through repeated use of a particular query, then there is an element of the contribution being seen not to be queried but to be rejected.

In this case, we are not coding the direct intention of the utterance (which is our purpose in most of the other categories) but rather the interpersonal level of the communication. Such a violation of this basis for information exchange could have implications for the remainder of the dialogue, for example whether the hearer at this point decides not to introduce other relevant information, as its quality has been questioned. In terms of a dialogue context where the transferral of information is all-important, this would be quite a significant risk.

Eaq4c2

#TA 67
Then come vertically down until you come to gold mine.

#TB 68
I don’t have a gold mine.

\(^5\)The difference parameter basically refers to the amount of difference the speaker assumes between their state of knowledge, and their participant’s state of knowledge. Assuming more difference means investing more effort in grounding new information.
You don't have a gold mine?

No.

You must have a gold mine somewhere. Right.

I don't, I didn't have that

Is ... is it below banana tree?

No.

No.

Rock fall, I've got.

Rock fall. Surely between rock fall and banana tree you've got a gold mine, no?

No. (laugh)

(laugh) Is it ... Is it anywhere at all near?
In the above example, such repeated questioning can be seen (turns TA71, TA73, TA77, TA79). The Follower is being asked again and again whether they have the *gold mine*, and it takes another ten turns before the Giver appears to accept the Follower's version. It is interesting to note here that the speaker whose contribution was questioned ceased introducing new features after this point, which does perhaps suggest that their behaviour was affected by this exchange.

In terms of coding, there is a binary choice in operation on this move attribute. Either such a belief is apparent (the assumption), or it is obviously not. Only the latter of these is coded, and that would obviously be negative. In practice, positive coding does not exist for this move attribute. This may not seem entirely logical, but the reasoning is simple: we know (or can easily work out) when speakers are adhering to this principle of partner belief as this is the default - we only need to label the deviations from that widely accepted norm.

It should also be noted that in the example above, the first query (TA69) about the differing beliefs is not negatively coded for this move attribute. It is only the *repeated* queries that are coded thus.

In principle, this move attribute could apply to any move type, but because of its nature, it will probably only occur in questions, particularly queries. We have few examples of this, and those we have all occur in queries, as above.

### 5.4.2 REPLY-MIN - Provide a Response

This move attribute describes *minimal* responses to questions, where the utterance fulfils the basic function of a 'response' move, but does not answer the question in the fullest sense. This can either be where the response is deemed to be too brief, or where the utterance simply does not answer the question being asked. This is low effort because of the lack of effort involved in producing an appropriate response, and because the move risks miscommunication and task errors.

Eaq3c2

#TA 33
Can you move down, eh, three inches.
Could you move to your left, ah, approximately six inches.

#TB 34
Am I allowed to say if I’m going to go into one of these obstacles?
It’s got right in the middle of the water. Six inches.

#TA 35
Mmhmm.
#G +REPLY-MIN -REPLY-FULL
Right, okay.
[Moves on to next dialogue segment]

In this extract, the Giver’s reply in TA35 does not address either of the queries which the Follower has made: neither the metacomment about the task, nor the current problem. Such a brief answer may be due to the need to provide an appropriate utterance within certain time constraints (Carletta, Caley, and Isard, 1995): in discourse terms, a minimal response is less marked than no response at all. Not replying to a comment by your partner is breaking one of the basic rules of the game, making a minimal reply is at least grounds for the continuation of the dialogue. So the coding of TA35 reflects this idea of it being a ‘stop-gap’ response, as positive coding for this move attribute is better than an equivalent negative coding had no response been made.

The risk taken in this case led to the incorrect negotiation of a landmark.

Note that as the REPLY-MIN coding labels the utterance as ‘insufficient’ in some way, the utterance should also be negatively coded for the standard RESPONSE coding: REPLY-FULL. (See Section 5.5.2 for a description of this coding.)

Eaq3c3

#TA 56
And go straight left to ... almost to the edge of the page.

#TB 57
Do you have a waterfall?

#TA 58
Yeah.
#G +REPLY-YN
Underneath the waterfall?

I want you to go up ... up past ... on the left-hand side of the waterfall, if you can

< But you ... You want me to go ... /

Right, I want you to go ...

I want you to go underneath it and then round the left-hand side of it from where you are now.

Okay, ...

In turn TA60 there is a different instance of a +REPLY-MIN coding. This is where the question of goals comes in. The Giver has made an answer which she feels answers the Follower’s question (we presume), but she has misinterpreted the aim of the question: she hasn’t paid sufficient attention to her partner’s goals. She has invested too little effort, and made a risky interpretation. Therefore, her answer is coded as +REPLY-MIN rather than +REPLY-FULL (see Section 5.5.2) This approach may seem problematic, as we have previously stated that it is speaker intention that we are coding rather than the dialogue result, and misinterpretation does not necessarily imply a ‘bad’ intention. However, we should
clarify what we mean by ‘intention’. If a Follower makes a suggestion about where the route might go next, but it is inaccurate, then it should still be coded as +NEW-SUGGESTION (Section 5.7.1.1) because it was intended as a suggestion and there is no way that the Follower could have known that it would be incorrect. If you like, it was a statement made in good faith. However, in the example in the extract above, the information about what sort of answer was required to the question was available to the Giver. Therefore it is a different type of error. If the necessary information is not available to the speaker when they make such an error, then their intention is taken to be good. If it is felt that the information was available, but for some reason they have failed to act upon it, then their utterance should be assessed at face value. In this case, this means a coding of +REPLY-MIN -REPLY-FULL.

The dual coding here represents the low effort (+REPLY-MIN) and the high risk (-REPLY-FULL).

The example below demonstrates under what circumstances a negative coding for REPLY-MIN would be used.

Eaq3c8

#TA 3

...  
Have you got a haystack, /

#TB 4

Yeah.

#TA 5

on your map? Right, just move straight down from there, then. >

#TB 6

Past the blacksmith?

#TA 7

Past the haystack to, eh ...

#G -REP-MIN

Have you got a carpenter's house?

#TB 8

Yeah, right, I know where you are.

(4)
In TA7, the Giver ignores the Follower’s question, and continues with the route instructions. In this case, the Follower did manage to negotiate the blacksmith without further help from the Giver, but this must be considered risky behaviour on the part of both speakers. Ignoring the Follower’s contribution was risky for the Giver because of the need of both the information about the blacksmith, and because of the effect this might have on the Follower’s decisions on proffering information in the rest of the dialogue: indeed, the Follower did not introduce any more features into the dialogue. For the Follower, the decision not to try the query again was risky because of the immediate affect on task success.

Because the Giver makes no attempt to answer the Follower’s question, this omission is coded as -reply-min: no response where one was required.

Anderson, Clark, and Mullin (1994) found that such omissions had an effect on task success: the more failures there were to answer questions in a dialogue, the worse the task result tended to be.

5.4.3 REPLY-YN - Answer Queries

This coding is also aimed at short responses to questions, but in this case, it is for those utterances which are appropriately brief. A typical example would be the response to a Y/N question, like the one below:

Eaq4c1

#TA 27
Is the Saxon barn on your map?

#TB 28
Yes.

#F +REPLY-YN

This type of utterance is not coded as a +REPLY-FULL because it is not supplying any extra information other than that directly solicited. It is low effort because it requires low input on the part of the speaker. It is not particularly high risk, as it is not insufficient in any respect, although including more information may have made it lower risk. This fuller type of utterance (a Y/N response plus extra information) would be coded as a +REPLY-FULL.
5.4.4 ACK-SHORT - Minimal Follow-up

This category accounts for many of the minimal responses made by the interactants: those utterances whose basic intent is to say “I heard you”, and contain no further information. This is in contrast to follow-up-type moves which are coded as +ACK-FULL, these must contain some comment on information offered in addition to (or instead of) the minimal response.

As for responses coded as REPLY-YN, this category is considered to be low effort because it requires low input on the part of the speaker.

Such a move often consists of a single word, such as “uh huh”, “aha”, or “yeah”. This is illustrated in the typical example in TA3 in the extract below.

Eaq3c2

#TA 1
... Have you got a starting point at the top left of your map corner?

#TB 2
Yeah.

#TA 3
Right.
#G +ACK-SHORT

(6)

In the following example, the use of negative coding for this dialogue attribute is demonstrated.

Eaq4c4

#TA 59
Then you go up vertically, til just above the alpine garden. Then you mak ... go right to the end of the alpine garden. Then vertically again to the top of the youth hostel.

#TB 60
There is no youth hostel.
Then about three inches.

Where was the youth hostel?

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.

IN TA59, the Giver makes the assumption that the youth hostel is shared, and introduces it within a route description. It isn’t shared, and the Follower is forced to object to this assumption. The Giver replans the route, and offers the Follower an alternative description in TA61. However, because the Giver has not acknowledged the Follower’s contribution, the Follower does not interpret TA61 as intended, and assumes that this instruction refers to the route beyond the youth hostel. Therefore, the importance of the acknowledgement in this instance is very clear, and this omission is coded as -ACK-SHORT: no acknowledgement where one was required.

This behaviour would be considered to be risky, as there is no guarantee the Follower would pursue the matter in order to clarify a possible miscommunication. The Giver also shows insufficient regard for the Follower’s wants at this point.

5.5 2 - Moderate Effort

- CHECK-UNDERSTANDING, CHECK-COMPLETION, CHECK-ALIGN
- REPLY-FULL
- ACK-FULL
- ACK-REPEAT
- FEATURE-INTRO
These categories require more effort, although they are either reactions to a participant's initiation, or are domain-specific attributes.

Fuller responses and follow-ups (REPLY-FULL, ACK-FULL, ACK-REPEAT) need more thought than briefer ones. They should be considering their partner's intention in formulating the previous move. Therefore, more planning effort should be involved. They are not considered to be higher Effort because they are reactions to preceding moves: they are not involved in the higher-level planning of tasks and sub-tasks.

Checks are included here as although they are technically questions, they are not information-seeking in the sense that a true question is. The initiator of a check assumes they know the Response which will be made: there is an assumption of mutual belief at some level. In addition, checks are entirely prompted by what has preceded them. Therefore, checking routines neither take real responsibility for the maintenance of mutual belief, nor are they part of higher-level planning. Therefore, we consider them to be lower effort categories.

FEATURE-INTRO is an attribute specific to this particular task, and is picked up quickly by most participants as a useful tool - almost equivalent to a checking move.

FEATURE-INTRO is included here because it is specific to this particular task. It is, perhaps, a more high effort category than is suggested by this grouping, because it requires thought about the co-participant's state of knowledge. However, this attribute is highlighted by the nature of the task, and is therefore a more obvious basic choice - almost equivalent to a checking move. We would suggest that when such utterance-forms are used frequently, they require less absolute Effort to plan and produce than might otherwise be the case. If you have been planning and producing a number of such utterances, then producing one more should be less effortful. One could liken this to the process of learning to drive. In the initial stages, changing gear whilst simultaneously steering or braking is very difficult, although after a while it becomes easy. Yet the same physical actions are required. On a less physical level, consider the process of performing a statistical test. The first few times the calculation is performed, it will probably seem quite tricky. After those first few negotiations, the same process will again seem to involve less Effort.

Morgan (1975) argues a similar case in respect to the difference between highly conventionalised conversational implicatures, and 'one-off' implicatures. He argues that it is highly unlikely that native speakers of English go through the whole
process of making an implicature, and making an inference for such conventional phrases as using the modal can in the request form. For example:

Can you tell me the time?

It is suggested that we simply associate such highly-conventionalised uses with the literal intended meaning, and do not even consider the non-literal meaning of ‘ability to perform the action’. He calls this process ‘short-circuiting’ the implicature. It is equivalent to hard-wiring an electrical circuit: only one choice is possible, so less ‘work’ is expended. It should therefore follow that interpreting a highly-conventionalised implicature should involve less Effort than interpreting a less-conventionalised one.

We argue that in this context +FEATURE-INTRO is, in effect, a conventionalised form because it is so routinely employed. Therefore, if we generalise Morgan’s argument, then it should be considered to involve less Effort in planning and production than might otherwise be the case. According, all other attributes which relate to knowledge about features (FEATURE-LOC, FEATURE-UNIQUE) are given a higher Effort grouping because they are not so routinely employed, and are therefore not conventionalised.

5.5.1 Checking Routines

The group of move attributes described here are those checking routines, often initiated by the Information Giver, which perform the function of ascertaining the successful completion of a particular sub-goal. In the case of the map dialogues, these would be the goal of mutual knowledge in respect of information given, completion of a sub-task, or agreement on current position. They are not requests for extra information, but rather a request for confirmation of what the initiator believes to be true. Their function could be summarised as “checking even when there is no apparent problem”. They are therefore low risk attributes.

These move attributes are typified by brief utterances, such as an exchange of “Ok”, or “Ok” followed by a y/n response. Often, one of these brief exchanges performs more than one of these functions.

Positive coding would be used here when one (or more) of the following move attributes is in evidence. Multiple coding is quite acceptable, +CHECK-UNDERSTANDING and +CHECK-COMPLETION co-occur quite frequently, though this should be either multiple positive coding or multiple negative coding, not a mixture of the
two. Negative coding would only be employed where there was an obvious omission, e.g. at the end of a route section, or after a potentially problematic utterance - a suggestion of how to circumvent an unshared feature, perhaps.

5.5.1.1 CHECK-UNDERSTANDING - Information Understanding

This asks for confirmation that the information proffered has been understood.

#TA 77
I want you to go west, making sure that you remain above the gold mine.
Understand?
#G +CHECK-UNDERSTANDING
So you go due west above the gold mine.

#TB 78
Okay

(8)

This is probably the most frequent type of check to be found in the corpus, and it often consists of an "Okay?". Example (8) is more explicit than most.

5.5.1.2 CHECK-COMPLETION- Subtask Completion

This type of check is to ascertain whether the current subtask (usually drawing a small section of route) has been successfully completed.

Eaq3c8

#TA 35
... Could you move to the bottom, just the corner of seven beeches, at the bottom left-hand corner of seven beeches.

7However, when the score for a dialogue is calculated, a given move is only allowed 1 checking attribute. Therefore, even if a move is coded for all three types of CHECK, the addition to the total will only be +2 - equivalent to the coding of one such attribute. This applies to both positive and negative coding. Otherwise the multiple positive or negative coding which could occur could skew the overall dialogue totals.
Note that the Follower has already confirmed understanding of the instruction before the Giver initiates her checking move. The check gives the Follower the opportunity to state the need for more time to complete the subtask. Without such a break, the Follower would have to interrupt the Giver's next utterance in order to gain that time. A failure to interrupt when the time is needed would mean that the Follower would be attempting to process new information whilst completing the previous subtask. This would probably lead to difficulties with one or other of those tasks, as neither are simple. Therefore, the Giver's check for completion before moving on allows the Follower's needs to be met without a potentially awkward recourse to interruption.

5.5.1.3 CHECK-ALIGN - Alignment

Alignment in this context means checking that you have agreement on your current position. Such checking would typically occur at the end of a route segment, as with the other varieties of checking routines (CHECK-UNDERSTANDING, CHECK-COMPLETION).

This move attribute is initiated by the Instruction Giver (in most cases), and is a posited current position, not a statement asserting a particular position. Statements of position are usually made by Instruction Followers, and are classed as 'introduction of appropriate new information' (see Section 5.7.1.3 below). The reason for this difference is the contrasting intentions behind them.

Both types of utterance are concerned with the Follower's current position. So, for the Giver, it is checking agreement on that position, whereas for the Follower it is a statement of (potentially) new and useful information. Givers do not seem
to give statements about their position, presumably because the focus of the task is the Follower's progress, and when Followers ask about the Giver's current position, they are not checking that they agree on the current position, but rather requesting new information (this would be coded as +NEW-QUESTION). These two types of utterance also differ in how effortful they are considered to be for the speaker. The check move is considered to be less effortful because it is not dealing with the introduction of new information, and predicts the reaction it will receive (agreement).

Eaq3c1

#TA 53
You should be on top of the Indian country, correct?
#G +CHECK-ALIGN

#TB 54
Um, and just south of the ravine.

#TA 55
That's right.

(10)

In this context, this move attribute is very important in ensuring the alignment of the two interactants in respect of the map. Explicit alignment checks enable speakers to discover misunderstandings, and then work towards rectifying them. In the following example, the Giver fails to check for such alignment.

Eaq3c4

#TA 70
Right.
Go right up to the top of the page.

#TB 71
Right to the top?
In this case, the Giver does not check that both participants mean the same thing by the phrase ‘the top of the page’. The Giver (presumably) means ‘level with the bottom of the pyramid’, which is in fact about three to four inches below the top of the page. The Follower, however, draws his line right to the top of the piece of paper (thus -CHECK-UNDERSTANDING). The Giver’s following instructions were then interpreted in relation to the Follower’s incorrect position, and errors were made. This problem could have been discovered had the Giver checked that the Follower was where she thought he was (thus -CHECK-ALIGN).

Not checking is a reasonably low risk failure - there is a reasonable chance that the Giver’s instructions have been understood correctly. In this case, the risk did not pay off, and led to two features being negotiated incorrectly.

5.5.2 REPLY-FULL - Be Appropriately Informative

An utterance coded in this fashion is deemed to have met the functions of both the minimal required response and the provision of information appropriate to the situation. It is probably the most frequent coding used in our corpus for the replies to non-Y/N questions, and is the default for that situation. In extract(3), turn TA64 (Section 5.4.2 above), the Giver provides the directly solicited information, plus some more information which she thinks is relevant to that subpart of the route. For that extra information, the utterance is coded as +REPLY-FULL. As pointed out in the previous Section, this coding can be used for replies to Y/N questions, but this coding is only used when the Y/N response is accompanied by some extra information. Another example of this is given below.
Okay, so I'm now at the left-hand side of the waterfall. Do you wish me to go north?

Are you ... yeah ...
I want you to go up north, to the top of the waterfall.

Right.

5.5.3 ACK-FULL - Follow-up Information Offered

An utterance coded as +ACK-FULL will be fuller than that coded as +ACK-SHORT, and will usually comment on the information offered by the previous speaker, whereas a move coded as +ACK-SHORT will usually be a one word utterance, such as "uh-huh" as a minimal follow-up.

The reason for differentiating between these two types of follow-up utterances is to separate between the type of utterance which simply says 'I've heard you', and the one which gives this same confirmation, but also provides some comment on the information offered. Commenting on your partner's contribution shows more obvious understanding of what was said, and also a willingness to take information given into account. It also demonstrates that you, as current speaker, recognise that your partner has offered you this information for a reason - thus you are taking account of their goals as well as your own, and are therefore investing more effort.

In most cases, ACK-FULL is extra rather than being required, so an ACK-SHORT coding does not necessarily imply a -ACK-FULL coding. This is in contrast to the use of the +REPLY-MIN -REPLY-FULL coding discussed in the previous Section, and this differential is caused by the different nature of RESPONSE and FOLLOW-UP moves.
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.

Right.

Vertically.

Right, I think I know where it is,

In this extract, the follow-up in TB64 is coded as +ACK-SHORT because it does no more than acknowledge that the information offered has been heard: it does not necessarily imply understanding. But this is not problematic in itself, the speaker could have produced a fuller follow-up at this point, but it was not necessary that he should do so.

In contrast, the follow-up-type move offered in TB66 performs the function of the one in TB64, plus it demonstrates that the information has not only been heard, but that it has been used by the hearer in constructing their model of the situation. So this receives a +ACK-FULL coding, but again, it was not absolutely necessary, and had a brief follow-up been employed at this point, then it would not have been negatively coded for ACK-FULL.

Then you move in an angle slo ... sloping slightly upwards, to approximately a third of the way across the page from the right, and stop.

So what, is this ... Eh, there's a flight museum here, on the
It's not marked on my map.

Oh right, ehm, it's, eh, on the same level as the start position, basically.

Slightly below. Approximately on level with the top of the camera shop.

Right,

The example above illustrates the difference between negative coding for ACK-FULL and ACK-SHORT.

In TA17, the Giver gives short shrift to the Follower's introduction of a new feature. She does acknowledge it to some extent, but it can scarcely said to be recognising the Follower's goals in introducing relevant information. Generally speaking, Followers introduce information which may be useful in the negotiation of route-drawing, in order to speed the process and make miscommunication less likely. the Follower here has made an effort to introduce potentially useful information, even though he has not been requested to do so. Indeed, this feature, flight museum, provides the pivot to the route section which the Giver is trying to describe. The Giver's response fails to appropriately acknowledge the Follower's effort in making the contribution, and therefore the utterance is coded for an follow-up which is insufficient to the dialogue context: +ACK-SHORT, -ACK-FULL.⁸

⁸This is an example where the interactional level of talk is important to the transactional level. We have claimed elsewhere that they cannot be entirely separated, and this is why. If a speaker makes an effortful contribution to a dialogue, they expect such a contribution to be acknowledged by their partner. Why else should they expend more of their time than they absolutely have to? This principle generalises very well to household chores. If your partner does not notice that you have spent your whole evening doing some particular chore (e.g. washing-
TA19, however, is coded as -ACK-SHORT. Why is this? Even though the utterance might appear to be following on from the Follower’s helpful rejoinder to the Giver’s previously unhelpful follow-up, it is in fact referring back to the Giver’s original instruction in TA15. The instruction in TA19 reiterates to the Follower where their line should go, and does not attempt to consider where this line might be in relation to the flight museum which has been described by the Follower. Effectively, the Giver has completely ignored the Follower’s contribution to the conversation in TB18. Therefore, as there is no attempt at a follow-up where one is required, -ACK-SHORT is used.

5.5.4 ACK-REPEAT- Self-Understanding Checks

Such a check is usually comprised of a partial repetition of the information in the previous utterance. It has the function of an acknowledgement, and also acts as a query, because the other participant could query the understanding which is shown in the partial repetition (although there is no guarantee that this would happen).

Eaq4c5

#TA 31

... so, ehm ... Is the avalanche to the right or the left of the pine grove? >

#TB 32
It’s to the right.

#TA 33
To the right. Okay.
#G +ACK-REPEAT

This move attribute has a binary coding, and either is null, or is positively coded (i.e. this attribute is used). There is no concept of a failure to do a self-understanding check. If there has been a failure to check on the part of the Information Follower, then it will be coded as -QUERY.

up, ironing), then a) their lack of consideration will be pointed out b) you are unlikely to make the effort next time. We would argue that the same applies to dialogue.
FEATURE-INTRO - Check Existence of New Features

This is perhaps the most fundamental of the entity-specific dialogue attributes, and certainly the most highly used. It demands that the speaker should check whether their partner can identify a particular feature before using it in a route description. This enables the other participant either to locate the feature on the map, or to inform the speaker that the feature is unshared. Failing to make such a check can lead to problems with unshared features, which will not necessarily be resolved. These utterances will always be INITIATE moves: the introduction of features which occurs in RESPONSE or more unusually FOLLOW-UP will not be coded as +FEATURE-INTRO because the highlighting of the new information is provided by the other person asking the question, rather than the actions of the current speaker.

Such introductions of new features have been termed question introductions (Anderson, Clark, and Mullin, 1991b; Anderson and Boyle, 1994; Anderson, Clark, and Mullin, 1994), and are often typified by certain syntactic constructions like the ones given below:

- Have you got a ..?
- Do you have a ..?
- Do you see a ..?
- Can you see a ..?
- Is there a ... on your map?

In their work, Anderson and her co-workers have demonstrated that the use of these types of introductions are associated with greater task success in the Map Task. They are therefore low risk attributes, whereas the failure to use them must be considered to be a higher risk.

Other constructions are also perfectly acceptable. Our definition of this category is somewhat wider than Anderson’s. For our purposes this category refers to an intention rather than a precise syntactic form, therefore we do not require an interrogative construction. The intention we wish to capture is the highlighting of a potentially new feature, so this coding is appropriate providing that the new feature is the focus of the utterance, rather than being incidental to its purpose. So, as an introductory mention, I’ve got a boathouse. would be coded
as +FEATURE-INTRO, but *Do I go to the left of the boathouse?* would be coded as -FEATURE-INTRO because the focus of this utterance is *the left of the boathouse* rather than the introduction of the *boathouse* itself. A typical example of a highlighted introduction is given below.

Eaq4c3

#TA 5
...
Okay, ehm, have you got a ravine?  
#G +FEATURE-INTRO [ravine]

#TB 6
Yeah.

#TA 7
Right,

(16)

The following extract illustrates a non-highlighted introduction. In this case, the feature is shared, so the Follower has no difficulty in identifying the referent. If the feature is unshared, using such an approach might have been problematic. Note that the negative coding still occurs, even though no difficulty is encountered. It is the intention - the Effort and Risk levels - we are coding, not just the result.

Eaq4c3

#TA 1
Right, ehm, you’ve got to take the line down
from the start, to just ... vertically ... to just ...  
to the left of burnt forest.  
#G -FEATURE-INTRO[burnt forest]

#TB 2
To the left of burnt forest?

#TA 3
Mmmmm.

(17)
It should also be observed, that unlike Anderson, we do not restrict this category rigorously to first mentions of a feature. The function of this type of utterance is to establish that the status (shared/unshared) of a feature is mutually known. If there is a significant gap between the first introduction of an item, and the eventual use of that item in a description, then this vital piece of information can be lost amongst the wealth of other information necessary to the task. Human participants can forget whether they have asked about a feature, or can fail to remember the outcome of the original question.

Because of the nature of the Map Task, such re-introductions do occur. Features can be introduced at a point where they are not strictly relevant to the route - especially by the Follower, who does not know where the route is headed.

In the following extract, the Follower introduces the *secret valley* at a point where he thinks it may be relevant to the route (TB152). It proves not to be, and by the time it is relevant, thirty turns later (TA189), the Giver feels that it is appropriate to re-introduce the feature by using a highlighting mechanism.

Eaq4c7

#TB 152
< Have you got secret valley or submerged /
#F +NEW-QUESTION
#F +FEATURE-INTRO[secret valley]
#F +FEATURE-INTRO[submerged rocks]

#TA 153
Yes.

#TB 154
rocks? >

#TA 155
So I’ve got secret valley.

#TB 156
< Okay, so you want me to go /

#TA 157
Nowhere near

#TB 158
above secret valley? >
Nowhere near secret valley.

--- BREAK ---

And you got a secret valley?

Yeah.

Can you go down from there down the left-hand side of the secret valley?

Mmhmm.

If we think about this issue in terms of work on focus space (e.g., Grosz and Sidner, 1986), then this potential problem would make sense: referents soon lose prominence in a conversation. Therefore, re-introducing a feature through a highlighting mechanism can have the same benefits as a first-time introduction.

The precise coding for this is the only one which varies according to speaker. This occurs because of our evaluation of what each speaker should be doing as a minimum for the completion of this task. Positive coding should be employed when this move attribute is used by either speaker. However, it was decided that Instruction Followers should be multiply coded as either +RELEVANT-INFO +FEATURE-INTRO or +NEW-QUESTION +FEATURE-INTRO, because they were introducing new information as well as introducing it in this form. The role of an Instruction Giver demands that they should provide new information in order to describe the route, so +RELEVANT-INFO or +NEW-QUESTION coding does not apply to them in this context (Remember that we are coding what speakers do above what is necessary.).

Negative coding is used when a new landmark is mentioned, but not highlighted in an interrogative form. Instruction Givers’ utterances would be coded as -FEATURE-INTRO, and Instruction Followers’ utterances would be coded as +RELEVANT-INFO/+NEW-QUESTION distinction refers to the syntactic form of the utterance: statement or question. See Sections 5.7.1.3 and 5.7.1.2 respectively.
5.6 3 - Medium Effort

- QUERY
- OBJECTION
- KNOWLEDGE-MISMATCH
- FEATURE-LOC
- FEATURE-UNIQUE

These five attributes are grouped at a higher level of effort because they either take some responsibility for the course of the dialogue, or they demonstrate a high level of thought about their partner's level of knowledge, and the potential pitfalls of the task.

KNOWLEDGE MISMATCH, QUERY and OBJECTION are all concerned with the avoidance or resolution of miscommunication. They involve the challenging of false assumptions, and therefore take responsibility for the maintenance of mutual belief. It could also be characterised as the current speaker taking on the responsibility for the achievement of the other speaker's goals. It does, at the very least, demonstrate more of an attempt to guide the direction of the dialogue, rather than leaving it to the other participant to determine its progress.

FEATURE-LOC and FEATURE-UNIQUE are concerned with information about the precise location of features. Their use in an utterance show an awareness of the particular problems associated with this domain, and an intention to avoid miscommunication by providing useful and readily interpretable information. Again, this is concerned with the issue of mutual belief. In this case, the speaker is attempting to ensure no false assumptions are made - either by themself or the other interactant.

5.6.1 QUERY - Queries

These are clarificatory questions, and unlike those questions which would be coded as +NEW-QUESTION (see Section 5.7.1.2 below), they do refer directly back to a
particular utterance. Often a query refers to the utterance immediately preceding it. Queries are considered to be less effort than +NEW-QUESTIONs because they are prompted by the dialogue context, rather than being entirely initiated by the current speaker.

Queries appear in two varieties. They are either a check that information was understood correctly, or they request extra information which wasn’t in the original utterance. Checks of understanding can include repeating back parts of the route which have been described by the other speaker. The other participant is being asked to confirm the accuracy of the current speaker’s rendering; it is not judged as a NEW-QUESTION because it is a repetition of information already given, and is not soliciting unknown information. The latter type of QUERY, requests for information, suggest that some information necessary to the intention of the utterance was omitted. For instance, the Information Giver may intend the Follower to draw a section of route from the information given in the utterance, but if that utterance fails to provide a piece of salient information (e.g. which side of a feature to go round, where precisely to stop, etc), then the Follower either has to guess this information, or ask a query.

Eaq4c5

#TA 7
I want you to take the line underneath the carpenter’s cottage and then go vertically upwards through carpenter’s cottage and the ravine.

#TB 8
Between the two?
#F +QUERY

#TA 9
Yeah.

#TB 10
Yeah.

(19)

---

10 This latter type are often used in response to a high risk utterance made by the other speaker, which made unfounded assumptions.
Eaq3c4

#TA 116
Come down just past the remote village.

#TB 117
Right.

#TA 118
And go to your right.

#TB 119
How far?
#F +QUERY

#TA 120
Just past the remote village.

#TB 121
Right.

Extract (19) is an example of a check of understanding. The Follower’s query at TB8 is a reformulation of the instruction in TA7, so he is requesting confirmation from the Giver that his reformulation is an accurate representation of where she wants him to draw the line. The next extract illustrates the other variety of query. Here, the Follower is requesting a piece of information missing from the instruction in TA118 which the Giver then provides.

In terms of coding, positive coding is used when an utterance meets the definition given above. Negative coding occurs when the analyst can see that there was an obvious problem (e.g. use of unknown feature in route description, information given which clashed with the Giver’s position, etc), but the current speaker failed to query. Null coding is appropriate for all other cases.

Such queries are low risk attributes, but failing to use them when they are necessary is high risk behaviour. They are necessary to avoid the misunderstandings and map errors which would otherwise inevitably occur.

143
The two turns in the extract above are the last two utterances of that dialogue; there was no further discussion. The utterance TA61 is problematic in a couple of ways: it covers more than one route segment, it introduces a new feature without highlighting it -FEATURE-INTRO, and it doesn’t specify an endpoint for this section of route. Yet, the Follower does not query any of these failings. TB62 was therefore negatively coded for not using a query as it seems an occasion where such clarification is necessary. In this particular case, a map error was caused by the lack of discussion, and the Follower’s high risk behaviour of not querying the Giver for more information did not pay off.

5.6.2 OBJECTION- Objection to information Given

Utterances coded as +OBJECTION have one of the same functions as those coded as +QUERY: to question some element of information used in the previous utterance. In this case, though, the utterance takes the form of a statement rather than a question - and it is this difference that we are identifying.

---

11 This lack of querying is probably a result of the uncooperative responses of the Giver to previous queries in this particular dialogue.
In some sense, the speaker is introducing new information about the situation, yet it cannot be coded as +RELEVANT-INFO (Section 5.7.1.3) because the query is directly caused by the preceding utterance. But then, neither is it a query, because it isn’t a question. So the +OBJECTION coding is used for statements of there being a problem, and such an utterance is directly prompted by the previous turn. This type of utterance never has a checking function, it is always pointing out some problem in the previous utterance - it is literally an objection to some assumption made\(^\text{12}\), hence the KNOWLEDGE-MISMATCH coding (see Section 5.6.3 below).

A similar relationship of effort exists between OBJECTION and RELEVANT-INFO as between QUERY and NEW-QUESTION. Those moves which are considered to be a reaction against the previous one use less effort because the speaker’s decision about what to say and how to say it is effectively made for them, whereas for the higher effort option, all aspects of the move are decided by the speaker.

### 5.6.3 KNOWLEDGE-MISMATCH - Querying Different Beliefs

The dialogues used for this study concerned negotiation for a task where there was a mismatch on knowledge. In such situations it is quite common for misplaced assumptions of knowledge to occur. In the Map Task, this is usually demonstrated by the use of an unshared new feature in a route description. It is important that the hearer should challenge this assumption, and initiate the process of resolving

\(^{12}\) The reader may have noticed that a virtually identical utterance was coded as +ACK-SHORT, -ACK-FULL in extract 14 (Section 5.5.3). This is because of the different roles each of these utterances is fulfilling. Here, the Follower is providing information to the Giver, in order that a problem might be avoided. Thus, the utterance is positively coded. In the previous example, the Follower provides information (introduces an unshared feature) and the Giver gives the comment “It’s not marked on my map.”, which is not helpful information. It is rather a disregard of the effort which the Follower had invested into introducing that feature. Hence in that case, the Giver’s utterance was negatively rather than positively coded. This demonstrates the importance of utterance role at a particular point in a dialogue, an aspect which was emphasised by Power (1974) (See Section 2.5.1 for a discussion of this.)
the knowledge mismatch. It is this challenge to the other speaker's assumption that we are coding.

Similarly to the other attributes above, this one is only considered a medium effort attribute because its use is prompted by the previous utterance. Not using it, however, is considered to be a high risk, because its use is necessary precisely where a potential miscommunication has already occurred.

Anderson, Clark, and Mullin (1994) found that the discovery of such differences in knowledge was important in terms of task success. This result agrees with what we would predict. If you do not discover the differences in the maps, then not only are you more likely to make a map error, but you are also going to be unaware of the problem, and thus less able to resolve it.

In extract (23) below, there is obviously some mismatch between to the interactants' positions on the map, otherwise the Follower would have been able to move the distance specified in TA17 without there being a problem. Yet, neither of the interactants query why such a difficulty has occurred, or try to align their respective map positions. Therefore, this sequence will be negatively coded. It is the Giver's utterance which receives the negative coding because her turn is the first point at which a problem has become apparent, although it would also be appropriate for the Follower to pursue the cause.

Eaq3c2

#TA 17
Could you go to your right another three inches?

#TB 18
Eh, well no. A bit off the edge of the page!

#TA 19
Right!
Well, whatever you can manage.
#G -KNOWLEDGE-MISMATCH -QUERY

#TB 20
Okay.

The query used by the Follower in TB10 below enables the Giver to replan that section of the route, excluding the landmark springboks from that description.
The Follower has recognised a problem (albeit with the features rather than the route directly) and has acted upon that immediately.

Eaq4c8

#TA 9
Then draw a line, ehm ... a kind of ... Go from there, to just below the ‘s’ of ‘springboks’.

#TB 10
Don’t have springboks.
#F +KNOWLEDGE-MISMATCH +OBJECTION

#TA 11
You don’t have ... That’s right. Eh, okay.
Right, in that case just draw a horizontal line to just below the ... eh, to the the left of the ‘h’ of ‘highest viewpoint’.

Positive coding would be appropriate where a false assumption by the other speaker has been made, and the current speaker then draws attention to the problem. Negative coding should be used when a similar assumption has been made, but no such challenge is forthcoming.

Due to the nature of what this attempts to code, this particular move attribute almost always co-occurs with another one. This move attribute represents a high-level intention, and does not denote the precise way in which the query occurs. Typical ways in which the challenge is made is either by an interrogative (QUERY), or a statement which indicates there is a problem (OBJECTION). In example (24) above, it occurs alongside a objection, and in example (23), the utterance is negatively coded for a query in addition to the negative coding for not querying the apparently different beliefs (-KNOWLEDGE-MISMATCH).

5.6.4 FEATURE-LOC - Don’t Refer to Unnegotiated Unshared Features

Once a feature has been established as unshared, the Instruction Giver has two options, either to replan the telling of that section of route, omitting that feature, or to negotiate with their partner the position of the feature, prior to the
description. If they choose to use the referent without locating it, then a problem is likely to occur. This is high risk behaviour.

In the following extract, the Giver refers to *youth hostel* as though it was either a shared feature, or its position had been negotiated. In fact, *youth hostel* had previously been introduced by the Giver, and denied by the Follower.

Eaq4c6

#TA 59
Well, you ... The next stage is you have to draw a diagonal line from the right-hand side of the youth hostel down to
#G -FEATURE-LOC[youth hostel]

#TB 60
I don't have youth hostel.

#TA 61
Oh, right, well from that point that I've just gone round,

#TB 62
Mmmmm.

Not only is this not the introductory mention of *youth hostel*, but also the Giver had not discussed the location of *youth hostel* either at the moment when it was introduced, or at a subsequent point in the dialogue\(^{13}\). Therefore, the Giver was referring to the landmark as a known entity even though its position had not been determined. This approach caused a problem which was resolved through a replan by the Giver. However, the risk taken by the Giver is only recovered through the actions of the Follower: it was up to him to point out the false assumption.

In the following extract, a similar context exists, where the Giver had introduced *graveyard*, and the Follower had denied it. The ensuing conversation, however, is quite different.

\(^{13}\)It should also be noted that there was not a long gap between the introductory and subsequent mentions
Due east, and then up to the left of the carved stones, that way you'll avoid the graveyard, which is to the right of the diamond mine.

#G +FEATURE-LOC [graveyard]

Fine.

(26)

Here, the utterance TA20 is coded as +FEATURE-LOC because the Giver has situated graveyard in relation to known features, so that the Follower is aware of how this unshared entity fits in with the remainder of the map.

In summary, an utterance will be coded as +FEATURE-LOC if it begins the process of negotiating the position of an unshared feature, or refers to such a negotiated unshared feature. Note that a speaker can begin the process of negotiating the position of their partner's unshared features, e.g. "Where's your totem pole?".

Negative coding occurs for two phenomena: employing an already mentioned unshared feature in a route description14, and failing to initiate the process of negotiation when their fellow participant introduces an unshared feature. This latter condition is the reverse of the positive coding used for "Where's your totem pole?" above.

Such negotiation can occur in either of the three main move types (INITIATE, RESPONSE, FOLLOW-UP), although it occurring in a FOLLOW-UP move is reasonably unlikely. It is probably most likely to occur within instructions from the Giver (INSTRUCT moves), as in the examples given above, or in other types of INITIATE moves, such as NEW-SUGGESTION (Section 5.7.1.1 below).

5.6.5 FEATURE-UNIQUE - Don't Assume Features are Unique

In the Map Task, each map has a duplicate feature: for example, two boathouses. One of these features will be relevant to the route, and one won't be. The In-

\[14\] Unless it establishes the feature's position within the same utterance.
struction Follower only has the irrelevant feature, but the Instruction Giver has both. If the participants don't realise this problem, then errors can easily occur. In this case, utterances are coded positively when they imply that identically named features may not be the same. For example: "So where's your diamond mine?". A negative coding of this attribute does not exist, because, like the high effort INITIATE moves discussed below (Section 5.7), the use of this attribute is an optional extra. Participants cannot be said to have failed to use an attribute essential to the utterance if behaviour coded as +FEATURE-UNIQUE is not used. In the following example, there is only one remote village, and they are in the same place on each of the two maps, but the two interactants can't necessarily take that for granted.

Eaq3c4

#TA 92
... All I want you to do is clear the remote village.

#TB 93
< I have cleared it.
Right, the remote village is like, ah /

#TA 94
W ...

#TB 95
four inches below the pyramid in this? >
#F +FEATURE-UNIQUE

#TA 96
Yes.

(27)

The two speakers manage to verify through this that their two remote villages are in fact the same entity, and that they can refer to it as such.

Such a 'failure' could cause a task error, but, the error could be avoided in other ways: checks or queries on behalf of Giver or Follower would prevent most. Also, the use of negative coding for this attribute would effectively mean coding many utterances in this way (as most are not coded +FEATURE-UNIQUE), which would mask our other results.
Like the FEATURE-LOC coding, this type of behaviour is not restricted by move type. It is most likely to occur in information-bearing moves such as INITIATE or RESPONSE moves, but might equally apply to a fuller follow-up move, such as one coded as +ACK-FULL. This move attribute and FEATURE-LOC discussed previously are parallel attributes: the utterance will most likely be coded for something else (concerned with its move type) as well as being coded for one of these attributes\textsuperscript{16}.

5.7 4 - High Effort

- High effort INITIATE moves
  - NEW-SUGGESTION
  - NEW-QUESTION
  - RELEVANT-INFO
- INFO-INTEG

The highest effort grouping contain those categories which require the greatest level of initiation on the part of the speaker. Utterances which contain these attributes move the dialogue in a new direction - they do not rely entirely on the preceding utterance for their direction. It is an offer of new information, a suggestion about the next move, a request for new information, or a use of new information offered: the emphasis is on new. Therefore, Effort is being invested in the higher-level planning of the dialogue. The speaker is reasoning about their approach to the task, and then planning and producing an appropriate utterance. This grouping is in some respects the reverse of the first grouping. We would claim that a dialogue which uses these types of utterance is likely to show a large degree of commitment by the speakers to the task in hand, and can lead to useful discussions, and the avoidance of problems. However, these types of utterances are by no means necessary to a successful dialogue: many dialogues in our corpus contain very few examples, yet still complete the task successfully.

\textsuperscript{16}Of course, the same applies to +FEATURE-INTRO for the Follower, with the combination of either +RELEVANT-INFO or +NEW-QUESTION. See Section 5.5.5 for an explanation of this.
5.7.1 High Effort INITIATE moves

This group of move attributes is perhaps aimed more at the Information Follower in the context of the Map Task, although not exclusively so. Because of the nature of the task, the Instruction Giver tends to be the more dominant in the conversation, and it is possible for the Instruction Follower to let their partner do much of the work, at least in the sense that the Follower is reacting to information provided by the Giver. The move attributes described here, however, all involve initiating a new game, and are thus not utterances which have been directly solicited or prompted by their partner’s previous utterance (although there is obviously some progression). The utterances which come into this section push the dialogue (and hopefully the task) onwards: they are an initiation not an attempt at completion.

A brief summary of the intent behind this grouping might be “Don’t leave it all to your partner - be willing to take the initiative.” As we have suggested before, where there is initiative there is effort, so these move attributes are also considered to be effortful.

Positive coding is employed when an utterance meets the descriptions of the move attributes below.

These are all high initiative move attributes, and therefore rarely appear in a negative form: speakers can’t be said to have needed to make them, so they aren’t a failure in any respect. Negative coding is only used when there is an identifiable lack in terms of either utterance content, or choice of move type. For move-type choices, the only INITIATE dialogue attributes which are likely to be negatively coded are QUERY/OBJECTION (Obvious problem which has been ignored) or a lack of checking routines (at the end of route sections). There are certainly many occasions in the dialogues when one of the move attributes described in this Section could have been used, but it is not these (missed) opportunities we are coding. By their very definition, these high initiative dialogue attributes are extra to what is required in these conversations, and while we consider them to be useful to the process of information transferral, not using them will not in itself directly prevent the process.

The exception to this is RELEVANT-INFO which is more likely to appear in negative form, as it is more possible to label an utterance as failing to introduce relevant new information, such as a landmark which might be unshared. Here, we can pinpoint a specific lack in terms of information transferral, rather than being able
to say no more than “Maybe asking a question, or making some suggestions might have helped...”.

These move attributes are therefore considered to be low risk and high effort. Not using them is generally a low risk approach, which is why so few negative codings of them exist.

Whether these move attributes do, in fact, make an impact on the dialogues and their level of task success is considered in Section 6.3.1. The Principle of Cooperation (not Gricean Cooperation) predicts that such high effort attributes should have an impact. However, Wilkes-Gibbs’ (1986) findings (Principle of Collaboration) suggest that the reverse is true.

5.7.1.1 NEW-SUGGESTION- Make Suggestions

An utterance coded as +NEW-SUGGESTION will make a suggestion of how a potential problem could be solved, rather than introducing factual information. For example:

- The Instruction Follower proposing a way of circumventing a particular feature
- Either of the speakers suggesting a way in which to envisage the position of an unshared feature

In the following extract, the Follower suggests an alternative way of envisaging the stopping point for that part of the route. In this case, the Follower’s suggestion is off-track, but it should be remembered that we are coding the speaker’s intent, not necessarily the result.

Eaq3c8

#TA 21
Right, could you move up to about, eh, the top of the carpenter’s house, you know the middle of the top about two inches above.

#TB 22
To about the level of the ... the top of the haystack?
#F +NEW-SUGGESTION
To the level of the bottom of the haystack.

Right.

5.7.1.2 NEW-QUESTION - Ask New Questions

These questions request new information from the hearer. They are not clarificatory questions, and do not refer directly back to a particular utterance. Queries are coded as QUERY (see Section 5.6.1 above).

You'll see a graveyard. See a graveyard on your map? To the right of the diamond mine?

No.

Right. They've obviously not marked the graveyard.

How far to the right of the diamond mine is it?

The graveyard is almost halfway in between ... Do you have carved stones??
The question in TB24 is not immediately prompted by the utterance before it, which is why it is coded as a new question rather than a query. It does refer to the information given in TA21, but it is requesting new information related to the original question, rather than checking the content of the original utterance. The Follower did not need to ask about the location of the diamond mine at this point, it was not essential to a route-part. Therefore, this utterance constituted extra effort.

5.7.1.3 RELEVANT-INFO - Introduce Appropriate New Information

This move attribute refers to slightly different situations for the Giver and Follower, because of their contrasting roles in relation to the task.

For the Giver, utterances are coded as +RELEVANT-INFO when they provide some explanation of what is happening, over and above the basic instructions. This can either take the form of telling the Follower what the next route-part will be (e.g. “Now we’re going to go from highest viewpoint to safari truck.”) or it tells the Follower why the current route-part is being approached in a particular way. Both of these explain the current plan to the Follower, which helps him fit the information given into his understanding of the present position\textsuperscript{17}. In extract (30) below we have an example of the latter of these two types (the Giver explaining his motivation), the Giver’s utterance in TA43 provides extra comment about how the route fits together – the directions for that part of the route had been agreed on. The extra explanation is unsolicited, but it helps ensure that the communication which had gone before was successful, as it gives the Follower another opportunity for a query. This extra information from the Giver is both unsolicited and optional, and therefore effortful. These are the three important aspects.

For the Follower, +RELEVANT-INFO refers to the introduction of unsolicited information. This variant rarely (if ever) applies to the Instruction Giver because introducing the basic information relevant to the route they are describing is part of their role. It would be impossible for the task to be completed without the Giver performing this function, whereas for the Follower it is optional and constitutes extra effort.

The types of utterances which are categorised as this variant of +RELEVANT-INFO

\textsuperscript{17}This move attribute type recognises the importance of The Follower in the overall task: it is recognising that there is a need to share information. This task is not just a case of the Giver telling the Follower what to do.
mostly come into the following two sub-groups:

- Introduction of (potentially) unshared features relevant to the route
- Statement of current position e.g., "I'm at Bandit Territory"

So, for both speakers, utterances which come into this category introduce new information, which potentially only they can introduce, and which they have not been requested to produce. The type of information introduced should also not be obligatory for the completion of the basic task: this means basic instructions produced by the Giver are not be assessed in terms of this move attribute.

Eaq3c1

#TA 43
So if you're almost on top of the carved stones then you'll totally miss the graveyard.
#G +RELEVANT-INFO

#TB 44
Okay.
Right, so I am now just directly north of the carved stones and slightly north of the start.
#F +RELEVANT-INFO

#TA 45
< That's correct. So you move east around /

Extract (30) shows an example of the second type of +RELEVANT-INFO coding, where in TB44 the Follower gives a statement of his current position. This is a potentially useful action because any statement of position gives the other interactant the opportunity to check that their interpretation of your status is the same as the one you have just offered. Any misconceptions can then be located and repaired.
In this extract, the Follower introduces a new feature, although he hasn’t been requested to do so. Therefore the utterance is coded for the introduction of new information (+RELEVANT-INFO) and for the highlighted introduction of a new feature FEATURE-INTRO. The reader is referred to Section 5.5.5 for an explanation of this double coding.

5.7.2 INFO-INTEG - Negotiation of Information

This coding usually co-occurs with either +REPLY-FULL or +ACK-FULL. It is an optional extra coding to mark for the explicit negotiation of information in either RESPONSE or FOLLOW-UP moves. But what do we mean by negotiation in this context? What we are trying to highlight is any occasion where the speaker is providing more information than they might reasonably be expected to give - especially where they use information apparent in the previous move (be it a question, a give information or a response) in their presentation of the extra information. They are then ‘negotiating’ in the sense that they are recognising and recycling the information offered, and adding something extra to it: so it is negotiation in that there is ‘give and take’ in the information transferral process.
The concept suggested here has some aspects in common with Clark and Wilkes-Gibb's (1986) idea of refashioning, where one speaker offers extra information to a posited referring expression, and the other then accepts that new information by refashioning the referring expression, and integrating the extra detail into it. In the case of the Map Task, it is more the case of integrating new route information into the instructions, but the same principles apply.

For an utterance to be coded as +ACK-FULL +INFO-INTEG, it has to both comment on the information offered, and use it explicitly whereas for +ACK-FULL, only a comment is required. Like the three INITIATE attributes discussed above, this attribute is low risk/high effort, and rarely incurs negative coding, because not using it is also a low risk behaviour.

Eaq3c8

#TA 75

... Right, could you, eh, then move to the edge of the paper, on your left about another, I'll bet about another two inches I would've said.

#TB 76

Right to the edge? I mean it'll have to be completely to the edge because of the crashed spaceship.

#TA 77

Right there's a crashed spaceship there, right I thought you were ... Right move right to the edge enough ... try and avoid that crashed spaceship.

#G +ACK-FULL +INFO-INTEG

#TB 78

Okay, I'll do my best.

Here, the speaker is not only commenting on their partner's offering, but is actively engaging with it. The Giver could equally have replied with a one word utterance e.g. "Aha", or "okay" (+ACK-SHORT), or a somewhat fuller follow-up, e.g. "That'll be fine", or some such. Either of these types of utterance would have been perfectly appropriate.
In terms of taking on one's partner's goals, not only is there recognition of the goals, but there is positive reinforcement for those goals in the speaker's use of the information. This could perhaps be seen as encouragement for the partner to offer more information in the future, as their contribution as been reinforced as being useful.

It should be pointed out that negative coding for INFO-INTEG rarely occurs, as utterances are only negatively coded if they have failed to make the least effort-requiring appropriate response. Even then, the negative coding will be for the least effort dialogue attribute, not a higher effort one, such as INFO-INTEG. In most such cases, only a negative coding for ACK-FULL or REPLY-FULL would be made, as a comment on information given would have been deemed sufficient. Negative INFO-INTEG coding should only be employed where negotiation of information is considered necessary (and this is rare).

5.8 An Outline of Move Attributes

Here we will list briefly the move attributes, in terms of the general class of move they would be associated with. This type of categorisation does not capture all the move attribute types, as some of them do not operate at the level of the move, or are not fixed to particular move categories.

Inter-Utterance These apply to the variety of dialogue attributes within a move-type.

1. INITIATE moves
   (a) Initiators of checking routines:
       • CHECK-UNDERSTANDING
       • CHECK-COMPLETION
       • CHECK-ALIGN
   (b) Initiators of other games:
       • NEW-SUGGESTION
       • NEW-QUESTION
       • RELEVANT-INFO
       • QUERY
       • OBJECTION
2. **REPLY moves**
   - **REPLY-MIN**
   - **REPLY-YN**
   - **REPLY-FULL**

3. **FOLLOW-UP moves**
   - **ACK-SHORT**
   - **ACK-REPEAT**
   - **ACK-FULL**

**Unfixed Move-Types** These move attributes are not limited to one move-type.

- **INFO-INTEG**
  - This move attribute refers to the explicit use of just-offered information, and therefore would typically apply to **REPLY** or **FOLLOW-UP** moves rather than **INITIATE** moves.

**Parallel Move Attributes** This refers to some particular content of the move (concerned with *features*), rather than the move type it appears in. The move in question may be coded for the move attribute and for one of the following attributes concurrently, hence the term *parallel*.

- **FEATURE-INTRO**
  - This occurs only in **initiate** moves, as it refers to the unsolicited new mention of a feature.

- **FEATURE-LOC**
  - Refers to the negotiation of unshared features. Typically found in **INITIATE** or **REPLY** moves.

- **FEATURE-UNIQUE**
  - Refers to any reference to feature duplication, or differing locations. Found in information-bearing moves - e.g. **INITIATE** and **REPLY** moves.
Higher Level Move Attributes These attributes are specified at a higher level than moves, and are not tied to particular move types. It is the intention behind the attribute that we are attempting to categorise here.

- KNOWLEDGE-MISMATCH
  - Refers to the highlighting of a mismatch in beliefs.

- PARTNER-BELIEF
  - Refers to the continual questioning of a statement made by a dialogue partner.

5.9 Summary

In this Chapter we have presented the Typology of Move Attributes, described its underlying motivation and the importance of the notion of Effort to its structure.

In the initial part of the Chapter, the relationship between the Typology and other current work on dialogue analysis is considered. Comparisons are made between the approach taken here, and that of Birmingham School Discourse Analysis, Conversation Analysis and Dialogue Games. The prime difference is the emphasis on the evaluation of interactants’ moves in addition to the process of categorisation.

The concept of Effort is also discussed at this point. Each of the move attributes in the Typology is categorised according to Effort, and therefore the definition and use of this term is very important. We describe four levels of increasing Effort in terms of both social and task-based goals. We argue that these two aspects are interrelated, and should not be separated.

Next, the process of coding is described. This is particularly important given the evaluative nature of the Typology. Here, the distinction between coding what has taken place (positive coding) and what has been omitted (negative coding) is explained.

The remainder of the Chapter is concerned with the description of each move attribute, and examples to demonstrate their usage. This is organised in terms of the four effort levels, ordered from least effort to greatest effort.

Further details of the coding process are given in the documentation for the replication study (Appendix B). A full explanation of the replication study itself can be found in Appendix A.
Chapter 6

Empirical Tests

6.1 Introduction

In Chapter Three we presented a series of empirically testable hypotheses derived from the conversational principles which were previously discussed. The aim of this Chapter is to examine the results of the empirical tests which were carried out, and discuss the implications of these results for our view of conversation.

6.1.1 Coding

The testable hypotheses that we discuss below use the dialogue coding scheme which is presented in Chapter Five. We will make reference to both positive and negative coding. These terms are explained fully in Section 5.2.2, but we will summarise them briefly here.

Positive coding is based on the types of move attributes which speakers have used in an utterance. If an utterance meets the definition of a particular attribute type, then it will be coded positively for that attribute type. Each dialogue turn will typically be coded for one or more attribute types, depending on the move-type involved.

Negative coding is based on what speakers don’t do. Interactants make a series of choices when they speak. However, we would suggest that they do not always make the best choice, and that an analyst can assess the utterance in the dialogue context. When speakers don’t make the best choice, they are considered to have made a dialogue error, or failure, and the utterance will be negatively coded for the attribute type which the analyst believes would have been most appropriate. Such negative codings are not as common as positive codings, and will only apply
to a small subset of turns in a typical conversation.

In the context of these results, positive and negative coding totals are not just raw totals: they are *weighted* according to the effort level of the attribute type. Effort here simply refers to the amount of effort invested to produce the various attribute types. These levels and the weighting system are explained fully in Section 5.3, but in brief, lower effort attribute types receive low positive weighting and high negative weighting, and the reverse applies to high effort attribute types.

### 6.2 Hypotheses

Before we consider the results in detail, we will summarise the empirically testable hypotheses (Arabic numerals). These have been grouped according to both the type and focus of the test. The significance of the result to each of the individual hypotheses (roman numerals) will be considered in the appropriate Section.

#### 6.2.1 Hypotheses based on the comparison of whole dialogues: effect of high effort

1. High effort would be associated with task success
   - Hypothesis xv (Cooperation) - effort *would* correlate with success
   - Hypothesis x (Collaboration) - effort *wouldn’t* be associated with success *(null hypothesis)*
   - Note conflict between the predictions of the two Principles

2. Speakers with equal commitment (whether high or low) should be associated with more task success
   - Hypothesis ix (Least Collaborative Effort)
   - (by implication ix suggests least *individual* effort)

#### 6.2.2 Hypotheses based on the comparison of whole dialogues: effect of low effort

3. Risks would be taken
   - Hypothesis xi (Principle of Parsimony) - risks *would* be taken
   - Hypothesis xvii (Cooperation) - risks are *unlikely* to be taken *(null hypothesis)*
   - Note conflict between hypotheses
4. Low effort would be associated with low task success
   Hypothesis xvi (Cooperation)

6.2.3 Hypotheses based on changes in the dataset over time: changes in dialogue

5. Dialogue strategies will change
   Hypothesis iii (Gricean Cooperation)
   Hypothesis v (Coordination)
   Hypothesis xiv (Principle of Parsimony)
   Hypothesis xviii (Cooperation)

6. Risks would decrease over time
   Hypothesis xii (Principle of Parsimony)
   This assumes empirical hypothesis 3 is supported: risks cannot decrease unless risks are taken

7. Effort will decrease - effort minimised
   Hypotheses i & iii (Gricean Cooperation)
   Hypotheses vii & vii (Collaboration / Least Collaborative Effort)
   (by implication vii & viii suggest least individual effort)
   This assumes that empirical hypothesis 6 is supported: effort cannot decrease unless behaviour changes

6.2.4 Hypotheses based on changes in the dataset over time: changes in task success

8. Task success will improve over time
   Hypothesis ii (Gricean Cooperation)
   Hypothesis iv (Coordination)
   Hypothesis xiii (Principle of Parsimony)
   Hypothesis xix (Cooperation)

6.3 Empirical Tests

In this part of the Chapter we will present the method and results for each of our hypotheses in turn. Each Section will set out a hypothesis, and describe the
motivation behind the experiment designed to test it. We adopt $p \leq 0.05$ as the criterion for a significant result.

6.3.1 Hypotheses based on the comparison of whole dialogues: effect of high effort

These hypotheses are all concerned with the effect of high effort on task success.

6.3.1.1 1a. High effort would be associated with task success

This hypothesis states that effortful dialogue should lead to good task results, the implication being that the more effort you invest in a dialogue, the greater the task success that you should gain.

The Principle of Cooperation predicts that there should be a correlation between Effort and Task success.

Method In order to investigate this hypothesis, we will base our study on four associated measures: total positive score, average positive score, total high effort attribute types and the percentage of high effort attribute types in the dialogue. We will perform the same statistic (Spearman’s Rank Correlation) on all these measures.

**Total positive score** is simply the total of all the positive codings, with the adjustments (weighting) for effort levels. The total for each dialogue was calculated, and the dialogues were ranked both by individual quad, and by the whole dataset.

**Average positive score** is calculated by dividing the total positive score by the number of turns in the dialogue. This test was only performed on the complete dataset, not on individual quads.

**Total high effort attribute types** is the total incidence of three high effort initiatory attribute types: NEW-SUGGESTION, NEW-QUESTION and RELEVANT-INFO. These three attribute types were selected because they represent extra effort, and were not solicited by the other speaker: they are all INITIATE moves. They therefore demonstrate a choice on the part of the speaker to engage in an effortful utterance. As above, the total for each dialogue was calculated, and the dialogues were ranked by individual quad, and by the whole dataset.
Percentage of high effort attribute types is the percentage of utterances in a dialogue which are coded as using a high effort attribute type.

All these rankings were then separately compared with the ranking based on IE scores, which represents the task success measure.

The test comparing total positive score and task result is equivalent to Clark and Wilkes-Gibbs' (1986) operational definition of cooperation, which measures the amount of 'talk' in a conversation\(^1\). The other three tests (average positive score, total high effort level & percentage high effort level) are more effort-specific, or at least consider another aspect of effort measurement. This is because they are not reliant merely on dialogue length, but rather consider the level of effort per utterance, or the use of high effort attribute types. As we argued in Section 3.2.3.2, Clark and Wilkes-Gibbs' measure is open to criticism because dialogues could be long, but not be particularly effortful on an utterance by utterance level, or be long and only use the lower effort attribute types. Therefore, our set of tests will investigate the relationship between effort and task success more fully. If all are significant, then our demonstration of the support for this hypothesis would be stronger than that originally given by Clark and Wilkes-Gibbs, because of the greater breadth of our empirical tests. If none, or only some are significant, then we will have a clearer indication of how the relationship between effort and task success functions.

**Results**  Total positive score does not correlate with task success either in the quads, or in the complete dataset, neither does average positive score (\(p > 0.1\), for all conditions). There is a correlation between total high effort attribute types and task success for one quad (EAQ7, \(r = 0.795, p < 0.025\)), but the remaining results are not significant. Neither is there a significant correlation between the percentage of high effort attribute types used and task success (\(p > 0.1\)).

We can therefore conclude that this hypothesis is not upheld: high effort is not associated with task success.

\(^1\)Total positive score is strongly correlated with dialogue length (\(r = 0.931, p \leq 0.0005\) for complete dataset, one-tailed test), dialogue length is not correlated with task success (\(p > 0.05\)), in fact the negative correlation is nearly significant.
6.3.1.2 1b. Collaboration/Effort wouldn’t be associated with task success

This hypothesis is derived from the Principle of Collaboration, as described by Wilkes-Gibbs (1986). We use her measures to test the relationship between Collaboration (an alternative view of Effort) and Task success.

Wilkes-Gibbs uses the following measures of Collaboration:

These are adapted from Wilkes-Gibbs (1986)

1. **Number of words and turns in a dialogue**
   
   This measures the amount of effort which the speakers are investing: more words and turn = more effort

2. **Mean number of words per turn**
   
   A low mean number of w.p.t. should indicate many changes of speaker, and a sharing of the workload

We will test the correlation of the following with Task success, and provide the results in a table below:

1. Number of turns in the dialogue
2. Number of words in the dialogue
3. Mean number of words per turn
4. Mean number of words per turn (Giver only)
5. Mean number of words per turn (Follower only)

**Results**

<table>
<thead>
<tr>
<th>Tests of Collaboration vs Incorrect Entity Score</th>
<th>r value</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turns</td>
<td>-0.313</td>
<td>NOT SIG</td>
</tr>
<tr>
<td>Words</td>
<td>0.197</td>
<td>NOT SIG</td>
</tr>
<tr>
<td>Words per Turn</td>
<td>0.276</td>
<td>NEARLY SIG</td>
</tr>
<tr>
<td>Words per Turn (Giver)</td>
<td>0.053</td>
<td>NOT SIG</td>
</tr>
<tr>
<td>Words per Turn (Follower)</td>
<td>0.061</td>
<td>NOT SIG</td>
</tr>
</tbody>
</table>
Neither the number of words nor the number of turns in a dialogue are positively correlated with task success. In fact, the number of turns in a dialogue almost show a negative correlation ($r = -0.313, p \leq 0.1$, 2-tailed test).

The relationship between the mean number of words per turn and task success is almost significant. However, we suspected that this was simply a chance result. This is because in our dialogues, the amount of talk is shared very unequally between Giver and Follower\(^2\), and thus to average two separate distributions in this way would be inappropriate statistically. We investigated this further by testing the correlations of the two separate distributions (Giver’s w.p.t. & Follower’s w.p.t) with task result. We felt this would be a fairer test of our data, and more appropriate statistically. Neither of these were significant, and showed almost random distributions.

We therefore conclude that there is no correlation between Wilkes-Gibbs’ measure of Collaboration and the degree of Task success achieved. This supports the prediction made by Wilkes-Gibbs.

6.3.1.3 2. Speakers with equal commitment (whether high or low) should be associated with more task success

The emphasis here is on the impact of an individual’s desires on what is essentially a joint activity. The argument here is that speakers who have the same level of commitment are less likely to have conflicting goals. This is demonstrated in better task results.

**Method** Unlike Wilkes-Gibbs (1986), our speakers are not put into the role of either high or low-criterion speakers. Any difference we might find will be self-selected. Therefore, we needed to find an appropriate way of distinguishing such self-selected differences. We concentrated on the usage of the three high effort (level 4) attribute types (+NEW-SUGGESTION, +NEW QUESTION and +RELEVANT-INFO) which we justified in the previous Section. These categories represent extra effort, and are INITIATE moves - the speaker has chosen to invest the effort in the conversation. This is the nearest match to the concept of ‘high criterion’ speech.

In order to make any comparison between the dialogues, we needed to develop a way of measuring equality (or otherwise) of input on these three high level

\(^2\)Mean turn length for Giver = 11.02, mean turn length for Follower = 4.97. Percentage of talk provided by Follower to a dialogue varies between 15% and 45%, mean = 28.97%
initiate attribute types. For each dialogue, the total number of these attribute types for Giver and Follower were calculated. These then were both expressed as a fraction of the overall total of these attributes for the dialogue (i.e. Giver’s total divided by overall total, Follower’s total divided by overall total). These two fractions should add up to one. The difference in proportion was then calculated by subtracting the Follower’s share from the Giver’s share. This gives a figure between 0 and 1: the higher the figure, the greater the difference in the number of these attributes used by each speaker.

<table>
<thead>
<tr>
<th>No.</th>
<th>Giver</th>
<th>Follower</th>
<th>Total</th>
<th>Giver Prop</th>
<th>Follower Prop</th>
<th>Diff</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>6</td>
<td>18</td>
<td>0.67</td>
<td>0.33</td>
<td>0.34</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>9</td>
<td>18</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>16</td>
<td>18</td>
<td>0.11</td>
<td>0.89</td>
<td>0.78</td>
<td>3</td>
</tr>
</tbody>
</table>

As can be seen from this example, where the two participants use the same number of high level attributes (dialogue 2), the differential works out to be zero. However, where the speakers vary in their input, the differential increases. In dialogue 1, the Giver uses a third more high level attribute types than their partner, and thus the differential is calculated as 0.34 (to 2 significant figures). In dialogue 3, where the Follower uses 8 times as many high level attribute types, the differential is 0.78: much larger than the figure for the previous dialogues, thus emphasising the difference in proportional usage. Note that the differential always has a positive value - we are not interested in the polarity of the difference: it’s the absolute differential that’s important.

This procedure was carried out on all the dialogues. We then ranked them (smallest difference first). A Spearman’s Rank Correlation test was then performed to see whether there was any relationship between the proportion of high effort ranking and the IE score ranking (task success). This test was carried out on both individual quads and the dataset as a whole.

Results All these tests produced a significant result. The individual quads produced significances between \( p \leq 0.025 \) and \( p \leq 0.01 \) (one-tailed tests). The result on the whole dataset was highly significant \( (r = 0.820, p \leq 0.0005, \text{ one-tailed test}) \), and would support the experimental hypothesis.
6.3.2 Hypotheses based on the comparison of whole dialogues: effect of low effort

This Section is different from the others, as we are going to consider two different empirical hypotheses within one Section. This is because both these hypotheses draw on the same group of tests relating to low effort. Therefore, in order to minimise any repetition, we will first set out a description of the group of empirical tests involved and their corresponding results. Then, separate Sections for each of the hypotheses will demonstrate the relevance of the tests in that context, and discuss the importance of the results.

1. Does total negative score correlate with task success? The total of the negative codings, weighted for effort, was calculated for each dialogue. These were then ranked, and their relationship to a ranking based on IE scores (task success) was tested using Spearman's Rank Correlation. This was performed both on individual quads and the dataset as a whole.

2. Does average negative score correlate with task success? The average negative score for each dialogue was calculated by dividing the total negative score by the number of moves in the dialogue. These scores were ranked as before, and the relationship with the ranking based on IE scores was tested using Spearman's Rank Correlation. This was only performed on rankings based on the whole dataset.

3. Are the occasions of dialogue failure associated with negative score? The aim of this question is to look at a more specific relationship between instances of negative coding and task failure. Correlating task success and negative scores at the level of whole dialogues can only tell us so much. Here, we consider the level of dialogue segments. The empirical question is: is there any relationship between the number of negative codings in a segment, and whether that segment contains a task failure (an incorrectly navigated feature).

---

3The observant reader may notice that average scores for positive and negative codings are calculated differently: turns vs moves. This is because the codings work in slightly different ways. When I embark on a turn of speech, it is entirely up to me whether I use more than one move - thus it is a function of the turn. For negative coding, however, we need to consider the **number of opportunities** where the speaker can fail to do something. Negative coding is based on the unit of the move: each move is assessed for failure, rather than each turn. Thus, for the calculation of the negative average, we use the number of moves rather than the number of turns.
Firstly, each dialogue segment was classified as either ‘error’ or ‘no error’: no account was taken of the number of errors in a segment. Secondly, the number of instances of negative coding was calculated for each segment. These scores were divided into two conditions (low score, high score) along the median. A Chi Square test was then performed to test whether high score segments were more likely to be error segments.

As we have discussed before, shared features are easier to negotiate than unshared ones (Section 4.6.5). Therefore, we separated our original data into two separate conditions: segments containing shared features, segments containing unshared features. This was to ensure that any result we found was not affected by the type of feature involved. The Chi Square test was repeated, using the same conditions as above, on both sets of data.

**Results** We will consider each result in turn.

1. **Does total negative score correlate with task success?** All the tests were significant. The result for individual quads varied between $p \leq 0.025$ and $p \leq 0.01$ (one-tailed tests). The result for the whole dataset was $r = 0.820, p \leq 0.0005$ (one-tailed test), which is highly significant. A table of all the results is given below:

<table>
<thead>
<tr>
<th>Quad</th>
<th>$r$ value</th>
<th>$p$ value</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>EAQ3</td>
<td>0.854</td>
<td>$p \leq 0.01$</td>
<td>SIG</td>
</tr>
<tr>
<td>EAQ4</td>
<td>0.833</td>
<td>$p \leq 0.01$</td>
<td>SIG</td>
</tr>
<tr>
<td>EAQ7</td>
<td>0.835</td>
<td>$p \leq 0.01$</td>
<td>SIG</td>
</tr>
<tr>
<td>EAQ8</td>
<td>0.798</td>
<td>$p \leq 0.025$</td>
<td>SIG</td>
</tr>
<tr>
<td>DATASET</td>
<td>0.820</td>
<td>$p \leq 0.0005$</td>
<td>SIG</td>
</tr>
</tbody>
</table>

2. **Does average negative score correlate with task success?** This test was highly significant, $r = 0.573, p \leq 0.0005$ (one-tailed test).

3. **Are the occasions of dialogue failure associated with negative score?** All these tests were highly significant. The results are listed below:

- **All features**
  Task errors are more likely to occur where there are more dialogue failures
  
  $df = 1, X^2 = 57.52, p < 0.0005$ (one-tailed test)
- **Shared features**
  Task errors are more likely to occur where there are more dialogue failures
  \[ df = 1, X^2 = 17.20, p < 0.0005 \text{ (one-tailed test)} \]

- **Unshared features**
  Task errors are more likely to occur where there are more dialogue failures
  \[ df = 1, X^2 = 21.02, p < 0.0005 \text{ (one-tailed test)} \]

**6.3.2.1 3. Risks would be taken**

This hypothesis states that interactants will take the risk of being misunderstood in dialogues, which could cause task errors. They are willing to take risks because risky behaviour is less effortful for the speaker.

**Method** Here, we make the assumption that ‘risks’ are equivalent to behaviours which are negatively coded in the dialogues. The justification for this view is given in the descriptions of all the various dialogue attribute types in the Typology (Chapter Five). Therefore, we wish to test whether there is any relationship between negative coding and task errors. The tests given in Section 6.3.2 above were used:

1. Does total negative score correlate with task success?

2. Does average negative score correlate with task success?

3. Are the occasions of dialogue failure associated with negative score?

**Results** All the tests we have performed - which consider several ways of comparing task success and risk - indicate strongly that speakers do take risks\(^4\). If behaviour which incurred negative coding was not associated with a detrimental affect on task success, then we would not be able to make this claim. Such behaviour must be considered risky if it affects the outcome of the dialogue. Therefore, we must conclude that these results support the predictions of the Principle of Parsimony and not the Principle of Cooperation. Interactants do take risks in dialogue, and they do make errors: this does not fit with the idea of speakers being ‘helpful’.

\(^4\)Statistically, all these experimental hypotheses were supported. See Section 6.3.2.
6.3.2.2 4. Low effort would be associated with low task success

As we have suggested previously (Section 3.4.5), this hypothesis is, in some ways, the converse of the previous one. If high effort is associated with good task results, then low effort would be associated with poor ones.

Method  In order to consider this question empirically, we need to have some measure of what constitutes ‘low effort’. The operational definition which we will use here is based on extreme low effort, rather than ‘moderate’ effort (which would be difficult to quantify, other than ‘not high effort’ and ‘not extremely low effort’!). We take low effort to be represented by the behaviour which is coded negatively: not doing what either your partner or the task requires you to do at any point must be considered to be low effort.

The empirical tests which are relevant here are those which are described in Section 6.3.2 above. We list them here:

1. Does total negative score correlate with task success?
2. Does average negative score correlate with task success?
3. Are the occasions of dialogue failure associated with negative score?

Results  All our results are significant, our experimental hypotheses are supported so therefore we can conclude that low effort is likely to have a detrimental impact on task success. However, as we argued in the original description of these hypotheses (Section 3.4.5), this hypothesis can only support the Principle of Cooperation if its associated hypothesis (that high effort is associated with task success) was also supported. This is not the case, so therefore we conclude that the result of this hypothesis is better explained by the Principle of Parsimony and the Risk-Effort Trade-Off.

6.3.3 Hypotheses based on changes in the dataset over time: Changes in dialogue behaviour

The hypotheses in this and the following Section (Section 6.3.4) are concerned with any alteration in the way speakers interact as they gain experience of the task. Any such changes can provide a great deal of information about how we talk as we can investigate the type of changes which occur, and we can see the
type of language behaviours that speakers converge on. The latter of these is particularly important as it should demonstrate our preferred ways of speaking. Before we consider the results of these hypotheses, it should also be pointed out that participants were neither given any feedback on task performance, nor allowed to talk to any one else about the task. Neither did they see the routes which were drawn. Any changes which occurred were due to the experiences of the participants as Givers and Followers on the task, rather than any other information.

6.3.3.1 Dividing the dataset

In order to test for differences over time, the dataset needed to be divided into two parts, so that a comparison could be made. It was split into two conditions: 1st Givings versus 2nd Givings. 1st Givings all occur in the first half of a quad, 2nd Givings occur in the second half, and involve the same maps, the same Giver, but a different Follower. Each of the 16 speakers involved would participate in two dialogues under each condition, once as a Giver, and once as a Follower. Therefore, the two conditions produced equal size datasets (16 dialogues), and were symmetrical in all but the different Follower for the second Giving of the map.

This division was used to look for any difference in the dependent variable by ranking the whole data set, dividing the dataset according to the conditions above, and then testing the difference in the sums of ranks by using the Wilcoxon-Mann-Whitney statistic.

6.3.3.2 5. Dialogue Strategies will change

In the previous Section we have considered global changes in measurements of the dialogues, in this Section we focus our attention on changes in two particular move attributes.

The hypothesis of behaviour change is based on the prediction that speakers learn. Therefore, changes in behaviour are likely to occur, as speakers move towards using a particular set of strategies. This assumes that learning is based mainly in the way in which speakers go about achieving a task. This hypothesis can be tested by looking at changes in use of particular move attributes.
Method  The attribute types chosen for this study are FEATURE-INTRO and the three checking routines (CHECK-UNDERSTANDING, CHECK-COMPLETION & CHECK-ALIGN), which will be referred to jointly as CHECKS. These were chosen because they are not reliant on the course the dialogue takes: types of reply and acknowledgements may be predicted by previous moves, QUERY and OBJECTION are also produced by the effects of previous moves. FEATURE-INTRO and CHECKS are used due to independent choice: the choice of how to introduce a feature or whether to use a checking routine is always an open one. They also have the advantage of being fairly common attribute types, and appear in all the dialogues, unlike the high effort attribute types which we considered previously.

For the purpose of clarity we will describe briefly what behaviours are positively and negatively coded for each of these attribute types. FEATURE-INTRO refers to the way a feature is introduced into the conversation. A positively coded utterance will introduce a feature as the main focus of that utterance - it will not be in the context of a route instruction, and it will also not have been elicited by the other speaker. Introducing a feature in the context of a route instruction would be negatively coded. CHECKS typically occur either at the end of route segments5, or after an attempt at negotiating a difficult part of the route. An omission of checking at either of these points would often lead to a ‘failure to check’ coding. A fuller description of coding for FEATURE-INTRO and CHECKS can be found in Sections 5.5.5 and 5.5.1 respectively.

In order to test our hypothesis, we needed to calculate the totals per dialogue for each of these attribute types. The total is simply the number of times each attribute type occurred. Four totals were calculated: FEATURE-INTRO total usage (positive coding), FEATURE-INTRO failure to be used total (negative coding), CHECKS total usage (positive coding), CHECKS failure to be used total (negative coding). The dialogues were then each ranked according to each set of totals, and the dataset was again divided into two. The same conditions of 1st and 2nd Giving were used. This was tested using the Wilcoxon-Mann-Whitney statistic.

For FEATURE-INTRO, there was also another condition which we wished to investigate. Beside the condition of Giving-type, there is additionally the condition of sharedness: the feature which is being introduced can either be shared or unshared. The question we wanted to investigate here was whether speakers treated shared and unshared features differently. In particular, we predict that speakers will use more feature introductions (+FEATURE-INTRO), and will decrease the

---

5See Section 4.6.1 for a definition of ‘route segment’.
number of failures to use this attribute type (-FEATURE-INTRO).

It should be pointed out that interactants do not know whether features are shared or not, and any difference found here will be due to speakers’ experience of the task and the particular maps. This applies particularly to Givers on the condition of 2nd Giving.

**Results** First, we will consider the effect of time on CHECKS. There was no significant difference between the positive totals for the two halves of the dataset - the number used was very similar. However, the failure to use this attribute type decreased \((z = 2.83, p = 0.0024, \text{one-tailed test})\), so there was some change in behaviour.

For FEATURE-INTRO, the results are more complex. There was no significant difference using Wilcoxon-Mann-Whitney on the condition of experience between either the two sets of positive totals, or the two sets of negative totals. However, the result of the Chi Square test did show a change in behaviour for the failure to use the attribute type \((df = 1, X^2 = 6.1, p < 0.01, \text{one-tailed test})\). The table is reproduced below:

<table>
<thead>
<tr>
<th></th>
<th>Shared</th>
<th>Unshared</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Giving</td>
<td>31</td>
<td>32</td>
<td>63</td>
</tr>
<tr>
<td>2nd Giving</td>
<td>40</td>
<td>16</td>
<td>56</td>
</tr>
</tbody>
</table>

By looking at the table, it is easy to see where the difference lies: the totals for each half of the dataset may not vary much, but their distribution does. In the first half of the dataset, the datapoints are equally split between shared and unshared features. In the second half, the datapoints are biased towards the shared features. Far more shared features are introduced into the dialogue without their mutual existence being checked first. Anderson and Boyle (1994) found a similar effect with their corpus of school children playing the Map Task\(^6\)

This result constitutes partial support for the Principles of Coordination and Cooperation, as their hypotheses were fairly weak. For them, it was a 2-tailed hypothesis; the prediction was merely a difference, the direction of the difference was not predicted. In statistical terms, this means that the probability of this relationship occurring by chance should be doubled \((p < 0.002 \& p < 0.02\) for the failure to use checks, and the difference in use of feature introductions with shared

---

\(^6\)Their definition of “question-introduction” is slightly different to our use of FEATURE-INTRO. This difference is explained in Section 5.5.5.
and unshared features). The hypotheses of both Gricean Cooperation (Efficient) and the Principle of Parsimony are both considered to be supported, because the results show a re-focusing of effort. Effort is channelled into the introduction of unshared features, and away from shared ones; effort is invested in checking routines where they’re needed (decrease in negative coding), but as the overall number of such routines doesn’t increase, speakers must be decreasing their use elsewhere. Effort could be seen here as a scarce resource, which needs channelling in the most effective way.

6.3.3.3 6. Risks would decrease over time

This hypothesis is concerned with changes in the corpus over time: what effect does experience have? In Section 6.3.2.1 we have shown that interactants take risks which affect the task result. The question here is whether the amount of risks taken (i.e. behaviours which detrimentally affect the task result) decreases over time.

Method Here, as before, we assume that behaviours which incur negative codings are equivalent to ‘risks’. Therefore, in order to test the above hypothesis we will see whether the negative coding totals decrease as speakers gain more experience of the task.

In order to perform this test, we ranked the dialogues according to their negative coding totals. Then, this dataset was divided in two according to the conditions of 1st and 2nd Giving, as described above. The sum of the ranks was calculated for each condition, and tested for significance using Wilcoxon-Mann-Whitney.

Results The result of this test supported the view that negative score (and thus, risks) decrease over time \( z = 2.09, p = 0.0183, \) one-tailed test). This result also has implications for the effect of experience on task success. In Section 6.3.2.2, we have shown that the amount of risks taken is related to task success: the more risks, the poorer the task success. This should imply that if the number of risks taken decreases over time, then task success should improve over time. We test this hypothesis in Section 6.3.4.1.
6.3.3.4 7. Effort will decrease - effort minimised

Again, this hypothesis is concerned with changes in the dataset over time. Here, we wish to see whether the amount of effort speakers invest in a dialogue decreases as the interactants repeat the task.

**Method**  Effort here is measured by the type of attribute types used, the number of them that are used and the length of the dialogue.

Two types of tests were used: one based on the coding of dialogue attribute types, the other simply based on dialogue length (measured in turns).

First, we will describe the tests based on the coding of attribute types. Two were performed, one based on the total number of coded attribute types (weighted according to their effort grouping), and the other based on the average effort-level per utterance for a dialogue (total codings as above, divided by the number of turns in the dialogue. Both of these measures were used in order to determine whether any effect was caused, or masked, by dialogue length. Therefore, each dialogue had two scores associated with it: total positive score, and average positive score.

Each set of scores was ranked, and, as previously, they were divided into two conditions: 1st Giving and 2nd Giving. Both sets of ranks were tested independently using Wilcoxon-Mann-Whitney to see whether there was any significant difference caused by the condition of experience in either set of data (total positive score or average positive score).

Secondly, a test was performed based on any change in dialogue length. The dialogues were ranked on their length, and the dataset divided into two as before. A Wilcoxon-Mann-Whitney statistic was used to test significance.

**Result**  None of the tests were significant ($p > 0.1$ for all tests). The data do not support the hypotheses.

This does not support the Principle of Collaboration’s prediction of a decrease in effort (Collaboration). This is perhaps not surprising as our task is not based on a repetition of almost identical tasks, nor do our interactants always converse with the same partner. However, we would argue that this demonstrates that their prediction does not generalise particularly well.

Although the absolute value of effort does not decrease, one could argue that
like in the previous hypothesis (changes in behaviour), effort may well be re-distributed, and thus some kind of effort minimisation does occur. We would suggest that this would imply support for Gricean Cooperation. However, as we have argued previously (Sections 3.2.3.3 & 3.2.5), this result would certainly provide support for the Principle of Parsimony and the Risk-Effort Trade-Off.

6.3.4 Hypotheses based on changes in the dataset over time: Changes in task success

In the previous Section we have looked at the changes which occur in the dialogues over time. Here, we shift our focus to the output of the dialogue: the route drawn by the Follower. If dialogue behaviour alters, how does this affect the task result?

6.3.4.1 Task success will improve over time

In order to test this hypothesis, we needed to see whether task success was affected by the dialogue’s position in the quad: would later dialogues show better task results, and thus some evidence of learning?

Method The dataset of 32 dialogues was ranked according to the Incorrect Entity score gained by the dialogue\(^7\). This dataset was then divided according to the two conditions (1st Giving, 2nd Giving) as described previously.

Result The differences between these conditions were tested using Wilcoxon-Mann-Whitney, and produced a significant result \((z = 2.11, p = 0.0179, \text{ one-tailed test})\). Task result does improve as the interactants gain experience of the task. This supports the hypotheses of the Principles of Gricean Cooperation (Rationalist/Efficient), Coordination, Parsimony and Cooperation.

6.4 Discussion

Here, we will summarise the results and the conclusions that they suggest.

---

\(^7\)This is a score based on the number of features negotiated incorrectly - the higher the number, the worse the drawn route. A detailed explanation is given in Section 4.6.5.2
6.4.1 Gricean Cooperation

<table>
<thead>
<tr>
<th>Principle</th>
<th>Supported</th>
<th>Not Supported</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gricean Cooperation</td>
<td>task success improvement over time</td>
<td></td>
</tr>
<tr>
<td></td>
<td>behaviour modification over time</td>
<td></td>
</tr>
<tr>
<td></td>
<td>decrease in effort over time</td>
<td></td>
</tr>
</tbody>
</table>

The three hypotheses which were suggested for the Gricean view have all been supported. We see evidence of a decrease in effort (or, at least, a refocusing), behaviour modification and an improvement in task success. One could also argue that this interpretation of Grice has much in common with the motivations behind the Principle of Parsimony and the Principle of Least Individual Effort.

Efficiency is concerned with the minimisation of Effort. Grice certainly doesn’t make explicit comments about Risk and Effort, hence there being no hypotheses concerned with this in the original set in Chapter Three. However, the concept of the Risk-Effort Trade-Off would be consistent with the general aim of Effort minimisation. Therefore, the support for the hypotheses associated with Parsimony and Least Effort might be considered to lend more credence to this interpretation of the Gricean view. We will see in Section 6.4.5 that these hypotheses are indeed well-supported.

In Chapter Three of this thesis we offered two differing interpretations of Grice: Rationalist and Efficient. In the resulting empirical work, the predictions for both approaches were the same. Why, then, did we keep both analyses in the discussion? It is precisely because Grice’s work is so difficult to translate from a philosophical to an empirical framework. Readers of this work will no doubt disagree with some aspect of either (or both) characterisations of Grice. By taking this approach we wished to demonstrate three things. Firstly, small differences in definitions may not make much difference to predictions (as is true here), especially when you are trying to define something as slippery as Grice’s Cooperative Principle. Secondly, it is probably this slipperiness (which we have bemoaned previously) that means these predictions were not as specific as we would have liked. Finally, and most importantly, the main aim here was not so much to produce an impeccable interpretation of Grice, but to distinguish his views from the concept of ‘Cooperation’ as it is often used in the linguistic literature. We would suggest that these empirical studies have shown Grice’s views to have a closer relationship to the concepts of Parsimony/Least Effort than to the folklinguistic notion of ‘Cooperation as helpfulness’. It is this distinction that is most important to the work described here.
6.4.2 Coordination

<table>
<thead>
<tr>
<th>Principle</th>
<th>Supported</th>
<th>Not Supported</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coordination</td>
<td>task success improvement</td>
<td></td>
</tr>
<tr>
<td></td>
<td>over time</td>
<td></td>
</tr>
<tr>
<td></td>
<td>behaviour modification</td>
<td></td>
</tr>
<tr>
<td></td>
<td>over time (partial)</td>
<td></td>
</tr>
</tbody>
</table>

It is perhaps not unsurprising that the two hypotheses which were suggested here are underpinned by the notion of learning. The Principle of Coordination states that speaker behaviour should converge, and is likely to converge on useful behaviour (thus greater task success). As we have argued previously, this Principle seems merely to be descriptive: what is interesting here is why we behave in this way, not the fact that we do. As we have argued previously (Section 3.4.2), the Principle of Coordination does not motivate this sufficiently.

6.4.3 Collaboration

<table>
<thead>
<tr>
<th>Principle</th>
<th>Supported</th>
<th>Not Supported</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collaboration</td>
<td>equal commitment</td>
<td>decrease in effort over time</td>
</tr>
<tr>
<td></td>
<td>both speakers contribute</td>
<td></td>
</tr>
<tr>
<td></td>
<td>effort ≠ task success</td>
<td></td>
</tr>
</tbody>
</table>

Here, a number of the hypotheses are supported, but is there sufficient evidence to uphold the Collaborative Principle? Our empirical tests show (unsurprisingly) that both speakers engage in the dialogue task: both speakers’ usage of turns and words is Clark and Wilkes-Gibbs’ (1986) operational definition of the term. However, it is arguable whether this is a useful definition, and even if it is, it again does not tell us why speakers interact in this way.

The importance of equal commitment to task success, and the lack of relationship between effort and task success are more interesting. The combination of these results behaving as predicted makes the claim much stronger: the importance of equal commitment is emphasised, as no effect from this can be attributed merely to a general relationship between effort and task success. However, we would suggest that this result does not necessarily support the Principle of Collaboration.

As we have argued previously (Section 3.2.3.3), this outcome seems a strong indicator of the importance of individual input to a dialogue: its outcome is dependent as much on one’s partner’s intentions as one’s own. We believe that this is in conflict with Clark and Wilkes-Gibbs’ notion of language as joint action, where speakers collaborate to jointly produce both the language and the task.
result. From our results, and from Wilkes-Gibbs' (1986) results, we suggest that although language is undeniably joint in some sense, it is also the product of individuals who have their own agendas, and are unwilling to adjust their behaviour\(^8\) - especially if that adjustment means an increase in effort. Dialogue outcome is reliant on the self-oriented commitment of either speaker, even though such action can have a detrimental affect on the level of task success.

### 6.4.4 Cooperation

<table>
<thead>
<tr>
<th>Principle</th>
<th>Supported</th>
<th>Not Supported</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooperation</td>
<td>task success improvement</td>
<td>risks unlikely</td>
</tr>
<tr>
<td></td>
<td>behaviour modification (partial)</td>
<td>high effort = greater task success</td>
</tr>
<tr>
<td></td>
<td></td>
<td>low effort = low task success</td>
</tr>
</tbody>
</table>

We would argue that the two supported hypotheses are outweighed by the importance of the three unsupported ones. The unsupported hypotheses are more central to the underlying motivations of the Principle. In addition, we would suggest that the two hypotheses which are supported are better explained by other Principles (see below). Speakers do not seem to try to be particularly ‘helpful’ - they do take risks with task errors, they don’t necessarily adjust to their partner’s needs, and can, in fact, impede them (equal commitment, Collaboration, above), and their effort is not particularly associated with task success.

The two hypotheses which are supported (task success improvement) & behaviour modification could be seen to suggest ‘helpfulness’. Speakers learn what behaviours are helpful, and then orient towards them. However, we have many alternative explanations for these hypotheses, some of which are better supported (see Sections 3.4.4 & 6.4.5). We would suggest that these behaviours have often been interpreted as being ‘helpful’, and this has then formed the basis for the view of conversation as a cooperative activity. However, these views have not been substantiated empirically either here or elsewhere. ‘Helpfulness’ is one possible explanation for the improvement in task success, but the other behaviours which would be necessary to support this view have not been substantiated here.

Jarvella and Engelkamp (1983) in their discussion of the presentation of Given and New information in dialogue, also take the view that there are phenomena which may be interpreted as evidence for ‘cooperation’, but that we should be

\(^8\)Or not able to perceive any need for adjustment.

182
wary of making such assumptions.

“A too cooperative view of linguistic communication runs the risk of failing to recognise that many linguistic choices are probably dictated by states of the speaker’s mind rather than the listener’s, or out of consideration for the listener.”

Jarvella and Engelkamp (1983:227)

Cooperative dialogue should be listener-centred, not speaker-centred. We have no evidence for that here.

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

<table>
<thead>
<tr>
<th>Principle</th>
<th>Supported</th>
<th>Not Supported</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parsimony</td>
<td>task success improvement</td>
<td></td>
</tr>
<tr>
<td></td>
<td>behaviour modification</td>
<td></td>
</tr>
<tr>
<td></td>
<td>risks would be taken</td>
<td></td>
</tr>
<tr>
<td></td>
<td>risks would decrease over time</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[decrease in effort (refocusing of effort)]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[low effort = low task success]</td>
<td></td>
</tr>
</tbody>
</table>

Parsimony and Risk-Effort All the hypotheses for the Principle of Parsimony were supported, in addition, several hypotheses which were not directly related to this Principle produced results which fitted better with the explanation provided by the Risk-Effort Trade-Off than the original suggested explanation. These hypotheses are in square brackets in the table above, and we explain their inclusion below. We would therefore argue that the pattern of behaviour we have found here is consistent with what the Risk-Effort Trade-Off would predict. The trade-off between Risk and Effort seems to explain much of the behaviour and changes in behaviour which we have described in our data. In Section 3.2.5 we suggested that Risk-Effort is an instantiation of the Principle of Least Individual Effort: we do the least we can to get the job done. This is the Principle that we would argue that our interactants orient towards. This does not mean that they do not work well together, nor that they do the job badly, just that they try to conserve effort.

The important hypotheses here are those related directly to Risk and Effort. Our results show that speakers take risks, but also that they learn what constitutes
risky behaviour in this context, and decrease the level of risk. These two behaviours provide an explanation of why task success improves, and behaviour changes over time. None of the other Principles which predicted the outcome for those two hypotheses have been able to support their explanation empirically.

The other two hypotheses (decrease in effort & low effort = low task success which were originally intended for the support of other Principles (Collaboration & Cooperation, respectively) also give results which support the Principle of Parsimony. As we have argued elsewhere (Section 6.3.3.4 & 3.2.3.3), speakers may not decrease their effort overall, but they do seem to refocus its use: speakers are more careful not to fail to use FEATURE-INTRO for unshared features rather than shared features; the number of risks taken over time decreases (e.g. CHECKS, total negative score), but the amount of effort does not increase overall. The result associating low effort with low rates of task success is also more appropriate to the Risk-Effort Trade-Off than to the Principle of Cooperation (Section 3.4.5). According to Risk-Effort, high degrees of risk are likely to lead to high degrees of failure, which is what we see here.

**Principle of Least Individual Effort**

<table>
<thead>
<tr>
<th>Principle</th>
<th>Supported</th>
<th>Not Supported</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least Effort</td>
<td>All hypotheses supporting parsimony/risk-effort</td>
<td></td>
</tr>
<tr>
<td></td>
<td>effort ≠ task success (collaboration)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>equal commitment (collaboration)</td>
<td></td>
</tr>
</tbody>
</table>

As we have argued previously (Section 3.2.5), the Risk-Effort Trade-Off is an explanation of how the Principle of Least Individual Effort might affect the workings of a conversation. Therefore, all the hypotheses which were supported for the Risk-Effort Trade-Off also provide support for the Principle of Least Individual Effort.

In addition to this, we have also suggested previously (Section 3.2.3.3 & Section 6.4.3 above) that the proven hypotheses for the Principle of Collaboration and Least Collaborative Effort are better explained by the Principle of Least Individual Effort than they are by the arguments provided by their authors (Clark and Wilkes-Gibbs, 1986; Wilkes-Gibbs, 1986). This is because the complementary hypotheses effort ≠ task success and equal commitment = task success emphasise
the power of the individual's contribution to a dialogue. The equality of effort across the two individuals seems to be more important than absolute effort, however you measure it. In the empirical tests above we used six different ways to measure absolute effort/collaboration, and none of them showed any relationship with task success. As an interactant in a dialogue, you are completely reliant on what your partner's interpretation of 'Least Individual Effort' is. No amount of effort you can invest in a dialogue will make up for a low criterion on the part of the other speaker.

Therefore, we conclude that there is strong support for the Principle of Parsimony and Least Individual Effort.

6.5 Summary

From the results that we have found in our empirical study, we would argue that the Principle of Least Individual Effort (Principle of Parsimony and the Risk-Effort Trade-Off) is the most strongly supported of all the behavioural Principles which we have considered.

All the hypotheses which were supported could be motivated by the Principle of Least Individual Effort:

- Speakers learn over time
- Speakers modify behaviour over time
- Speakers take risks
- Speakers take less risks over time
- Speakers decrease (refocus) their effort input over time
- Low effort leads to low task success
- High effort does not lead to good task success
- Equal commitment by speakers leads to task success

Parsimony/Risk-Effort account for the first five of these hypotheses. The remainder were originally formulated to test the Principle of Collaboration and Least Collaborative Effort, but we have argued that the results could be explained more effectively by the Principle of Least Individual Effort.
The patterns of behaviour which we have tested here can be summarised as follows:

Speakers take risks which cause errors, but they learn how to focus their effort appropriately for the new task, and improve their level of success as they gain experience. In terms of effort usage, mismatches of effort mean clashes of individual agendas, and worse task results because low criterion speakers are unwilling to increase the effort which they invest.
Chapter 7

Conclusions and Further Work

7.1 Introduction

This final chapter has three parts. Firstly, we will summarise our conclusions with respect to our analysis of Dialogue Principles. Secondly, we will assess the contributions of this work, and finally, we will consider further research which could be undertaken to promote deeper understanding of this area.

7.2 Summary of Findings

This thesis has been concerned with the question of how interactants operate in dialogues. As users of language, we are all aware of the myriad choices available to speakers, but we have less information about how and why those decisions are made. It is this issue of motivation that we have concentrated on.

Several Dialogue Principles have been suggested to account for these speaker choices. We have concentrated on the following five: Gricean Cooperation, Cooperation, Coordination, Collaboration, Parsimony. These terms are discussed and evaluated in Chapter Three. We developed operational definitions of each of the Principles, which were then tested using task-oriented dialogues from the HCRC Map Task Corpus.

The results most strongly supported the Principle of Parsimony and the Principle of Least Individual Effort, which, we have argued, are equivalent. Speakers take risks which cause failures, but as they learn which risks aren’t worth taking (in terms of the Risk-Effort Trade-Off), they decrease the risks they take, and the level of task success improves. Although the amount of absolute effort doesn’t decrease, we do see evidence of behaviour modification, as speakers focus the
effort they are willing to invest in the most productive manner. Additionally, we also argued that the lack of relationship between absolute high effort and task success, and the correlation between equal effort (even if low) and task success also support Parsimony/Least Individual Effort. This is because these complementary predictions demonstrate the importance of individual input to a dialogue. If two interactants have mismatched effort criteria, then regardless of the absolute effort invested in the dialogue, they are unlikely to produce as good a task result as partners with matched effort criteria, even if their absolute effort input is relatively low. These two results originated as supportive hypotheses for Clark’s notions of Collaboration and Least Collaborative Effort, and were the only two hypotheses accepted for these Principles. But as we argue in Section 3.2.3.3, we cannot see how these behaviours can be explained within his theory. The behaviour which Clark terms Least Collaborative Effort would be better described by the Principle of Least Individual Effort. Yes, language is a joint production in some respects. However, it would seem we are at the mercy of our partner’s beliefs and corresponding effort input rather than necessarily being aided by them.

The notion of ‘Cooperation’ is more complex. Firstly, we divided this term into two: Gricean Cooperation, Cooperation. This was to clarify the difference we perceived between Grice’s intentions in defining the Cooperative Principle (Grice, 1975), and the way in which the term ‘Cooperation’ is often used in the discussion of natural language dialogue. In brief, the difference might be summarised as the assumption of rationality/conventions of language (Gricean Cooperation) versus being ‘helpful’ (Cooperation). In terms of the empirical testing, Gricean Cooperation was better supported than Cooperation.

Because of the difficulty involved in translating Grice’s philosophical analysis of indirect utterances into empirically testable hypotheses, we originally suggested two different interpretations of his work: rationality versus efficiency. These two interpretations are closely related; our the empirical definitions were unable to distinguish them. Indeed, efficiency could be seen as a form of rational behaviour, and essentially our empirical hypotheses reflected the efficiency interpretation. The concept of rationality in our evaluation is retained because of Grice’s emphasis on this as a guiding force in human behaviour (Grice 1975:47-48). In the end, they both predicted the same sort of changes in behaviour, task improvement and decrease in effort as the Principle of Parsimony/Principle of Least Individual Effort.

1It is possible that other types of data may have afforded a way to separate them. Ours did not.
Both rational and efficient speakers would be expected to learn and thus modify their behaviour accordingly to take less risks. We would also expect to see a decrease in effort - the concept of efficiency, in particular, would predict that. Our data showed a re-focusing of, rather than an absolute decrease in, effort. This is consistent with a speaker learning where they can take risks (effort minimised), and where the risks are not worth taking (effort increased).

In fact, this interpretation of Grice could be seen as another aspect of the ‘Least Individual Effort’ argument. The concept of Efficiency as derived from Grice (1975, 1978) does not make explicit predictions about such issues as Effort and Risk, nor the relationship between these and task success. Hence their lack of inclusion in the original set of hypotheses for Gricean Cooperation. However, the type of relationships which the Principle of Parsimony/Principle of Least Individual Effort predict are in keeping with the concept of Efficiency. Indeed, even the emphasis on the importance of the individual contribution is obvious in Grice’s work: the design of indirect utterances mainly benefits the speaker - the addressee has to work harder to access the intended meaning rather than the literal one.

Whether one can push the interpretation of Grice’s work quite this far is probably an individual decision. Certainly, the possibilities are there. This lack of clarity lies in the problem of interpreting and generalising Grice’s work. However, the aim of this work was primarily to differentiate the Gricean view from the use of the term ‘Cooperation’. Whichever of the above interpretations of Grice the reader favours, we believe that we have demonstrated that the Gricean view is closer to the ideas of Parsimony and Least Effort than the folklinguistic notion of ‘Cooperation’ often espoused in the linguistic literature. However you define Grice, the difference between his technical use of the word ‘Cooperation’ and its everyday meaning should be clear.

Cooperation in this everyday meaning of ‘helpfulness’ gained very little support. Its prediction that task success would improve was upheld, but the motivation for that prediction was not. Cooperation suggests that task success would improve because the speaker would learn what was helpful, and then use more of that behaviour. If one equates ‘effort’ with ‘helpfulness’ as we have done here, then you would expect to see a relationship between higher effort and task result. This was not substantiated. The reverse of that relationship - low effort being

---

2 Or any alternative she wishes to suggest
3 Where ‘effort’ could be thought of the amount of thought and design you put into an utterance. See Section 5.3 for further discussion of this concept.
correlated with low task success - was supported, but we argue that these two hypotheses must both be upheld before the motivation behind them can be accepted. If unhelpfulness is correlated with poor results, then helpfulness should be correlated with better results (Section 6.4.4). At the very least, cooperative interactants should not impede their partners, yet the relationship between equal commitment and good task result, and the converse lack of relationship between high effort and task result suggests that speakers do not even manage to achieve this. In addition, cooperative speakers should also avoid behaviours which are associated with errors. At least, they should learn which behaviours are associated with errors, and use more 'helpful' alternatives. However, our speakers often take such risks, and they continue to do so throughout the dialogues. Admittedly, there is a decrease in these behaviours associated with errors as the speakers gain experience of the task, but they do not disappear. Speakers continue to take risks which jeopardise task success; we would argue that this is not 'helpful' behaviour. We would therefore suggest that there is little evidence to support the notion of Cooperation which we have used here.

The final Dialogue Principle which we considered was Coordination. This concerns the 'matching' of one speaker's behaviour to another's. Clark (1985) describes it in terms of the set of moves we might go through in order to avoid someone else on the street: if you move left and I move right, then we won't solve our problem. He suggests that we coordinate at all levels of language: from the marking of Given and New entities, to turn-taking, to common ground. However, this Principle appears to be mostly descriptive. Yes, we may behave in this way, but the concept of Coordination does not explain how or why. It is no surprise, then, that the two hypotheses we were able to develop essentially predict a change in behaviour, but do not motivate it other than to suggest that speakers should converge on successful behaviours. Yes, speakers do modify their behaviour, and they do produce better task results, but Coordination does not tell us why.

7.3 Evaluation

The research described in this thesis provides a useful empirical study of philosophical notions. These concepts like 'Cooperation', 'Collaboration' and 'Coordination' are too often taken for granted, and used overly frequently in linguistic analysis, without any real consideration of their meaning. We have evaluated each of these Dialogue Principles according to the evidence offered by their proponents, and according to an empirical study. This has provided both some evidence of
what speakers do, and also the motivations behind the choices which they make. Of course, this is not to say that this research has been unproblematic. In order to apply empirical techniques to a philosophical area, we have had to make practical interpretations of each of the Dialogue Principles with which others may not agree. This includes our decision to treat these Principles as competing theories of dialogue structure. We can only reiterate why we made those particular decisions, and how we believe they are justified.

The dialogue analysis which we have used has also made a contribution to this area of linguistics. There are many methods of dialogue analysis, but as far as we are aware, they all are descriptive. For certain purposes, using an **evaluative** system seems most appropriate. The descriptive system seems to make the assumption that human dialogue is always effective and successful, whereas our starting point is that speakers do not always make the best choices. Our aim in this system is to enumerate the important available choices, and study their effects on task success. After all, it is fairly obvious that speakers use utterances which lead them into miscommunication and repair sequences. Our system provides some explanation of why this might happen.

Using a system such as this to generate data for empirical testing will always be open to criticism. Any dialogue labelling method will have some element of subjectivity, however much you try to tighten the classification system. Dialogue analysis is not an exact science, and we would argue that an analyst has to be aware the problematic nature of their task, but that in itself shouldn’t prevent them from using quantitative methods. In this work, we have tried to demonstrate that our coding system is reliable by using a replication study. This shows that other coders can be taught to make the distinctions which we believe are important in dialogue.

The results we have found, with their empirical support for the Principle of Parsimony, Least Individual Effort and the Risk-Effort Trade-Off have implications for pragmatics, dialogue analysis, and their computational implementations. More information about how we act, and why we might make those decisions is interesting from theoretical, social, and practical points of view. However, this is only the first evidence for the importance of this Principle (Parsimony) rather than those which have been more traditionally used. From our evidence, Parsimony seems to make stronger and clearer predictions than Gricean Cooperation, and quite different ones to those predicted (and not supported) for the folklinguistic notion of Cooperation. Further work needs to be done in order to strengthen the
support for this analysis of dialogue.

7.4 Further Work

Several aspects of this work require further investigation. Firstly, it would be useful to replicate this study using a different set of dialogue data. Alternative task-oriented dialogues producing similar results would strengthen the overall findings. In the long term, we would want to extend the study to different types of dialogue\(^4\). Broadening the research would lead to better indications of speaker behaviour in different situations, and it would also demonstrate the flexibility of the dialogue analysis method which we have presented.

More thorough testing of the Principles could also be achieved. In this study, the operational definitions of the Principles were based on what features we were able to test in the data. Now, while we would argue that the HCRC Map Task Corpus is fairly well-suited to such an investigation, it might be possible to achieve a more in-depth study by collecting data dedicated to this particular task. This would allow operational definitions closer to theoretical ones, as there would be fewer constraints on their production.

Given the current support for the Principles of Parsimony and Least Individual Effort, it would also be interesting to study the Risk-Effort Trade-Off in more detail. In this work we have concentrated more on the risks speakers have taken rather than how they recover from them. We have hypothesised that some dialogue risks/problems are easier to recover from than others, but we have not empirically tested this. Carletta (1992) built recovery strategies into her dialogue system, JAM, which complemented her agents' communicative posture: high or low risk (Section 2.5.2). This implementation is a useful beginning to this project, but it does not distinguish between low effort and high effort recovery strategies. To gain a true idea of this trade-off between Effort and Risk, and how effective interactants are at finding this balance, a detailed analysis of miscommunications and the effort of the recovery sequence would have to be undertaken. Our coded dataset would provide some information on this area, but the analysis system would have to be further refined in order to distinguish all the important elements.

In this study we have concentrated on three broad categories of move: INITIATE, RESPONSE and FOLLOW-UP. We have not considered INSTRUCT moves. This is

\(^4\)For example, naturally occurring data, maybe casual conversation.
problematic in studying dialogues which are concerned with information transfer-
rereal. Why did we do this? Initially, the Typology of Move Attributes did contain
a set of attributes which referred to the content of INSTRUCT moves. These
considered three things:

1. Is there too much information in the move?
   - Is there too much new information?
   - Is there too large a section of route being covered?

2. Are good start and end points provided for the route description in the
   move?

3. Is the move well-structured?

However, these aspects of discourse content and structure proved to be beyond
the scope of this work. Within the constraints of this project we were unable to
define these concepts in such a way that other coders could be trained to reliably
agree on the coding used. Therefore, all these categories were excluded from our
final analysis. It should be noted that this exclusion did not affect any of our
results. This compromise of removing these categories is not very satisfactory,
and we believe that any further extension of this work should re-instate this aspect
of the analysis. This area in particular would provide rich study for the question
of the Risk-Effort Trade-Off, as many potential miscommunications are contained
within INSTRUCT moves.

7.5 Conclusion

The concepts of Coordination, Collaboration, Parsimony and Cooperation, in
particular, have been much talked about in the literature of pragmatics, psy-
cholinguistics, discourse analysis and computational linguistics. However, these
terms are rarely defined in a meaningful way, and thus any analysis which uses
them is, in effect, problematic. In this thesis we have tried to address this lack of
clarity in two different ways.

6These categories are described briefly in the documentation which was provided in our
replication experiment - see Appendix B.
7In all cases, it neither affected whether a result was significant or not, nor the level of
significance achieved.
Firstly, we have tried to evaluate each of these Principles theoretically. This showed an important difference between the Gricean Cooperative Principle and the more general use of the term ‘Cooperation’: Grice is concerned with the exploitation of language conventions and the minimisation of effort, not with some notion of ‘helpfulness’. We also questioned the validity of Clark’s Principle of Least Collaborative Effort. The evidence of the work in this area suggests that dialogue behaviour would be better explained by a Principle of Least Individual Effort. Dialogue may be a joint product, but it is engaged in by interactants who have their own agendas. When these agendas clash, the task result is affected. One speaker’s investment of effort is unlikely to overcome the problem of mismatched effort criteria. Finally, we suggested that Coordination is a descriptive rather than a motivating principle: speakers coordinate because it is less effort to do so.

Secondly, we developed definitions for each of these Principles which could be tested experimentally. The following hypotheses were supported by the statistical tests:

- Speakers learn over time
- Speakers modify behaviour over time
- Speakers take risks
- Speakers take less risks over time
- Speakers decrease (refocus) their effort input over time
- Low effort leads to low task success
- High effort does not lead to good task success
- Equal commitment by speakers leads to task success

These results give the most credence to those Dialogue Principles which are concerned with minimising effort. This is primarily the Principle of Parsimony/Principle of Least Individual Effort, which we have argued to be essentially the same. However, in general terms it also supports the Gricean view. Grice’s work, being more philosophical in approach, makes fewer explicit hypotheses, and is thus harder to test experimentally. But, we would argue that the concept of Least Effort is consistent with Grice’s approach to implicatures. This outcome underlines the basic
difference between Grice’s Cooperative Principle and the folklinguistic notion of ‘Cooperation’ which we argued for in Chapter Three.

The aim of this thesis was to clarify the use of a particular set of technical terms, which also had everyday, non-technical meanings. Our argument was that the non-technical uses often blurred the technical distinctions. We believe that we have achieved this aim, at least in part. Our findings have the potential to be useful to the many areas of research which employ dialogue analysis.
References


Appendix A

Replication Study

A.1 Introduction

This Appendix describes a study which was carried out to test the replicability of the Typology of Move Attribute Types\(^1\). By showing the reliability of the scheme, we can demonstrate that the coding system is sufficiently objective and consistent to be taught to naive coders. This would further support the statistical results which are described and discussed in Chapter Six.

A.2 The Replication Study

This study is intended to test whether the coding scheme described in Chapter Five (the Typology of Move Attribute Types) can be taught to naive coders.

The Typology, like most dialogue analysis schemes, is subjective, and such schemes are always subject to the charge that they are 1) ill-defined and 2) not replicable by other coders. This study was designed to demonstrate that our coding scheme is replicable by previously untrained coders. In essence, we are trying to show that naive coders can be trained to code data using our system, and that the resulting coding should be statistically similar to that produced by an expert.

A.2.1 Subjects

The experiment involved two volunteer coders, and one expert coder. The naive volunteers were both PhD students in the Department of Linguistics at the Uni-

---

\(^1\)When this study was carried out, this system was called 'The Typology of Dialogue Attribute Types'. This is the terminology used in the material given to subjects, which can be found in Appendix B and Appendix C.
versity of Edinburgh. Both were familiar with the Map Task, although neither used the data from the HCRC Map Task Corpus in their research work. One coder was also familiar with dialogue game coding (Kowtko et al., 1992), and had game coded part of the HCRC Map Task Corpus.

The expert coder was myself. I re-coded the dialogues after a delay of approximately 18 months. This was to demonstrate the stability of the coding system over time.

A.2.2 Materials

The test data consisted of:

- Two practice dialogues, with accompanying Giver and Follower maps
- Two test dialogues, with accompanying Giver and Follower maps

The subjects were provided with the following information about the Typology:

- A full guide to the Typology, equivalent to the description given in Chapter Five
- A set of instructions, which included:
  - A guide to the coding process
  - A summary of the Typology (reproduced in Appendix B)
- A tabular summary of the coding system (reproduced in Appendix C)
- A tree diagram of the coding system

A.2.2.1 Presentation of the Materials

The dialogues were presented in a form designed to focus the coder's attention on the issues of paradigmatic choices, rather than allowing them to become involved in other issues of utterance coding, for example, the placing of move boundaries and broad move types. Thus, the dialogue was divided into moves prior to presentation, and its broad move type was stated. The coders were asked to write the appropriate codings (both positive and negative) on the lines provided.
This is a short example of how the dialogues were presented and coded. It is adapted from the example in the *Instructions for Replication Study*, which was used by the participants, and reproduced in Appendix B.

\[T(A,B) = \text{turn label } (A=\text{Giver}, B=\text{Follower})\]
\[G = \text{Coding for Giver}\]
\[F = \text{Coding for Follower}\]

Example turn as presented for coding:

Eaq3c3.text

\[\#TB 6\]
Past the forest fire?
\[\#F7 \text{ INITIATE}\]
\[\#F7\]
\[\#G7\]

Example turn as it should be coded (Codings in bold type):

Eaq3c3.text

\[\#TB 6\]
Past the forest fire?
\[\#F7 \text{ INITIATE +QUERY}\]
\[\#F7 -\text{FEATURE-INTRO} [\text{forest fire}]\]
\[\#G7\]

Explanation of coding:

\[\#TB6\]
<text>
\[\#F7 <\text{move type}> <\text{positive coding}>\]
\[\#F7 <\text{negative coding}>\]
\[\#G7 <\text{negative coding for other speaker}>\]

A.2.3 Method

The experiment was in four stages:

- Introduction to the system
- Practice coding of two dialogues
• Feedback on practice coding
• Coding of two test dialogues

The first stage involved describing the system and how it worked to the participants, and answering any immediate queries. They were shown a sample of coded dialogue, and the process of coding was described. This was an informal session. At the end of this session, they were given two practice dialogues to take away and code; this constituted the second stage.

After the participants had completed the practice coding, another meeting was organised. In this session, the participants compared their coding with the original coding, and any discrepancies and misunderstandings were discussed. This, again, was an informal session, and the participants were encouraged to air any uncertainties that they had.

As the fourth stage, the participants were asked to code two test dialogues. They were told to work on their own, and not to collaborate with another coder.

A.2.4 Method of Analysis

By undertaking this experiment, we produced 4 sets of codings:

1. Naive coder 1
2. Naive coder 2
3. Expert coder (18 month gap)
4. Original coding

From this, we intended to make two comparisons:

• Compare Naive coders with Original coding
  Coder reliability
• Compare Naive coder with Expert coder
  Coder stability

Two methods of analysis were used in the interpretation of the data: the kappa statistic and pairwise analysis.
A.2.4.1 Kappa

The kappa statistic is a measure of agreement, which takes into account both the number and proportion of categories, and chance agreement. Measurements like pairwise analysis (discussed below) do not adjust for either of these aspects, and therefore their measures of reliability are less easy to interpret.

"The kappa coefficient of agreement is the ratio of the proportion of times that the raters agree (corrected for chance agreement) to the maximum proportion of times that the raters could agree (corrected for chance agreement)."

Siegel and Castellan (1988:265)

The equation for kappa is:

\[ K = \frac{P(A) - P(E)}{1 - P(E)} \]

where \( P(A) \) is the proportion of times that the coders agree (roughly equivalent to the pairwise analysis described below), and \( P(E) \) is the number of times that the coders would be expected to agree by chance. \( K \) ranges from \( K = 0 \) (equivalent to chance) to \( K = 1 \) (perfect agreement).

However, kappa only measures agreement between coders. This is fine if it is only the level of agreement between the coders in which you are interested. However, if you wish to see how often a set of coders match a particular coding, that is, can they replicate a coding which has already been produced, this is more problematic. You do not want to include coder agreements which conflict with your original codings: this would give a higher value to kappa than is justified by your results.

In our calculation of kappa, therefore, agreement which didn’t match the original coding was ignored. For example, if two coders agreed on +NEW-QUESTION, the other coder used +QUERY, and the original code was +QUERY, then the level of agreement on this datapoint was considered to be one rather than two.

**Interpreting Kappa** There is some disagreement in the interpretation of Kappa. Siegel and Castellan (1988) look for statistical significance in the value of \( K \): is it greater than that would be expected by chance? Carletta *et al.* (in preparation)
(after Krippendorff 1980) argues that it is better to interpret the scale of agreement, as the acceptable value of $K$ varies according to the type of data involved. Krippendorff (1980) suggests a reasonable 'rule of thumb' for associations which rely on subjective distinctions is to require $K > 0.8$ with $0.67 < K < 0.8$ allowing tentative conclusions to be drawn.

We will be using these recommendations for the analysis of our data.

<table>
<thead>
<tr>
<th>Value of $K$</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K &lt; 0.67$</td>
<td>No reliable agreement</td>
</tr>
<tr>
<td>$0.67 &lt; K &lt; 0.8$</td>
<td>Tentative agreement</td>
</tr>
<tr>
<td>$K &gt; 0.8$</td>
<td>Reliable agreement</td>
</tr>
</tbody>
</table>

A.2.4.2 Pairwise Analysis

Pairwise analysis compares agreement between each pair of coders. It was used by Pitrelli et al. (1994) in order to analyse a replicability study for the ToBI system of intonational analysis. For each datapoint, the codes of each coder are compared against the codes of every other transcriber, rather than comparing the codes of individual coders against the group.

The number of pairwise comparisons for $k$ coders is:

$$\sum_{i=1}^{n-1} i = \frac{n \times (n - 1)}{2}$$

Therefore, for 5 coders there are 10 possible pairings:

A B, A C, A D, A E = $(n - 1)^+$
B C, B D, B E = $(n - 2)^+$
C D, C E, = $(n - 3)^+$
D E = $(n - 4) = 10$

For example, if four out of five coders code a move as +NEW-QUESTION and the last coder uses +QUERY instead, then the level of agreement is considered to be 60% (6 pairs of agreement out of a possible 10 pairs) rather than 80% (4 out of 5 labelled correctly). This because 6 pairs of coders agreed on this analysis. Pitrelli et al. (1994) claims that this is a more stringent mode of analysis because it is concentrating on inter-coder agreement rather than a right/wrong distinction.
This metric is calculated by working out the percentage agreement for each datapoint (how many pairs agreed on each potential coding point), summing these percentages, and dividing them by the total number of datapoints (N). This gives a result reflecting the average pairwise agreement over the dataset.

Disadvantages with this analysis  Pairwise analysis has been criticised for not allowing for either chance agreement or for the distribution of coding choices (Carletta, 1996a). If two coders labelled utterances randomly using two codes, we would expect them to agree 50% of the time. When analysed critically, a result of 70% agreement in such circumstances would be less than impressive, but on the surface, it would look reasonable. If there are two exclusive categories, and one is expected to be used 80% of the time, then this choice, too, should be easier to make correctly. Therefore, the results of pairwise analysis should be interpreted carefully.

Why use pairwise analysis?  The kappa statistic, explained above, has many advantages over pairwise analysis in that it accounts for both chance agreement and the distribution of coding choices. However, it is not suitable for all types of data. Kappa works best when it is simple to locate when a choice has to be made, and what the set of choices is. For example, it is entirely appropriate to use kappa for analysing the dataset of RESPONSE moves. We know exactly which moves it applies to, and what choices (REPLY-MIN, REPLY-YN, REPLY-FULL). However, for testing the replicability of -FEATURE-LOC, which could occur in any move type at any time, such an approach is impossible. It is in these cases that we have chosen to use pairwise analysis.

A.2.5  Organisation of Tests

In this section we will briefly outline the tests which we applied to the replication data. As we have described previously, not all our data lends itself easily to such tests, so we will explain the reasoning behind each test.

We will describe the set of tests to be analysed using kappa first, followed by the pairwise analysis tests.
Tests employing kappa

1. INITIATE moves
   This tests the ability of the coders to choose correctly from the subset of codes available for INITIATE moves. These codes are:
   +NEW-QUESTION, +NEW-SUGGESTION, +RELEVANT-INFO, +CHECK, +QUERY, +OBJECTION
   This used all the coded INITIATE moves in the dataset.

2. RESPONSE moves
   Similarly to test 1, this tests the ability to choose correctly between the available subset of response moves:
   +REPLY-MIN, +REPLY-YN, +REPLY-FULL
   This used all the coded RESPONSE moves in the dataset.

3. FOLLOW-UP moves
   As above, this tests the accurate coding of FOLLOW-UP moves from the subset available:
   +ACK-SHORT, +ACK-REP, +ACK-FULL
   This used all the coded FOLLOW-UP moves in the dataset.

4. +INFO-INTEG
   This tested the use of +INFO-INTEG. This code is only available in conjunction with either +ACK-FULL or +REPLY-FULL. It is therefore a two-stage coding process: the coders must have correctly coded the move as either +ACK-FULL or +REPLY-FULL in order to consider the appropriateness of the +INFO-INTEG coding. In terms of the kappa statistic, this process is reflected in the selection of the dataset. Only those moves which all coders have coded as either +ACK-FULL or +REPLY-FULL are used. This is because it would be meaningless to make a comparison on coding when a coder has not achieved the first stage: they are bound to get the second stage wrong!

5. +KNOWLEDGE-MISMATCH
   This tested the use of +KNOWLEDGE-MISMATCH. This code is only available in conjunction with either +QUERY or +OBJECTION. As above, the coding reliability was tested on all those INITIATE moves which all coders had coded as either +QUERY or +OBJECTION.
6. **-REPLY-FULL**
   The -REPLY-FULL code is only used when a RESPONSE move has been chosen, but it is considered inadequate in that situation. Clearly, the coder has to make this decision of adequacy on every RESPONSE move. So, here we test the reliability of -REPLY-FULL on all RESPONSE moves.

7. **-ACK-FULL**
   Like -REPLY-FULL, -ACK-FULL is only used for an inadequate FOLLOW-UP move. Therefore the same type of test is performed as above. The coders are tested in their ability to accurately code -ACK-FULL on all FOLLOW-UP moves.

**Tests employing pairwise analysis**

The following codes are tested:

- +FEATURE-INTRO, -FEATURE-INTRO
- -RELEVANT-INFO
- -CHECK
- -REPLY-MIN
- -ACK-SHORT
- -FEATURE-LOC
- +FEATURE-UNIQUE
- -QUERY
- -KNOWLEDGE-MISMATCH

We are unable to test these using kappa, because we cannot constrain the dialogue environment in which they occur. All of these codes could apply to any move type, and therefore we would have to test on the whole dataset. This would skew any kappa result, because you would have many agreements of no-coding which would be meaningless: there would be no question of using that particular code in that particular situation. Therefore, the most sensible mode of analysis open to us is that of pairwise analysis.
8. +FEATURE-INTRO, -FEATURE-INTRO
   For any feature mentioned in a conversation, the coder has to decide whether it is being introduced (first mention), and if so, whether it has been a highlighted introduction (+FEATURE-INTRO) or a non-highlighted introduction. Therefore three coding possibilities are available:
   +FEATURE-INTRO, -FEATURE-INTRO, no-code
   A pairwise analysis was performed on this three-way distinction.

9. -RELEVANT-INFO
   A pairwise analysis was made on the two-way distinction:
   -RELEVANT-INFO, no-code

10. -CHECK
    A pairwise analysis was made on the two-way distinction:
    -CHECK, no-code

11. -REPLY-MIN
    A pairwise analysis was made on the two-way distinction:
    -REPLY-MIN, no-code

12. -ACK-SHORT
    A pairwise analysis was made on the two-way distinction:
    -ACK-SHORT, no-code

13. -FEATURE-LOC
    A pairwise analysis was made on the two-way distinction:
    -FEATURE-LOC, no-code

14. -FEATURE-UNIQUE
    A pairwise analysis was made on the two-way distinction:
    -FEATURE-UNIQUE, no-code

15. -QUERY
    A pairwise analysis was made on the two-way distinction:
    -QUERY, no-code

16. -KNOWLEDGE-MISMATCH
    A pairwise analysis was made on the two-way distinction:
    -KNOWLEDGE-MISMATCH, no-code
A.3 Results

In this section we will report the results which we found. We are reporting only the comparisons between naive coders and the expert coder. This is because there was no real difference between the expert coder re-coding the data (after 18 months) and the original coding. The kappa values and percentage pairwise agreements were extremely similar, and had no impact upon levels of reliability.

In the first section, we will report the results from 1-7 (Kappa results), and the second section will report the results from 8-16 (pairwise analysis results).

A.3.1 Kappa results

The term dataset refers to the number of agreements which we are testing: how many coding comparisons were made? Note this is a reflection of the number of coding opportunities available, given the context a particular code may occur in. The dataset therefore includes occasions where there are no codes; in such cases we are assessing the agreement between coders that no code was required.

The term datapoint refers to the number of instances that coding occurred in the dataset.

For example:

For test 1, there are 153 initiate moves in the dataset. For test 4, there are 49 points where a INFO-INTEG could have occurred, but only 7 actually did occur.

<table>
<thead>
<tr>
<th>Test</th>
<th>Code tested</th>
<th>Dataset (N)</th>
<th>Data-points</th>
<th>Kappa value</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>INITIATES</td>
<td>153</td>
<td>n/a</td>
<td>0.84</td>
<td>Reliable</td>
</tr>
<tr>
<td>2</td>
<td>RESPONSES</td>
<td>109</td>
<td>n/a</td>
<td>0.85</td>
<td>Reliable</td>
</tr>
<tr>
<td>3</td>
<td>FOLLOW-UPS</td>
<td>71</td>
<td>n/a</td>
<td>0.76</td>
<td>Tentative</td>
</tr>
<tr>
<td>4</td>
<td>+INFO-INTEG</td>
<td>49</td>
<td>7</td>
<td>0.69</td>
<td>Tentative</td>
</tr>
<tr>
<td>5</td>
<td>+KNOWLEDGE-MISMATCH</td>
<td>30</td>
<td>3</td>
<td>0.76</td>
<td>Tentative</td>
</tr>
<tr>
<td>6</td>
<td>-REPLY-FULL</td>
<td>109</td>
<td>19</td>
<td>0.86</td>
<td>Reliable</td>
</tr>
<tr>
<td>7</td>
<td>-ACK-FULL</td>
<td>71</td>
<td>4</td>
<td>0.72</td>
<td>Tentative</td>
</tr>
</tbody>
</table>

In this first set of tests, those results which do not reach $K = 0.8$ are mainly tests where very small numbers of datapoints are available. This applies particularly to the tests on +KNOWLEDGE-MISMATCH (3 instances) and -ACK-FULL (4 instances). It is difficult to gain a high degree of reliability on such small numbers, because one error by a coder has a large impact on the result, and may not be indicative of overall unreliability.
For the test on +INFO-INTEG we would suggest there is a different type of problem. In a typical dialogue, this coding occurs very rarely. The coders would have been aware of this fact, and would have seen very few instances in their practice dialogue. Consequently, it is not surprising that they have failed to recognise as many of these codings as they should: they would not expect to find that many in two dialogues. This is an example of the problem of atypical dialogues which we discuss below.

The poor result for FOLLOW-UPS reflects one coder’s difficulty in distinguishing +ACK-SHORT and +ACK-FULL. We believe this could be remedied with further training.

### A.3.2 Pairwise analysis results

For the pairwise analysis results, we have taken 80% as an acceptable level of agreement. We believe that this allows sufficiently for the degree of chance agreement.

<table>
<thead>
<tr>
<th>Test</th>
<th>Code tested</th>
<th>Datapoints</th>
<th>Agreement</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>+FEATURE-INTRO, -FEATURE-INTRO</td>
<td>23</td>
<td>85.9%</td>
<td>Good</td>
</tr>
<tr>
<td>9</td>
<td>-RELEVANT-INFO</td>
<td>10</td>
<td>80%</td>
<td>Good</td>
</tr>
<tr>
<td>10</td>
<td>-CHECK</td>
<td>8</td>
<td>83.3%</td>
<td>Good</td>
</tr>
<tr>
<td>11</td>
<td>-REPLY-MIN</td>
<td>5</td>
<td>66.7%</td>
<td>Poor</td>
</tr>
<tr>
<td>12</td>
<td>-ACK-SHORT</td>
<td>7</td>
<td>81%</td>
<td>Good</td>
</tr>
<tr>
<td>13</td>
<td>-FEATURE-LOC</td>
<td>11</td>
<td>81.8%</td>
<td>Good</td>
</tr>
<tr>
<td>14</td>
<td>+FEATURE-UNIQUE</td>
<td>1</td>
<td>100%</td>
<td>?</td>
</tr>
<tr>
<td>15</td>
<td>-QUERY &amp;</td>
<td>1</td>
<td>100%</td>
<td>?</td>
</tr>
<tr>
<td>16</td>
<td>-KNOWLEDGE-MISMATCH</td>
<td>1</td>
<td>100%</td>
<td>?</td>
</tr>
</tbody>
</table>

In this second set of tests, the only one which could not be considered reliable was -REPLY-MIN. Again, the numbers were very small: only five instances in the dataset. There was one instance of this coding where 2 out of 3 coders disagreed with the original coding, and this decreased the reliability percentage considerably. If that datapoint is excluded, then the reliability would be 83.3%, which would be considered acceptable.

The ratings for +FEATURE-UNIQUE and +QUERY & +KNOWLEDGE-MISMATCH are also questionable. there is only one example of each of these, and this is scarcely sufficient to base a reliability rating on. We have reported the percentage agreement which we have found, but have not reported a reliability level.
Codings not tested:
It was not possible to test all the codings we discussed in Chapter Five. Some codings occur very infrequently, and it was not practical to use data that had examples of these codings. These are the codings which were omitted:

- \texttt{+FEATURE-LOC}
- \texttt{-PARTNER-BELIEF} [extremely rare - excluded from study]
- \texttt{-INFO-INTEG} [extremely rare]

As these codings are so infrequent, we do not believe that our not replicating their usage affects the reliability of our results. This is because firstly, they have little impact on the overall results, and secondly, neither of our coders did use either \texttt{+FEATURE-LOC} or \texttt{-INFO-INTEG}, even though these codes were explained, and were available to them. That in itself must demonstrate some degree of reliability.

A.4 Discussion

In this section we will briefly evaluate the results which they have found, and consider whether they constitute support for the replicability of the coding system.

Although the results that we have found are not always strictly reliable, we would argue that given the constraints of this type of study, they are sufficient to support the statistical results reported in Chapter Six.

Experimental design

The basic design of the experiment was satisfactory, but it was very difficult to provide sufficient examples of all the coding categories, as our results above demonstrate. Obviously, we did try to choose dialogues which provided a broad coverage of coding types, but this in itself is problematic. As we have suggested above, there is a risk of using atypical dialogues in order to gain a wide spread of codings, especially where negative coding is concerned. We did not want to base a reliability study on data which was not fairly typical of most of our data. This would make little sense, and certainly would not show that our coding was, in general, replicable.
Other studies have avoided such problems by using dialogue extracts: you can then target your excerpts appropriately to gain the breadth of coverage, without being tied to a whole dialogue. However, because our coding system is very reliant on the discourse context, we could not utilise this method either. For the coding system to operate properly, the coders were required to code whole dialogues.

Ideally, the experiment would have included another practice stage in order to remediate any misunderstandings, and would also have involved coding more test dialogues, in order to increase the size and coverage of the dataset. This was not practical for this project. Such an extension would have meant a much greater investment of time from the coders: coding dialogues in this fashion is extremely time-consuming\(^2\). This was not possible under the circumstances.

**Evaluation of results**

Although the results we have found are not always strictly reliable in terms of the kappa statistic or pairwise analysis, we believe that these results are sufficient to justify the statistical analysis which is described in Chapter Six. It is notoriously difficult to achieve good results on such subjective judgements, and the levels we have achieved, we would consider acceptable.

In addition, we would argue that considering the size and complexity of the coding system, and the small amount of training time, the degree of reliability that was achieved is impressive. With a more intensive training programme, we believe that all the categories would be shown to be replicable.

\(^2\)And greatly tests the goodwill of friends!
Appendix B

Instructions for Replication Study

B.1 Background

This study is intended to investigate whether the coding system described in the *Brief Guide to the Typology of Dialogue Attribute Types*\(^1\) can be replicated by naive coders. You will be asked to code four dialogues, two for practice, and the remaining two as the data on which the study will be based. You will be provided with explanatory material, alongside the dialogues and associated maps.

The aim of this document is to provide an outline of the study, and to explain the process of coding a dialogue in more detail.

B.2 Procedure

The study will have two parts, firstly, we will introduce you to the coding scheme using some example dialogues, and secondly, you will be asked to code a set of test dialogues.

1. I will meet with all of you, and hand out this document, the test data, the examples and the other explanatory materials

   - You will be expected to code the test data during this period
   - You should contact me if there are any difficulties
   - You should not discuss the coding with the other experimentees

\(^1\)This contains the same information as Chapter Five in this thesis. It was presented in a slightly different form to the experimental subjects.
We will set the time for the individual meetings to follow

2. I will meet with individuals, discuss any problems with the test data, and hand out the experimental data

- You will be expected to code the experimental data, and return it by an agreed date
- You should contact me if there are any difficulties
- You should not discuss the coding with the other experimentees

B.3 Materials

You will receive the following items:

- This set of Instructions
- Brief Guide to the Typology of Dialogue Attributes
- Two coded dialogues, as an example
- Two dialogues and maps, as the test set
- Two dialogues and maps, as the data set

B.4 The Coding Process

In this section we aim to outline the steps you should take when coding an utterance. First we explain some of the terminology, and then we will move onto the coding process.

B.4.1 Coding

By coding we mean ‘flagging’ or ‘labelling’ particular features in utterances. We want to indicate all instances of a set of features so that we can look at their distribution, and see how their usage interacts with other aspects of the dialogue. Coding can be used at all levels of language: phonetics, syntax, referring expressions, and so forth. Here, we are interested in the choices speakers make in the way in which they tailor their utterances to the needs of their partner. This could
be in terms of the presentation of information, or how they reply to a question, or acknowledge some information offered by their partner.

Codes are in the form of short labels, which represent the meaning of that particular category. They are presented in SMALL CAPS to distinguish them from the rest of the text. One such code is QUERY which refers to the usage of what might be termed check questions.

Short definitions of each of these categories can be found in this document. Longer explanations and examples can be found in the Brief Guide.

These codes can either be positive or negative, a distinction which we will explain in the next section.

B.4.2 Positive and Negative Coding

In coding dialogues we are equally interested in what people do and don’t do. Consequently, we want to code utterances for both occurrences of features (positive), and failure to employ these features where appropriate (negative). A good example is that of FEATURE-INTRO, this category is concerned with how the interactant introduces a new landmark into the dialogue.

- A positive coding of this dialogue attribute type would occur when the new item is highlighted in the utterance, e.g.:
  1. “Have you got a burnt cottage?”
  2. “Do you see the burnt cottage?”

- A negative coding would occur when a new item is introduced, but is not made the focus of that utterance. e.g.:
  1. “Go past the burnt cottage.”
  2. “Go down as far as the burnt cottage.”

- No coding for this dialogue attribute type will occur, if no appropriate new item is introduced in an utterance.

We are interested in what aspects of speaker behaviour affect task success. Identifying the (potentially) positive and negative features of a dialogue allows us to see how these features interact with performance in the task. We are particularly interested in both the use of the more effortful attribute types (+NEW-
SUGGESTION, +NEW-QUESTION, +RELEVANT-INFO), and the points where speakers fail to use appropriate attribute types (all negative coding).

B.4.3 Coding Utterances

Here we will simulate the process of coding an utterance. Basically, the coder asks two simple questions about an utterance:

1. What has the speaker done?
2. What has the speaker failed to do?

B.4.3.1 What the Speaker has Done

The first step to take is to decide the supertype of move which the speaker has made. In fact, for this study, this decision has already been taken for you. The first coding line following the utterance will be labelled INITIATE, RESPONSE, FOLLOW-UP (this is our superset), or INSTRUCT (an important subtype of INITIATE). Your first task is to decide what subtype of that particular supertype this utterance constitutes. For example, INITIATE moves can be INSTRUCT (these are labelled for you), give information or solicit information moves, RESPONSE and FOLLOW-UP moves vary according to the amount of information or feedback they contain.

A full list of these move-specific dialogue attribute types can be found in Sections B.5.1 to B.5.4 of this document.

Secondly, there are several more specialised dialogue attribute types which you may also have to consider. These concern either the treatment of landmarks (as these are particularly important in the context of these dialogues), or speaker beliefs when asking questions (particularly those coded as +QUERY or +OBJEC-

B.4.3.2 What the Speaker has Failed to Do

Once you have coded what the utterance attempts to do, the next step is to consider whether the supertype or subtype of move chosen was appropriate for the dialogue context.
1. The first question is whether the *subtype* of move is appropriate: it is a *RESPONSE* move, but is it suitable for the dialogue context?

2. The second question is whether the *supertype* of move is appropriate: was a *FOLLOW-UP* move required here, before moving on to the next exchange?  

Section B.5 suggests alternate subtype and supertype choices for each move supertype (INITIATE, RESPONSE, FOLLOW-UP). It also covers the more specialised categories which are concerned with the treatment of entities, or speaker beliefs. It is suggested that you should use these possible alternatives as a checklist, in order to make the coding process easier.

The following list is just intended as a summary of the sort of behaviours which you are likely to encounter (and what coding possibilities there are).

Speakers can fail to (move-specific):

- Give instructions appropriately
  -NEW-FEATURE, -ROUTE-SECTION, -ROUTE-MARKER, -SIGNPOSTS
- Ask questions/raise objections
  -QUERY, -NEW-QUESTION, -OBJECTION
- Use checking routines
  -CHECK-UNDERSTANDING, -CHECK-COMPLETION, -CHECK-ALIGN
- Introduce information/make suggestions
  -RELEVANT-INFO, -NEW-SUGGESTION
- Make a reply
  -REPLY-MIN
- Make an appropriate reply
  +REPLY-MIN -REPLY-FULL, -INFO-INTEG
- Make a follow-up
  -ACK-SHORT
- Make an appropriate follow-up
  -ACK-FULL, -INFO-INTEG

---

3In this case, the negative coding will be of the appropriate subtype, rather than the supertype, for example, -ACK-SHORT rather than -FOLLOW-UP
3This acts as one unit of coding - see the *Brief Guide* for an explanation of this.
Speakers can fail to (entity specific):

- Unambiguously locate landmarks in terms of Given/New or physical position on the map
  -FEATURE-INTRO, -FEATURE-LOC, -FEATURE-UNIQUE

Speakers can fail to (belief-specific):

- Question/object to knowledge mismatches
  -KNOWLEDGE-MISMATCH

**B.5 Introduction to The Typology**

The aim of this section is to outline the process of coding an utterance according to the Typology described in the *Brief Guide*. This is not intended as a substitute for the descriptions and definitions of each of the dialogue attribute types given in the Typology, but rather a method of limiting the subset of possibilities which should be considered for any given utterance.

This document covers the dialogue attribute types in much the same order as the fuller description, covering the general groups of INITIATE, RESPONSE and FOLLOW-UP moves, and then moving on to the two non-move specific categories: Entity Specific and Higher Level attribute types.

This guide should be used by checking the section appropriate for the move-type of your utterance (I, R, F), and then checking the remaining sections, which might also apply. The dialogue attribute type labels in SMALL CAPS are the same as those used in the full description.

Remember that this is only intended to be an outline guide, and you should check the detail of attribute type definitions to ensure the right decisions are made.

**B.5.1 INITIATE Moves**

In this section we will cover all types of INITIATE moves: INSTRUCTS, checks and other game-type initiators.
B.5.1.1 What Move Type is It?

In this section we wish to determine which strategy conditions your current utterance satisfies.

- **INSTRUCT moves**
  See Section B.5.1.2

- Initiator for routine checking:
  (Checking other person’s understanding, of information offered by current speaker.)
  1. +CHECK-UNDERSTANDING
  2. +CHECK-COMPLETION
  3. +CHECK-UNDERSTANDING +CHECK-COMPLETION
  4. +CHECK-ALIGN

- Initiator for check questions and queries:
  (Checking information offered by other speaker.)
  1. +QUERY
     Syntactically a question
  2. +OBJECTION
     Syntactically a statement

- Initiators of other game types:
  (See full document for explanations of these, and restrictions on their usage.)
  1. +NEW-SUGGESTION
  2. +NEW-QUESTION
  3. +RELEVANT-INFO

B.5.1.2 INSTRUCT Moves

Because of the importance of instructions in a task-oriented dialogue, we have produced a list of dialogue attribute types which apply solely to this type of move. Note that due to the nature of the task, only GIVERS can give instructions: FOLLOWERS do not have the requisite knowledge (of the route) to do so. Any seeming instruction a Follower makes should be evaluated in terms of a *suggestion* (+NEW-SUGGESTION).
• In terms of number of new features used:
  1. No more than one - +NEW-FEATURE
  2. No features used in description - no coding appropriate
  3. More than one - -NEW-FEATURE
     (Should consider SIGNPOST coding at this point.)

• In terms of amount of route covered:
  (See route sections as listed for each separate map - in another document)
  1. Less or equal to a route segment - +ROUTE-SECTION
  2. Not relevant to this instruction - no coding
  3. More than one route segment - -ROUTE-SECTION
     (Should consider SIGNPOST coding at this point.)

• In terms of information about the route section given in the route section, for example:
  
  - Start point for route section
  - End point for route section
  - Relationship of route to adjacent landmark - e.g. which side

Which should be coded as:

  1. Does not omit an obvious route delimiters (as above) - +ROUTE-MARKER
  2. Not relevant to this instruction - no coding
  3. Omits obvious route delimiters (as above) - -ROUTE-MARKER

NB Note that you cannot have both positive and negative coding for this strategy. If there is one omission, then the utterance should be coded as -ROUTE-MARKER

• In terms of information organisation within an utterance
  NOTE: only applies to those utterances which have already been coded as -NEW FEATURE and/or -ROUTE-SECTION:

  - Uses discourse markers to structure information
  - Uses route pivots (e.g. fixed items on the map - landmarks, parts of the page, etc) to make the instruction more explicit
Which should be coded as:

1. Uses route pivots and markers - +SIGNPOSTS
2. Is not assessed as having failed to use pivots and markers - i.e. utterance is probably not long enough for this category to apply. (no coding)
3. Omits route pivots and/or markers - -SIGNPOSTS

**B.5.2 Is There a More Appropriate Move?**

Now that we have established the type of INITIATE move the current utterance constitutes, we can consider what type of move (INITIATE or otherwise) it might have been.

**B.5.2.1 Alternate INITIATE Moves**

Here we consider which other INITIATE moves may have been more appropriate for the current speaker to employ at this point in the dialogue.

1. Is there an obvious problem not addressed by the current move?
   - If yes then: -QUERY -KNOWLEDGE-MISMATCH
   - If no then: no coding

2. Is there some (potentially) relevant information which should have been introduced at this point? (e.g. feature on Follower’s map)
   - If yes then: -RELEVANT-INFO
   - If no then: no coding
3. Has the dialogue just moved on from a completed route section or problematic route-part *without* checking that it has been achieved successfully?

- If **yes** then consider the following possible codings:
  
  (a) Most likely: general checking required:
      - CHECK-UNDERSTANDING - CHECK-COMPLETION
  
  (b) Infrequent: specific alignment checking required:
      - CHECK-ALIGN

- If **no** then: no coding

**B.5.2.2 Alternate Move Types**

In this section, we consider whether a different *type* of move would have been more appropriate at this point.

1. Should the current move have been a RESPONSE or FOLLOW-UP to the participant’s previous turn?
   i.e. Is there a RESPONSE or FOLLOW-UP move missing?

   Note: this does not apply to QUERY or OBJECTION because although they are not technically RESPONSE or FOLLOW-UP moves, they constitute an *appropriate reaction* to the previous move, and thus override the need for a RESPONSE or FOLLOW-UP move at this point.

   If you judge there to be a missing RESPONSE or FOLLOW-UP move, then:

   - If a RESPONSE move is missing: -REPLY-MIN
   - If a FOLLOW-UP move is missing: -ACK-SHORT

**B.5.3 RESPONSE Moves**

The process for coding RESPONSE moves is somewhat different to that for coding INITIATE moves, this is because there does not appear to be as much *variation* within the general move-type. INITIATE moves can begin a variety of different tasks, RESPONSE moves, by their nature, are responding to a goal set by another person, so the field of replying is already constrained. The RESPONSE moves identified appear only to differ in terms of the *fullness* of information, and how that information relates to the other speaker’s goals. Therefore, the coding in this
section is split into just two categories: The coding of the nature of the RESPONSE move (Section B.5.3.1), and the consideration of other more suitable moves which could have occurred at this point (Section B.5.3.2).

B.5.3.1 Coding a RESPONSE Move

In the list below, we give an outline of the possible RESPONSE move codings (both positive and negative), and where they should be used. For a fuller definition of each of these categories, please refer to the Typology Description. These codings cover the appropriateness (or otherwise) of the RESPONSE move used. Please see the next section for the coding of substitutable moves.

A note on the use of negative coding is appended after this list.

No Response where one would be expected

• -REPLY-MIN
  No response provided where one would be expected

Response insufficient:

• +REPLY-MIN -REPLY-FULL
  Inappropriately brief (or otherwise problematic) response

Response sufficient:

• +REPLY-YN
  Yes-No response to Yes-No question

• (+REPLY-MIN) +REPLY-FULL
  Response to WH-question

Response fuller than that directly solicited:

• (+REPLY-MIN) +REPLY-FULL
  Fuller response to Yes-No question

• (+REPLY-MIN) +REPLY-FULL +INFO-INTEG
  Full response, plus explicit use of information offered

227
Note that the coding +REPLY-MIN is bracketed when it co-occurs with +REPLY-FULL. This is because this coding is redundant here as the category +REPLY-FULL presupposes satisfaction of the conditions for +REPLY-MIN. For logical reasons, we have shown it in the example coding here to represent the progression through the categories, but when coding dialogues, the coding in brackets should be omitted.

For the response categories(REPLY-MIN, REPLY-YN, REPLY-FULL), negative coding will only occur on its own when not even a minimal response has been made, and that will be coded as -REPLY-MIN. If a minimal response has been made, and a fuller response would have been more appropriate, then the utterance should be coded as +REPLY-MIN -REPLY-FULL. These are the only combinations for negative coding that are currently in use.

**B.5.3.2 Alternate Move Type**

The main choice which speakers appear to demonstrate at this point is between one of the types of RESPONSE move (as outlined above) or a query of some description. Therefore, the main substitutable move to check for here is the check question.

1. Is there an obvious problem which has not challenged?
   For example:
   - Use of an unshared, unnegotiated landmark in an instruction
   - An instruction incompatible with the speaker's position on the map
   - An instruction which would cause the Follower to draw through a feature

   If YES then: -QUERY -KNOWLEDGE-MISMATCH

**B.5.4 FOLLOW-UP Moves**

Like RESPONSE moves, FOLLOW-UP moves the range of possible moves is constrained by the previous move. Again, the possible coding combinations are on a gradient, moving from brief to full.
B.5.4.1 Coding a FOLLOW-UP Move

Here we give an outline of the possible positive and negative codings, and the type of FOLLOW-UP move that they would be used to describe.

There are a number of similarities between the way in the RESPONSE and FOLLOW-UP moves are coded, but note that here a minimal acknowledgement (+ACK-SHORT) does not immediately imply that the utterance is problematic. Here, utterances can be appropriately or inappropriately short, and should be coded accordingly. See the main document for an explanation of why this is judged to be the case, and a definition of what is deemed to be suitably, or unsuitably, brief.

An outline of codings:

No Follow-up where one would be expected:

- +ACK-SHORT
  
  No acknowledgement provided where one would be expected

Follow-up insufficient:

- +ACK-SHORT -ACK-FULL
  
  Inappropriately brief acknowledgement

Follow-up sufficient:

- +ACK-SHORT
  
  Appropriately brief acknowledgement

- (+ACK-SHORT) +ACK-FULL
  
  Fuller acknowledge move

Follow-up fuller than that directly solicited:

- (+ACK-SHORT) +ACK-FULL +INFO-INTEG
  
  Full acknowledge, plus explicit use of information offered

Note that as with the RESPONSE coding in Section B.5.3.1, +ACK-SHORT coding is presupposed by +ACK-FULL coding, and is bracketted in the examples given above. This coding is included here for logical completion, but should be omitted when coding, as it is essentially redundant.
B.5.4.2 Alternative Move Types

Again, like RESPONSE moves, the choice between move-types seems to be limited to one of the FOLLOW-UP moves outlined above, or a query of some description. Therefore, the main substitutable move to check for here is the check question.

Check questions occur less frequently at this position in the exchange (as compared to the RESPONSE move position). This is presumably because such checks usually occur after an information-bearing move, and INITIATE moves carry much of the weight of information transferral in this type of dialogue.

However, the question of whether a check-question should have been employed should still be examined.

1. Is there an obvious problem which has not challenged?
   For example:
   
   • Information which conflicts with the current speaker’s beliefs

   If YES then: -QUERY -KNOWLEDGE-MISMATCH

B.5.5 Entity-Specific Moves

The aim of this group of attribute types is to code how the utterance deals with giving information about a feature. Landmarks are the single most significant source of information for the route, and therefore it is important for the Typology to take account of the way in which features are introduced and discussed.

The questions given here below should be checked for any utterance which is concerned with giving information about a feature. The set of codings with which the questions are concerned are not move-specific, although in practice +/-FEATURE-INTRO coding tends only to apply to INITIATE moves, but that is because of its definition, rather than a desire to restrict it to that move-type.

1. Does the utterance introduce/re-introduce a feature using a question or otherwise highlighted construction?
   NB This introduction should be unsolicited.
   
   • If yes then +FEATURE-INTRO
   • If not relevant (i.e. no new feature) then: no coding
   • If no then -FEATURE-INTRO

230
2. Does the utterance negotiate the position of an unshared feature?
   NB This negotiation should be unsolicited.
   • If yes then +FEATURE-LOC

3. Does the utterance fail to negotiate the position of an unshared feature?
   E.g. Previous speaker introduces unshared feature, current speaker fails to mutually establish it’s location (or at least start this process).
   • If yes then -FEATURE-LOC

4. Does the utterance use a negotiated unshared feature in a route description?
   • If yes then +FEATURE-LOC

5. Does the utterance use an unnegotiated unshared feature in a route description?
   NB The feature should have been established as unshared.
   • If yes then -FEATURE-LOC

6. Does the utterance refer to a feature in a unique fashion?
   (This means referring to a landmark in such a way that it might suggest your cottage and my cottage may not be the same thing.)
   • If yes then +FEATURE-UNIQUE
   • If not relevant then: no coding
   • If no then
     – Has it caused a problem?
       * If yes then -FEATURE-UNIQUE
       * If no then has there been a -FEATURE-UNIQUE coding previously in the dialogue?
         • If yes then -FEATURE-UNIQUE
         • If no then: no coding

It will be noticed that the coding system for +FEATURE UNIQUE is somewhat different to the approach taken for the rest of the coding scheme, as negative coding is only ‘triggered’ when a problem occurs in the dialogue, rather than
always being coded for. This is because of our reasoning in coding this particular strategy.

One of the aims of coding dialogues in this sort of detail was to investigate how speakers strategies change as they adapt (or don’t adapt) to the needs of this particular task. Although the coding scheme does give a broad-brush view of these changes, it cannot pinpoint the moment at which a speaker chooses to change strategy, or the precise reason why they should change approach at that point. We can only suggest where these changeover points might be, and what caused them to change strategy.

However, the duplicate features on the map task provide an opportunity to investigate one particular strategy change. When doing the Map Task, speakers make assumptions about what the maps share, in terms of all the features being shared, or all the features being in the same place on the two maps, or there not being any duplicate features. Most of the time, speakers can safely assume that there are no duplicates, and that a reference to burnt forest will uniquely pick out that feature on the map. However, from time to time, a misunderstanding will be caused by the one duplicate feature on each set of maps. We are interested to see whether the discovery of a problem obviously caused by a lack of unique reference will cause the speakers to refer to new landmarks in a unique fashion. In other words, will this problem now make them check that any landmarks which have the same name are in fact in the same place?

We only code -FEATURE-UNIQUE when either a problem has been caused or such a problem has already occurred in this dialogue. Positive coding occurs as you would expect, throughout the dialogue. It may not seem logically coherent to only trigger the negative coding under certain circumstances, but we can justify this approach in two respects.

Firstly, virtually every mention of a feature in each of the dialogues would have to be coded as -FEATURE-UNIQUE, and this would both increase the amount of coding substantially, and would be inherently redundant. By default, if a feature mention is not coded as +FEATURE-UNIQUE then it is -FEATURE-UNIQUE - it is an either/or coding. By using other codings available on the Map Task database, it is straightforward to work out how many times any feature was mentioned, and

---

4Each map has one pair of identical features - one of which is relevant to the route, one of which is irrelevant. The Giver has both these features, The Follower only the irrelevant one. Depending on how close the irrelevant feature is to the route as a whole (this varies) it is possible for the speakers to confuse these two landmarks, and thus cause an error in the route, which may or may not be discovered.
thus what proportion were +FEATURE-UNIQUE. Therefore it is not necessary for us to use this extra coding, as this part of the information is already available through analysis carried out on our data.

Secondly, it should also be pointed out that such confusions caused by duplicate features only occur in a subset of the data used in this research. It is only worth our while to invest in the extra coding that this category requires when it may contribute to studying the phenomenon of strategy change.

B.5.6 Higher Level Attribute Types

This section is concerned with the coding of one attribute type: KNOWLEDGE-MISMATCH. This is coded *alongside* any other codings which have been used for the utterance⁵.

B.5.6.1 Highlighting Information Mismatches?

1. Has the utterance been coded as +QUERY, +OBJECTION?
   - If yes then does the utterance *highlight* information mismatches caused by the other interactant?
     - If yes then +KNOWLEDGE-MISMATCH
     - If no then: no coding

2. Has the utterance been coded as -QUERY?
   - If YES does the utterance *fail to highlight* an incorrect assumption made by the other speaker?
     - If yes then -KNOWLEDGE-MISMATCH
     - If no then: no coding

---

⁵One other such coding (PARTNER-BELIEF) does exist, but we shall not consider it in the context of this study.
B.6 Layout of Data

The dialogues are presented in the following way:

Eaq3c3.text

#TB 6
Past the forest fire?
#F7 INITIATE +QUERY
#F7 -FEATURE-INTRO [forest fire]
#G7

<previous turn>

#TB 6
<text>
#F7 <move type> <positive coding>
#F7 <negative coding>
#G7 <negative coding for other speaker>

<nexxt turn>

#T\{A,B\} = turn label  (A=Giver, B=Followe)
#G = Coding for Giver
#F = Coding for Follower

So each full turn has at least three coding lines associated with it:

1. This line states the move type (INSTRUCT, INITIATE, RESPONSE, OR FOLLOW-
   UP), and has any positive coding for that utterance
   - You can have multiple positive codings
   - Don’t worry if you have more than one line of these

2. This line contains any negative coding appropriate for that utterance
   - Again, don’t worry if you exceed the line

3. This line contains any negative coding for the other speaker (notice that it
   is labelled differently)
The number on the coding lines uniquely identifies that set of coding lines, which makes it easier to compare across the coding done by various coders. This is also useful because turns may consist of more than one move, and then it is necessary to have more than one set of coding lines. An example of this is given below.

Eaq3c6.text

#TA 44
Go just due left of the white mountain and keep going right the way up, at ... until about the level of the bottom of the pyramid.
#G43 INSTRUCT +ROUTE-SECTION +ROUTE-MARKER
#G43 -FEATURE-INTRO [pyramid]
#F43

#TB 45
Okay.
#F44 FOLLOW-UP +ACK-SHORT
#F44
#G44
that's past the lemon grove.
#F45 INITIATE +RELEVANT-INFO
#F45 -FEATURE-INTRO [lemon grove]
#G45

B.6.1 Adding Your Coding

In this section we wish to describe where you will add your coding decisions to the dialogue layout.

If we return to our previous example:

Eaq3c3.text

#TB 6
Past the forest fire?
#F7 INITIATE
#F7
#G7

235
This is how the dialogue data will be presented to you. Only the supertype label is entered on the first line. It is up to you to add appropriate positive and negative codings for the Giver and Follower.

**Positive Coding** goes on the same line as the supertype label

**Negative Coding** goes on the second line

**Negative Coding for the Other Speaker** goes on the final coding line.

Have a look at the example coded dialogues, and you should notice the following things:

- Many of these coding lines are empty, especially the final line. But, of course, when you are coding a dialogue for the first time, all these possibilities should be open to you.

- Most (but not all) moves will be positively coded for one or more dialogue attribute types.

- Negative coding is not appropriate for every move, in fact in most dialogues it is used reasonably rarely.

- Negative coding for the other speaker is very rare - it is mostly used for missed opportunities for checking routines.

- Not all separate turns have coding lines. This is usually for one of the following reasons:

  1. The turn is a continuation of the previous one (and the coding lines are located at that point)

  2. The turn is subject to an interruption, and the coding lines can be found at the speaker’s next turn

  3. The turn was effectively abandoned, so there is nothing to code.

  4. The content of the turn is irrelevant to the actual task, it is metacomment (e.g. personal interplay between speakers) and it is not relevant to the behaviours which we are trying to code here.
B.7 Summary

In this document we have detailed the task which we are asking you to attempt. We have described the coding process, the terminology, and the way in which the coding should be added to the data.

The Introduction to the Typology (Section B.5) is intended to lead you through the process of coding an individual utterance. To use it, you need to know both what move type the utterance is (INSTRUCT, INITIATE, RESPONSE, FOLLOW-UP), and its context. In this introduction, we try to make explicit how the coding system works, and where a particular category is and isn’t relevant.

This is a very brief guide, and the intention is for the user to choose the likely codings, and then check them against the fuller documentation of the Typology. We do not attempt to justify any of our categories here; again, this discussion is given in the fuller document. To include much more detail here would make this document too long to fulfil its function.
Appendix C

Tabular Summary of Typology

C.1 INSTRUCT moves

CODING FOR THE MOVE-TYPE

<table>
<thead>
<tr>
<th>INSTRUCT Move Used</th>
<th>Preferred INITIATE</th>
<th>Preferred Alternate Move?</th>
</tr>
</thead>
<tbody>
<tr>
<td>+/- NEW-FEATURE</td>
<td>-NEW-QUESTION</td>
<td>-REPLY-MIN -ACK-SHORT</td>
</tr>
<tr>
<td>+/- ROUTE-SECTION</td>
<td>-RELEVANT-INFO</td>
<td></td>
</tr>
<tr>
<td>+/- ROUTE-MARKER</td>
<td>-NEW-SUGGESTION</td>
<td></td>
</tr>
<tr>
<td>+/- SIGNPOSTS</td>
<td>-QUERY</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-CHECK</td>
<td></td>
</tr>
</tbody>
</table>

CODING FOR TREATMENT OF FEATURES

<table>
<thead>
<tr>
<th>Description</th>
<th>YES</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is the feature introduced or re-introduced in the move?</td>
<td>next question</td>
<td>coding N/A</td>
</tr>
<tr>
<td>Is the feature introduction the focus of the move</td>
<td>+FEATURE-INTRO</td>
<td>-FEATURE-INTRO</td>
</tr>
<tr>
<td>Is there an attempt to locate the feature?</td>
<td>+FEATURE-LOC</td>
<td>next question</td>
</tr>
<tr>
<td>Should there have been an attempt to locate the feature?</td>
<td>-FEATURE-LOC</td>
<td>coding N/A</td>
</tr>
<tr>
<td>Is the feature identified uniquely?</td>
<td>+FEATURE-UNIQUE</td>
<td>next question</td>
</tr>
<tr>
<td>If not, did it cause a problem?</td>
<td>-FEATURE-UNIQUE</td>
<td>coding N/A</td>
</tr>
</tbody>
</table>
## ADDITIONAL CODINGS

<table>
<thead>
<tr>
<th>KNOWLEDGE MISMATCH</th>
<th>YES</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is the move coded as -QUERY?</td>
<td>next question</td>
<td>coding N/A</td>
</tr>
<tr>
<td>Is this coding due to a failure to point out a failed assumption?</td>
<td>-KNOWLEDGE-MISMATCH</td>
<td>coding N/A</td>
</tr>
</tbody>
</table>
C.2 INITIATE moves

CODING FOR THE MOVE-TYPE

<table>
<thead>
<tr>
<th>INITIATE Move Used</th>
<th>Preferred INITIATE</th>
<th>Preferred Alternate Move?</th>
</tr>
</thead>
<tbody>
<tr>
<td>+NEW-QUESTION</td>
<td>-NEW-QUESTION</td>
<td>-REPLY-MIN</td>
</tr>
<tr>
<td>+RELEVANT-INFO</td>
<td>-RELEVANT-INFO</td>
<td>-ACK-SHORT</td>
</tr>
<tr>
<td>+NEW-SUGGESTION</td>
<td>-NEW-SUGGESTION</td>
<td></td>
</tr>
<tr>
<td>+QUERY</td>
<td>-QUERY</td>
<td></td>
</tr>
<tr>
<td>+CHECK</td>
<td>-CHECK</td>
<td></td>
</tr>
<tr>
<td>+OBJECTION</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CODING FOR TREATMENT OF FEATURES

<table>
<thead>
<tr>
<th>Description</th>
<th>YES</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is the feature introduced or re-introduced in the move?</td>
<td>next question</td>
<td>coding N/A</td>
</tr>
<tr>
<td>Is the feature introduction the focus of the move?</td>
<td>+FEATURE-INTRO</td>
<td>-FEATURE-INTRO</td>
</tr>
<tr>
<td>Is there an attempt to locate the feature?</td>
<td>+FEATURE-LOC</td>
<td>next question</td>
</tr>
<tr>
<td>Should there have been an attempt to locate the feature?</td>
<td>-FEATURE-LOC</td>
<td>coding N/A</td>
</tr>
<tr>
<td>Is the feature identified uniquely?</td>
<td>+FEATURE-UNIQUE</td>
<td>next question</td>
</tr>
<tr>
<td>If not, did it cause a problem?</td>
<td>-FEATURE-UNIQUE</td>
<td>coding N/A</td>
</tr>
</tbody>
</table>

ADDITIONAL CODINGS

<table>
<thead>
<tr>
<th>KNOWLEDGE MISMATCH</th>
<th>YES</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is the move coded as +QUERY or +OBJECTION?</td>
<td>next question</td>
<td>next section</td>
</tr>
<tr>
<td>Is this coding due to highlighting a failed assumption?</td>
<td>+KNOWLEDGE-MISMATCH</td>
<td>next section</td>
</tr>
<tr>
<td>Is the move coded as -QUERY?</td>
<td>next question</td>
<td>coding N/A</td>
</tr>
<tr>
<td>Is this coding due to a failure to point out a failed assumption?</td>
<td>-KNOWLEDGE-MISMATCH</td>
<td>coding N/A</td>
</tr>
</tbody>
</table>
C.3 RESPONSE moves

CODING FOR THE MOVE-TYPE

<table>
<thead>
<tr>
<th>RESPONSE Move Used</th>
<th>Preferred RESPONSE</th>
<th>Preferred Alternate Move?</th>
</tr>
</thead>
<tbody>
<tr>
<td>+REPLY-MIN -REPLY-FULL</td>
<td>-QUERY</td>
<td>-ACK-SHORT</td>
</tr>
<tr>
<td>+REPLY-YN</td>
<td>(-INFO-INTEG)</td>
<td></td>
</tr>
<tr>
<td>+REPLY-FULL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(+INFO-INTEG)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CODING FOR TREATMENT OF FEATURES

<table>
<thead>
<tr>
<th>Description</th>
<th>YES</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is the feature introduced or re-introduced in the move?</td>
<td>next question</td>
<td>coding N/A</td>
</tr>
<tr>
<td>Is the feature introduction the focus of the move?</td>
<td>+FEATURE-INTRO</td>
<td>-FEATURE-INTRO</td>
</tr>
<tr>
<td>Is there an attempt to locate the feature?</td>
<td>+FEATURE-LOC</td>
<td>next question</td>
</tr>
<tr>
<td>Should there have been an attempt to locate the feature?</td>
<td>-FEATURE-LOC</td>
<td>coding N/A</td>
</tr>
<tr>
<td>Is the feature identified uniquely?</td>
<td>+FEATURE-UNIQUE</td>
<td>next question</td>
</tr>
<tr>
<td>If not, did it cause a problem?</td>
<td>-FEATURE-UNIQUE</td>
<td>coding N/A</td>
</tr>
</tbody>
</table>

ADDITIONAL CODINGS

<table>
<thead>
<tr>
<th>KNOWLEDGE MISMATCH</th>
<th>YES</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is the move coded as -QUERY?</td>
<td>next question</td>
<td>coding N/A</td>
</tr>
<tr>
<td>Is this coding due to a failure to point out a failed assumption?</td>
<td>-KNOWLEDGE-MISMATCH</td>
<td>coding N/A</td>
</tr>
</tbody>
</table>
C.4 FOLLOW-UP moves

CODING FOR THE MOVE-TYPE

<table>
<thead>
<tr>
<th>FOLLOW-UP Move Used</th>
<th>Preferred FOLLOW-UP</th>
<th>Preferred Alternate Move?</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ACK-SHORT</td>
<td>-ACK-FULL</td>
<td>-QUERY</td>
</tr>
<tr>
<td>+ACK-FULL</td>
<td>(-INFO-INTEG)</td>
<td>-OBJECTION</td>
</tr>
<tr>
<td>(+INFO-INTEG)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CODING FOR TREATMENT OF FEATURES

<table>
<thead>
<tr>
<th>Description</th>
<th>YES</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is the feature introduced or re-introduced in the move?</td>
<td>next question</td>
<td>coding N/A</td>
</tr>
<tr>
<td>Is the feature introduction the focus of the move?</td>
<td>+FEATURE-INTRO</td>
<td>-FEATURE-INTRO</td>
</tr>
<tr>
<td>Is there an attempt to locate the feature?</td>
<td>+FEATURE-LOC</td>
<td>next question</td>
</tr>
<tr>
<td>Should there have been an attempt to locate the feature?</td>
<td>-FEATURE-LOC</td>
<td>coding N/A</td>
</tr>
<tr>
<td>Is the feature identified uniquely?</td>
<td>+FEATURE-UNIQUE</td>
<td>next question</td>
</tr>
<tr>
<td>If not, did it cause a problem?</td>
<td>-FEATURE-UNIQUE</td>
<td>coding N/A</td>
</tr>
</tbody>
</table>

ADDITIONAL CODINGS

<table>
<thead>
<tr>
<th>KNOWLEDGE MISMATCH</th>
<th>YES</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is the move coded as -QUERY?</td>
<td>next question</td>
<td>coding N/A</td>
</tr>
<tr>
<td>Is this coding due to a failure to point out a failed assumption?</td>
<td>-KNOWLEDGE-MISMATCH</td>
<td>coding N/A</td>
</tr>
</tbody>
</table>
# C.5 Summary of Positive Codings

<table>
<thead>
<tr>
<th><strong>INSTRUCT</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>+NEW-FEATURE</td>
<td>Move contains one new feature</td>
</tr>
<tr>
<td>+ROUTE-SECTION</td>
<td>Move discusses one route-section</td>
</tr>
<tr>
<td>+ROUTE-MARKER</td>
<td>Move contains all necessary markers for part of route described</td>
</tr>
<tr>
<td>+SIGNPOSTS</td>
<td>Move is too long (coded as -NEW-FEATURE and/or -ROUTE-SECTION), but is well-organised</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>INITIATE</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>+NEW-QUESTION</td>
<td>Asks question not directly prompted by previous utterance</td>
</tr>
<tr>
<td>+RELEVANT-INFO</td>
<td>Introduces new, unsolicited information (‘new’ in terms of focus, potentially relevant to route section)</td>
</tr>
<tr>
<td>+NEW-SUGGESTION</td>
<td>Makes unsolicited suggestion about where route might go next (need not be correct)</td>
</tr>
<tr>
<td>+QUERY</td>
<td>Question prompted by previous utterance either because of information problem, or checking self understanding (check if +KNOWLEDGE-MISMATCH is appropriate)</td>
</tr>
<tr>
<td>+OBJECTION</td>
<td>Statement prompted by previous utterance, concerned with information problem (check if +KNOWLEDGE-MISMATCH is appropriate)</td>
</tr>
<tr>
<td>+CHECK</td>
<td>Question which solicits other understanding of information already offered</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>RESPONSE</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>+REPLY-MIN -REPLY-FULL</td>
<td>Insufficient or inappropriate response</td>
</tr>
<tr>
<td>+REPLY-YN</td>
<td>Yes-No reply to Yes-No question</td>
</tr>
<tr>
<td>+REPLY-FULL</td>
<td>reply to WH-question, or full reply to Yes-No question</td>
</tr>
<tr>
<td>(+INFO-INTEG)</td>
<td>Additional information offered (Move should be coded as +REPLY-FULL) [RARE]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>FOLLOW-UP</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>+ACK-SHORT</td>
<td>Appropriately brief follow-up</td>
</tr>
<tr>
<td>+ACK-FULL</td>
<td>Full follow-up</td>
</tr>
<tr>
<td>(+INFO-INTEG)</td>
<td>Additional information offered (Move should already be coded as +ACK-FULL) [RARE]</td>
</tr>
</tbody>
</table>

Continued over page...
Summary of Positive Codings Continued

<table>
<thead>
<tr>
<th>FEATURE-SPECIFIC codings</th>
</tr>
</thead>
<tbody>
<tr>
<td>+FEATURE-INTRO</td>
</tr>
<tr>
<td>+FEATURE-LOC</td>
</tr>
<tr>
<td>+FEATURE-UNIQUE</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HIGHER-LEVEL codings</th>
</tr>
</thead>
<tbody>
<tr>
<td>+KNOWLEDGE-MISMATCH</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
## C.6 Summary of Negative Codings

<table>
<thead>
<tr>
<th>INSTRUCT</th>
<th>FREQUENCY</th>
</tr>
</thead>
<tbody>
<tr>
<td>-NEW-FEATURE</td>
<td>Move contains more than one new feature</td>
</tr>
<tr>
<td>-ROUTE-SECTION</td>
<td>Move covers more than one route section</td>
</tr>
<tr>
<td>-ROUTE-MARKER</td>
<td>Move omits important markers for described route section</td>
</tr>
<tr>
<td>-SIGNPOSTS</td>
<td>Move is both long and lacking organisation (coded as -NEW-FEATURE and/or ROUTE-SECTION)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>INITIATE</th>
<th>FREQUENCY</th>
</tr>
</thead>
<tbody>
<tr>
<td>-NEW-QUESTION</td>
<td>Not applicable</td>
</tr>
<tr>
<td>-RELEVANT-INFO</td>
<td>Failure to introduce useful information when necessary</td>
</tr>
<tr>
<td>-NEW-SUGGESTION</td>
<td>failure to make a suggestion</td>
</tr>
<tr>
<td></td>
<td>(This behaviour is potentially helpful rather than necessary, and therefore failure is rare)</td>
</tr>
<tr>
<td>-QUERY</td>
<td>Failure to indicate information problem</td>
</tr>
<tr>
<td>-OBJECTION</td>
<td>Collapsed into above category, as syntactic differences do not occur in unrealized options</td>
</tr>
<tr>
<td>-CHECK</td>
<td>Failure to check other's understanding of information offered</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RESPONSE</th>
<th>FREQUENCY</th>
</tr>
</thead>
<tbody>
<tr>
<td>-REPLY-MIN</td>
<td>No response given when required</td>
</tr>
<tr>
<td>+REPLY-MIN -REPLY-FULL</td>
<td>Reply too short, or inappropriate</td>
</tr>
<tr>
<td>(-INFO-INTEG)</td>
<td>More information necessary</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FOLLOW-UP</th>
<th>FREQUENCY</th>
</tr>
</thead>
<tbody>
<tr>
<td>-ACK-SHORT</td>
<td>No follow-up given when necessary</td>
</tr>
<tr>
<td>-ACK-FULL</td>
<td>Inappropriately brief follow-up</td>
</tr>
<tr>
<td>(-INFO-INTEG)</td>
<td>More information necessary</td>
</tr>
</tbody>
</table>

Continued over page...
### Summary of Negative Codings Continued

<table>
<thead>
<tr>
<th>FEATURE-SPECIFIC codings</th>
<th>FREQUENCY</th>
</tr>
</thead>
<tbody>
<tr>
<td>-FEATURE-INTRO</td>
<td>New feature introduced, but not highlighted</td>
</tr>
<tr>
<td>-FEATURE-LOC</td>
<td>Failure to start negotiation process for unshared (typically) feature</td>
</tr>
<tr>
<td>-FEATURE-UNIQUE</td>
<td>Failure to identify feature uniquely, which causes problem (duplicate feature)</td>
</tr>
<tr>
<td></td>
<td>Very rare</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HIGHER-LEVEL codings</th>
<th>FREQUENCY</th>
</tr>
</thead>
<tbody>
<tr>
<td>-KNOWLEDGE-MISMATCH</td>
<td>Move fails to point out mistaken assumption</td>
</tr>
<tr>
<td></td>
<td>Should be coded as -QUERY</td>
</tr>
</tbody>
</table>
