Interaction in dialogue: The effects of partner feedback on speakers, addressees and overhearers.

Ciara M. Catchpole

Ph.D.

University of Edinburgh

2006

Contents

Figures

Tables

Declaration

This thesis has been composed by myself, and the research presented herein is my own. No portion of the work has been submitted for any other degree or professional qualification.

Ciara Catchpole

Acknowledgements

This thesis could not have been completed without help from a great number of people, to whom I would like to show my appreciation. The greatest of thanks go to Martin Pickering and Holly Branigan, who have excelled in their wisdom, efficiency, kindness and encouragement, and have generally been superb in every way. I could not have asked for a better pair of mentors, and consider myself extremely fortunate to have been under their supervision. I remain grateful to my undergraduate dissertation supervisor, Wayne Murray, for introducing me to this field and showing me what an exciting world of research lay ahead. Jen Pardo and Bob Krauss graciously allowed me to work with them at Columbia University for 3 months during this period of doctoral study; the findings resulting from that time are not included in this thesis (largely for reasons of continuity), but this allowed me to enjoy a different department, culture and continent; an experience I will never forget, and for which I am very grateful. New York hasn't seen the last of me yet!

In terms of practical help, I am indebted to the Economic and Social Research Council (ESRC) for their generous funding of this work, and to Lucy MacGregor and Kimberley Bodner for helping me run Experiment 2. I would also like to thank Manon Jones for inspiring conversations on Experiments 5 and 6, Janet McLean for sharing her mastery of statistics with me on Experiment 2, and Sarah Haywood for advice, references and encouragement. I secretly suspect that Janet and Sarah together might know all there is to know in the world.

The support staff in the Psychology department have been extremely helpful at all times, and I particularly thank Davy Wilkinson for not chasing me when I kept DAT recorders for too long, Roy Welensky for graphics help, Mike Allerhand for help with E-Prime, Bill Robertson for keeping my wine glass full, and Jimmy Duncan for ceaselessly winding me up.

My office-mates (and, I hope, still friends!) have been fantastic and offered advice on everything from statistics to recipes; thanks to Susie Flett, Matt Watson, Anna Hatzidaki, Gaurav Malhotra, Claudine Raffray and Miki Tanaka for much laughter and support. Continuing on this personal note, many friends have cheered me (in both senses) along the way, most notably Louise Wilkes, Karen Hewitt, Shilpa Grewal, Siobhain Murdoch, Bregje de Kok, Pab Roberts, Kathryn Poolman, Heather Apsley, Susanna Dick, Anna Crangle, Julie Bryan and Helen Byers. I adore them all, and thank them gratefully.

Finally, I offer endless thanks and love to mum Oonagh, Conor and Patrick, for being the most supportive, encouraging family I could have asked for, and for patiently believing in me, often more than I believed in myself. I would like to dedicate this thesis to them, and to the memory of my dad, Owen, who I believe would have been as fascinated by Psycholinguistics as his daughter is.

'Words constitute the ultimate texture and stuff of our moral being, since they are the most refined and delicate and detailed, as well as the most universally used and understood, of the symbolisms whereby we express ourselves into existence. We become spiritual animals when we become verbal animals. The fundamental *distinctions can only be made in words. Words are spirit'.*

(Iris Murdoch, 1972, 'Salvation by words')

Abstract

The primary aim of this thesis is to analyse how interaction with a conversational partner affects the performance of speakers, their addressees, and those who overhear the discourse. Speakers frequently appear to adjust their speech to accommodate their addressees' needs (although the extent to which this is deliberate is a topic of much debate), and this often seems to occur as a direct response to the feedback they receive from their partners. The bulk of the research in this area has focussed upon the manner in which speakers use feedback to facilitate the overall success of the interaction (in terms of completing the task at hand, for example), but fewer studies have investigated how feedback specifically affects detailed characteristics of the speakers' speech. This thesis attempts to investigate such a topic, whilst also examining the benefit to addressees of being able to give feedback, and the benefit to overhearers of hearing addressees' feedback.

The thesis begins by examining the effect of feedback on speakers' and addressees' performances. It first reports a referential communication task which investigated some of the notable differences between speakers' speech in monologue and dialogue contexts (Experiment 1). I focussed in particular on how detailed aspects of language production, such as the length of object descriptions and the repetitiveness of the language used, varied between these two feedback conditions. I also looked at how the ability to give feedback aided the addressees' performances on the given task. Experiment 2 then analysed how the amount of feedback received by the speaker related to, firstly, the shortening of their item descriptions over repetitions, and secondly, the increasing consistency of descriptions (in terms of lexical overlap) during the experiment.

The second section of the thesis focuses primarily on the benefit for overhearers of hearing other people's feedback. It first reports two experiments that replicated and expanded on a study by Fox Tree (1999) which showed that overhearers identified tangrams more accurately when they overheard a dialogue rather than a monologue. Experiments 3 and 4 tested two explanations proposed by Fox Tree for this result; firstly, the potential presence of additional perspectives in dialogue (for example, one partner viewing a tangram as a 'chicken', and the other calling it an 'ice skater'), and secondly the more numerous discourse markers in dialogue in comparison with monologue. Additionally, the use of repeated descriptions by the speakers allowed me to analyse the effect of interlocutors' 'conceptual pacts' on overhearers' task performance. Finally, Experiments 5 and 6 assessed how the benefit of, firstly, giving feedback (for addressees) and secondly, hearing feedback (for overhearers) was affected by the difficulty of the task at hand. Overall, this thesis provides evidence that whether people are producing, receiving or simply overhearing feedback, it influences their performance in a positive manner, meaning that, in general, a greater amount of interaction leads to more successful communication for everyone involved.

Chapter 1: Introduction

1.1 Chapter overview

This chapter briefly introduces the themes of the thesis. Beginning with a discussion of why research into dialogue is important, I move on to discuss the specific questions tackled in the next six chapters, concluding with an overview of the experiments and topics presented herein.

1.2 The importance of studying dialogue

The idea of studying dialogue, rather than monologue, is still a slightly novel one; since the beginning of psycholinguistics (ca.1950), the majority of the experimental studies and theories in this area have focussed on monologue. Although there were a few notable dialogue studies in the 1960's (e.g. Krauss & Weinheimer, 1966; Krauss & Glucksberg, 1969), it is really only since the 1980's that this area has evolved, largely due to work by Herbert Clark and colleagues (e.g. Clark & Marshall, 1978, 1981; Clark & Wilkes-Gibbs, 1986; Clark & Schaefer, 1989). These contributions developed into a theory of dialogue that has informed much of the ensuing research. More recently, however, another model of dialogue was proposed by Pickering and Garrod (2004), and this seems set to influence a new wave of research. It studies dialogue from a slightly different, more mechanistic perspective, focussing more on seemingly automatic processes such as priming (roughly speaking, imitation of oneself or a partner) and less on the interlocutors' intentional goals, as were the focus of Clark's model. One thing that both these models have in common, however, is the study of dialogue as distinct from

monologue; a method of communication which, whether intentionally or automatically, is influenced by the presence of a partner, and where utterances are not just influenced by what the speaker intends to say, but also by what he and his interlocutor have previously said, and to some extent the knowledge that both partners hold. Thus dialogue is understandably more complex to study than monologue; it is difficult enough to take account of people's syntactic structures, lexical choices, prosody and so on, without having to also consider their knowledge of their interlocutor's previous utterances, beliefs, culture, experience, and their response to the feedback they receive from that person. But consider this we must, for a model of language will never be complete while it studies the comprehension and production of monologue in an isolated context, with no 'real' interaction in sight.

For a theoretical rationale for the study of dialogue, we need go no further than Chomsky's (1965) distinction between speakers' knowledge of a language (their 'competence') and their actual implementation of it (their 'performance')[1]. If the goal of psycholinguists is to study performance, rather than competence, then presumably we should focus upon the typical use of language in everyday life; the language which is used to buy apples at the greengrocers, to summarise the storyline of a film, or to discuss an upcoming holiday with friends; language complete with disfluencies, fractured sentences and mistakes both in content and in articulation. This, then, is the aim of dialogical research, and particularly the aim of this thesis; to examine 'real' language, in as close to actual communicative settings as we can get. While the *most* naturalistic study of language would involve analysing corpora, I stopped short of this in this thesis simply because of the specificity of the given topic; while studies of syntax and prosody can benefit

---

[1] Although Chomsky introduced this concept to refer to our knowledge and use of grammatical structures, its potential applications are widespread.

greatly from such sources, for the present studies I required more control over the *aims* of the speakers, and the resulting *performances* of the addressees and overhearers. To this end I employed a number of 'referential communication' tasks (introduced by Krauss and Weinheimer, 1966), which aim to simulate natural dialogue, whilst determining the intentions of both interlocutors as far as possible.

1.3 Interacting in dialogue

Two people in a dialogue do not produce autonomous, disconnected contributions; rather, they *interact*; a process which has practical repercussions on the speech and behaviour of both partners. In particular, it affects the way speakers refer to objects. One key finding from the field of dialogue is that when speakers refer to objects they have previously mentioned, their descriptions become shorter with repetition (Clark & Wilkes-Gibbs, 1986; Krauss & Weinheimer, 1966; Isaacs & Clark, 1987) and they use more definite determiners (e.g. 'the') rather than the previous indefinite determiners (e.g. 'a'; Clark &Wilkes-Gibbs, 1986; Hupet & Chantraine, 1992; Wilkes-Gibbs & Clark, 1992). These findings, along with others, have been taken to suggest that speakers alter their speech according to their partners' knowledge (a process referred to as 'audience design'); knowledge which they determine either from their prior beliefs about the addressee, or from their addressees' online feedback, or both. Certainly a number of other studies have also shown that speakers adjust their speech in ways that directly reflect their addressees' needs, in terms of *what* items the speakers refer to (Lockridge & Brennan, 2002), *how* they refer to them (Brennan & Clark, 1996), and how *much* information they provide their addressees with (Isaacs & Clark, 1987). But less is known about how the presence or absence of feedback from an addressee affects these adjustments of speakers' speech; whether they are dependent upon the speaker's receipt of feedback, or occur even in its absence.

1.4 Overhearing interaction in dialogue

If little is known of the effect of feedback, even less is known of how it affects people other than the person producing it (the addressee) and the person for whom it is intended (the speaker). Although it seems that overhearers benefit less from overhearing a dialogue than addressees do from participating in it (Schober & Clark, 1989), but that even overhearing feedback is better than hearing none at all (Fox Tree, 1999), almost no research has focussed on where the benefit of hearing feedback lies for overhearers. The later chapters of this thesis will attempt to examine such a topic.

1.5 Research questions addressed in this thesis

The first half of this thesis examines speakers' manners of referring to objects; in particular how these references change (or do not change) over time, and how this relates to the feedback received. Krauss and Weinheimer (1966) found that when speakers refer to the same items again and again, they use fewer words with each description, and that although this happens both with and without feedback, it occurs most notably with feedback. But what happens when the same speaker refers to *new* items during a task; are the descriptions of these items also shortened in dialogue (that is, with feedback) in comparison with monologue (without feedback)? If so, this may indicate that speakers transfer presumptions about their addressees' understanding from old items to new items as a result of feedback, a finding which was suggested by Isaacs and Clark (1987).

Another question of interest relates to what exactly happens in the shortening of descriptions with repetition. Are the same words used again and again? And do

descriptions which are shortened (after a number of repetitions) tend to be more 'consistent' with previous descriptions than those which aren't shortened? We can relate the idea of consistency to the concept of 'conceptual pacts' in dialogue (Brennan & Clark, 1996), and, in a different sense, to 'alignment' of lexical or conceptual representations (Pickering & Garrod, 2004). I will also look at how both these processes - of shortening and increasing consistency – are affected by feedback from an interlocutor.

The second half of the thesis moves onto an analysis of how feedback affects overhearers of a dialogue. Firstly, I test two hypotheses for why it might be useful for overhearers to overhear feedback, and in particular, if it is hearing the actual feedback that benefits the overhearers, or if its benefit stems from the way in which it affects the speaker. Additionally, we can see if the conceptual pacts reportedly created in dialogue have an impact on overhearers: do they prove to be a disadvantage to people who were not involved in their construction, or does the input of two people make descriptions more comprehensible to overhearers than if they were created by just one person? In this way we can perhaps come closer to understanding exactly what it is about overhearing feedback that benefits overhearers.

Towards the end of the thesis, I look at if and how the benefit of feedback for addressees and overhearers is affected by the difficulty of the task. It is possible that the ability to give feedback will be more useful to addressees when the cognitive burden on them is greater. But will this advantage also extend to overhearers? Also, we know that it is more useful to participate in a dialogue than to overhear it (Schober & Clark, 1989), but is this also true of monologue; is it more useful to be an addressee in monologue than to be an overhearer of it?

If this is the case, it may suggest that the advantage found by Schober and Clark relates more to the participatory status of an addressee than to any specific action of theirs.

1.6 Thesis overview

Chapter 1 is a review of the current literature on dialogue, beginning with a distinction between the language-as-product and language-as-action traditions of research (Clark, 1996). I then describe Clark's Collaborative model of language production, with a focus on collaboration, common ground and grounding, before moving onto a brief discussion of the evidence for the use of audience design by speakers. This leads to an overview of the relevant findings on syntactic, lexical and conceptual priming, followed by a description of Pickering and Garrod's (2004) Interactive Alignment model. The second section of the literature review focuses on feedback, and after a description of the possible definitions and purposes of feedback, I examine the effect it has on addressees', overhearers' and speakers' performances, primarily in the context of referential communication studies.

The next four chapters present six experiments which employ two different experimental paradigms. Chapter 3 describes Experiment 1; a exploratory referential communication study in which pairs of participants worked together to allow the 'builder' to create a Lego model similar to the one held by the 'describer', with no visual contact and under three feedback conditions: full feedback (free verbal interaction between the describer and builder), no feedback (only the describer was permitted to speak) and minimal feedback (the builder was only allowed to produce one-word utterances). A full analysis of the speakers' speech allowed me to assess the effect of varying levels of feedback on the length of noun phrases describing Lego blocks, the definiteness of determiners, the number of

features used in noun phrases, the particular nouns used to refer to Lego blocks, the number of disfluencies and the repetitiveness of the speakers' speech as a whole.

Chapter 4, which describes Experiment 2, examines the idea of consistency in speech along with the shortening of noun phrases with repetition. Again this study employed a referential communication task, this time using tangram figures instead of Lego blocks, and in only two feedback conditions: full feedback and no feedback. This allowed an analysis of the relationship between shortening and consistency, and the effect that feedback had (or didn't have) on both these elements.

Chapter 5 introduces a new paradigm, where recordings of the tangram descriptions produced in Experiment 2 were played to overhearers in two new experiments: Experiments 3 and 4. This chapter investigates the finding by Fox Tree (1999) that it is more beneficial for overhearers who are attempting to complete a tangram identification task to overhear a dialogue between two people discussing how to complete the task, than a monologue. After discounting a number of plausible explanations for this effect, she suggested it might be a result of hearing more than one perspective on the tangrams in the dialogue condition (for example, seeing an ambiguous figure as either an ice-skater or a chicken), or alternatively, that it might be due to the additional discourse markers (for example, 'I mean' and 'you know') in the dialogue descriptions. This chapter describes the testing of both these hypotheses, by excising the addressee's speech (Experiment 3) or the speakers' discourse markers (Experiment 4) from the descriptions in the dialogue condition, and comparing them with the original descriptions. These studies also gave me the opportunity to assess how speakers' later references to tangrams compared with earlier ones in terms of their usefulness to overhearers in the two feedback conditions, and to relate this to the idea of 'conceptual pacts' in dialogue (Brennan & Clark, 1996).

Chapter 6 describes two studies (Experiments 5 and 6) which examined the effect of task difficulty (in this case, represented by the 'discriminability' of tangrams) on tangram matching by addressees (Experiment 5) and then, in a separate study, by overhearers (Experiment 6). I also analysed if the feedback condition (no feedback; full feedback; restricted feedback) interacted with task difficulty in this study. Finally, I compared being an addressee of a referential communication task with being an overhearer of the same task (similarly to Schober and Clark's (1989) findings on dialogue), assessing in particular if the differences between participants' task success in these roles was affected by the presence or absence of feedback.

The concluding chapter, Chapter 7, summarises the findings from the six experiments which form this thesis, analyses what these experiments as a whole can tell us about the benefits of feedback, and discusses ideas for future research.

Chapter 2: Literature Review

2.1 Chapter Overview

This chapter will give an overview of the current literature on dialogue, with a particular focus on how speakers are influenced by their addressees, and more specifically, the effect that feedback from these addressees has on speakers, addressees and overhearers. It begins in Sections 2.2 and 2.3 by examining the necessity of research into dialogue, and the distinction made by Clark (1992) between the language-as-product and language-as-action traditions of research. Sections 2.4, 2.5 and 2.6 give an overview of Clark's Collaborative model, focussing on how interlocutors in dialogue decide upon terms of reference for particular items, and how this affects their subsequent references. Sections 2.7, 2.8 and 2.9 introduce the ideas of audience design and priming in dialogue, and then Section 2.10 looks at Pickering and Garrod's (2004) Interactive Alignment model, relating it to the findings on priming. Section 2.11 makes a comparison of the effects of these two processes. Section 2.12 narrows down our examination of dialogue to focus on feedback; after giving a brief account of what feedback is and how it varies between cultures and genders, I look at the reported effects on addressees of being able to give feedback, on speakers of receiving feedback, and on overhearers of overhearing it.

2.2 Introduction

Although at first glance dialogues may seem to be little more than sequences of alternating monologues, there is increasing evidence that this is not an accurate

picture. It could be argued that *dialogue*, rather than monologue, is the most natural form of language use; it is certainly the means by which we all learnt to speak, and, for most people, comprises their primary means of communication. Because we typically master the art of conversing at a very young age, we tend to take our ability to do this for granted, underestimating the many skills involved in its execution. It is only when we turn to look at dialogue from a psycholinguistic perspective that we realise just how much information the language production and comprehension systems must take account of at any one time, and the many aspects of this information which must be employed in constructing contributions to a dialogue.

Despite the apparent similarities between monologues and dialogues, there are many differences intrinsic to their production. One of the most obvious is that a speaker in dialogue needs to concurrently comprehend an interlocutor's speech, planning his own speech and listening to his partner's almost simultaneously, unlike speakers in monologue[2]. Multi-tasking on related tasks has been shown to have dramatic effects on performance; Hermer-Vazquez, Spelke and Katsnelson (1999) demonstrated that participants concurrently completing a language task (verbal shadowing) and a task which involves language to a lesser extent (people using spatial memory to reorientate themselves after becoming disorientated[3]) showed a substantial detrimental effect on the second task, more so than dual-tasking between two less related tasks (*non*-verbal shadowing and spatial

---

[2] Although even in monologue a speaker will monitor his own speech, which may involve similar processes, according to Levelt (1983).

[3] The authors concluded from this experiment that spatial orientation depends on language to some extent.

reorientation). This suggests that the commonalities between the two tasks increased the difficulty of multi-tasking in the first instance. It is probable, then, that multi-tasking on two explicitly language-based tasks will affect speech production to an even greater extent. Further evidence for this comes from Hohlfeld, Sangals and Sommer (2004), who found that while a language task (comparing meanings of pairs of nouns) showed some interference from a secondary spatial task (foot pedal pressing according to symbols shown on a computer screen), even more interference was demonstrated when the secondary task was also language-related (foot pedal pressing according to the letters 'L' (left) and 'R' (right) on the screen). These findings concur with Baddeley and Hitch's (1974) model of working memory, which contains two domain-specific (verbal and visuo-spatial) 'slave' systems that can each carry out only one task at a time. This means that when two tasks compete for the same mechanism (for example two language tasks), they will interfere. These studies demonstrate the potential difficulty for speakers in producing and comprehending speech within the same rough time-frame, and yet this is what we do apparently effortlessly every time we participate in a dialogue.

Even the comprehension of an addressee's speech by a speaker in dialogue is not as straightforward as it might initially seem; difficulties in comprehension may result from the large number of disfluent, interrupted and otherwise fragmented utterances that characterise dialogue (described in Fernandez and Ginzburg, 2002). Further to this, the speaker is not only expected to comprehend his addressee's speech, but, in an ideal situation, to adjust his own productions in the light of her response, be it verbal (Clark & Krych, 2004) or gestural (Bavelas, Coates & Johnson, 2002), or even a lack of response where one was expected (Bavelas, Coates &

Johnson, 2000; Kraut, Lewis & Swezey, 1982[4]). The extent to which this actually occurs is a topic of debate, and will be reviewed in Section 2.4.1 of this chapter.

In addition to these challenges, an interlocutor in dialogue must also construct a longer-term semantic plan to account for the progress of the discourse; extensive links need to be formed between what has just been said by person A to person B and how B responds to A's utterance, then how A will reply to B's response and so on. That is, there needs to be a running script of the previous dialogue in both (or all) the interlocutors' minds; one which is constantly updated upon the introduction of new speech. Along with this, there will be a second script that represents how the speaker expects the dialogue to pan out, and what kind of contributions he plans to make to it; this must be highly flexible, given that he cannot know for sure what his interlocutor is going to say next. Contrast this with monologue, where the speaker only needs to consider his own productions (although he may also have a model of how well he thinks his listener will understand his speech). The fact that neither speaker in dialogue knows exactly how the conversation will proceed only adds to the difficulty. This has bearings on not only the content of what the interlocutors say, but also the timing; a number of papers illustrate some of the difficulties intrinsic to deciding not only *what* to say in dialogue (e.g. Ginzburg, 1996), but *when* to say it (e.g. ten Bosch, Oostdijk & Boves, 2005).

The characteristics outlined above would indicate that dialogue should be much more difficult for speakers than monologue, but this belies the truth: contributions in an interactive dialogue are reported to be more expressive and informative than

---

[4] Although these papers did not explicitly report this finding, it is likely that the effect they described was a result of the speakers expecting feedback from an addressee and not receiving it; we will discuss this further in Section 3.2.2.

those in monologue (Bavelas, Coates & Johnson, 2000; Kraut, Lewis and Swezey, 1982)[5], despite their often fragmented nature. Additionally, dialogue instinctively 'feels' much easier for speakers than monologue, which may be supported by the findings that overall speech rate is faster in dialogue than in monologue (Riggenbach, 1989, on non-natives, reported in Bell, 2003) and individual words are articulated more quickly in dialogue than in monologue (McAllister, Potts, Mason & Marchant, 1994). It is also more useful for people in referential communication or description tasks to participate in dialogues than to overhear monologues (Kraut, Lewis & Swezey, 1982; Schober & Clark, 1989). It seems, then, that having an interlocutor benefits our language production system in terms of communication, if not in terms of style.

2.3 The language-as-product and language-as-action traditions

In some senses, the increase in dialogue research in recent years is attributable to a proposal by Clark (1992) that language must be understood in context. Clark made a distinction between what he termed the 'language-as-product' and 'language-as-action' traditions of psycholinguistics, and in doing so, demonstrated why research on monologue alone is not sufficient to comprise our whole understanding of language, and why explicit theories of dialogue are a necessary addition to this field.

The language-as-product tradition could be considered to be the original strain of psycholinguistics, and centres around the production and comprehension of monologue. This tradition, Clark pointed out, focuses on only the bare linguistic

---

[5] But see Section 2.12.4 for a caveat to these findings

elements of language, primarily examining how it is encoded and decoded by speakers in autonomous processes. Being a somewhat mechanistic approach, it abstracts language away from all other aspects of the context, such as the time, place, ultimate conversational goal and in fact everything except for the actual words produced by speakers. Data is often collected by presenting, or soliciting the production of, decontextualised language in unnatural, solo environments. This sanitised version of language avoids all the 'problems' that characterise natural speech, such as disfluencies, ellipses, interruptions and so on, which cause variability in spontaneous speech and make it difficult to study. This type of research has produced a huge body of knowledge over the years, which has been massively beneficial to our understanding of the language production and comprehension systems, and has informed countless theoretical models (e.g. Garrett, 1975; Fodor, Bever and Garrett, 1974). However, due to the experimental restrictions placed upon the participants, the conclusions drawn from these studies may not be entirely applicable to the main type of language use we engage in on a daily basis, that is, spontaneous dialogue.

The more recent language-as-action tradition involves studying language in its role as a means to an end, that is, the successful accomplishment of communicative goals. This tradition differs from language-as-product in two main respects: firstly, it advocates the analysis of speech in its natural context, which is far from the near-perfect monologues produced in soundproof booths which have been the foundation of psycholinguistic experimentation for decades. Instead this tradition examines the 'real' use of language, complete with disfluencies, false starts and so on. Secondly, it is assumed that language can be understood only as part of the *social* context in which it occurs, with all the intentions and goals that that involves, and for this reason language-as-action focuses upon dialogue as its primary model, rather than monologue. In his book, 'Using Language' (1996), Clark presented language as a *joint* tool which is employed in actions (but is not necessarily in itself the sole basis of those actions), stating that, "We cannot study language use

without studying joint activities and vice-versa" (p387). Instead of viewing dialogue as the combined outputs of two autonomous information processors, then (as the language-as-product tradition might dictate), it is seen as a joint accomplishment of the interlocutors, worthy of research in its own right.

Drawing us further into the idea of language being a joint project, Clark suggested that in conversation, each participant is affected by their partner's performance (for the most part, by their speech and gestures), and most likely adjusts their own performance or utterances in response to this. As Clark and Brennan (1991) noted, such collaboration is intrinsic to many forms of human interaction:

> "It takes two people working together to play a duet, shake hands, play chess, waltz, teach, or make love. To succeed, the two of them have to coordinate both the content and process of what they are doing...Communication, of course, is a collective activity of the first order" (p127-8).

Some of the first proponents (if implicitly) of this interactive concept of dialogue were Emmanuel Schegloff and colleagues, who, approaching dialogue from a sociolinguistic standpoint, produced a vast number of papers outlining turn-taking strategies and other practices involved in dialogue (e.g. Sacks, Schegloff & Jefferson, 1974; Schegloff, 2000). Clark and Krych (2004) describe how in these 'bilateral' accounts (also referred to as 'dialogic' models by Krauss and Fussell, 1996), firstly, the structure of a conversation is determined jointly by the speaker and the addressee, and secondly, the speaker takes account of his addressee's level of knowledge. They contrast these with more traditional, 'unilateral' accounts (e.g. Horton & Keysar, 1996) in which a speaker is seen to produce utterances as solo endeavours, and almost without reference to his addressee's state of mind (at least during initial processing). Unilateral accounts allow little provision for the speaker to monitor the addressee's level of understanding; rather, according to these, speech

production is an almost entirely egocentric process. Clark argued that the majority of psycholinguistic research in the past has implicitly presumed that dialogue is unilateral, and the truth of this statement is obvious when one studies the models of language production that have long been most influential in this field (e.g. Garrett, 1975; Levelt, 1989). The assumption here is that this research limitation has restricted our understanding of dialogical interaction to a significant extent; a misunderstanding that is only now, with the current abundance of research into dialogue, being rectified.

2.4 Clark's Collaborative Model

Before focussing specifically on feedback in dialogue, I will take a broader view of how partners interact in dialogue, and in particular, how speakers take account of their addressees' knowledge in dialogue (or fail to do so), even without the presence of feedback.

One model that has been extremely influential within the language-as-action tradition is Clark's Collaborative Model (Clark & Brennan, 1991; Clark & Schaefer, 1989; Clark & Wilkes-Gibbs, 1986). This section will give an overview of the main tenets of this theory: collaboration, common ground and grounding.

2.5 Collaboration

In the course of conversations which involve references to hard-to-describe objects, interlocutors must come to an agreement about what to call those objects. It is of course possible that such agreement might come about simply because one

partner dogmatically repeats a term with such frequency that their partner has little choice but to use the same word out of courtesy, but in everyday conversations where browbeating is not the norm, some concessions will usually be made by both speaker *and* addressee before they settle on a particular word or phrase to describe an entity. This process can tell us a great deal about how interlocutors work together to reach a common goal.

Producing a seminal paper in this field, Clark and Wilkes-Gibbs (1986) adapted the old Chinese game of Tangrams (which involves putting together geometrically-shaped cards to make pictures of people or animals) to create a type of referential communication experiment, a technique inspired by Krauss and Weinheimer (1966). Referential communication experiments typically involve one person (often called the director, speaker, leader or some variant on these) telling another person (the matcher, listener, addressee, follower) how to perform a task determined by the experimenter. This task often involves the manipulation of objects by the participants, with speakers telling addressees how to build complex items out of their components or how to repair items (for example, building Lego models by Clark and Krych, 2004; fixing bicycles by Kraut, Fussell and Siegel, 2003). A particularly common version of this paradigm involves speakers telling addressees how to arrange picture cards in a particular order (Krauss & Weinheimer, 1966) or pairs of participants playing a game involving cards (Hadelich, Branigan, Pickering & Crocker, 2004). Often the experimenter manipulates the partners' communication in some way, so that their visual communication is restricted (Hadelich et al, 2004), or the addressee is prevented from giving normal feedback to the speaker (Krauss & Weinheimer, 1966). On other occasions, the actual items used in the experiment are manipulated by the experimenter, so that the addressee cannot see all the items that the speaker can see (Keysar, Barr, Balin & Brauner, 2000), or the addressees' and speakers' items are not identical, unbeknownst to them (Chen & Krauss, 1991; Bard & Aylett, 2004). The referential communication task

has become a mainstay of dialogue research, and there are almost as many variations on task design as there are experimenters using it.

In Clark and Wilkes-Gibbs' study, the director had a particular arrangement of pre-made 'tangram' figures (images taken from Elffers, 1976) laid out in front of him and had to tell the matcher, who couldn't see either the director or his tangrams, how to arrange her own tangram figures in the same pattern. An example of a tangram is shown below in Figure 1.

Figure 1: Example tangram picture (Elffers, 1976)

Successful completion of this type of tangram-matching task largely depends upon the accuracy of the director's description of the tangrams. The advantage of using tangrams as experimental stimuli rather than, say, cartoons or photographs, is that tangram shapes are often quite ambiguous and difficult to describe, which makes the task more difficult and means it will take longer for the pair to settle on terms of reference.

Clark and Wilkes-Gibbs found that the pairs of participants typically worked together to create noun phrases to describe each tangram, adjusting, expanding or clarifying their descriptions until they reached a version that both of them were happy to accept. They referred to this process as *collaboration*. Clark and Brennan

(1991) propose that there are typically two recurrent phases to linguistic contributions in collaboration; firstly, a Presentation Phase, where a speaker puts forward a suggestion of some type, and then an Acceptance Phase, where their partner confirms their understanding of this suggestion. Often the acceptance phase actually involves the partner making another contribution at the presentation level, which still presupposes acceptance[6].

Not only do partners work together in collaboration to ensure quick, efficient communication, but Clark and Brennan suggest that speakers in a dialogue even intentionally plan their first descriptions to reduce the amount of effort required in the subsequent collaboration. Thus it is proposed that they follow the *Principle of Least Collaborative Effort:*

> "In conversation, the participants try to minimize their collaborative effort - the work that both do from the initiation of each contribution to its mutual acceptance" (1991, p135).

This principle does not mean that each partner will necessarily use the minimum possible number of words to describe objects. Rather, it might involve the speaker being somewhat more explicit than first seems necessary. If he does this, his addressee will be less likely to ask for clarification (saving her effort), so then the speaker will not have to expand into another utterance (thus also saving himself effort). This process could be seen either as a self-serving strategy or as a display of

---

[6] 'Acceptance' of a suggestion does not necessarily imply agreement with the sentiment, only that the partner understands the suggestion.

partner consideration, depending on how it is viewed, although it is possible that both aspects are involved (and see Brown and Dell (1987) for a discussion on how what looks like partner consideration may not always be so).

The idea of least collaborative effort partly concurs with Grice's (1975) maxims of quantity ("Make your contribution as informative as is required for the current purpose of the exchange, but do not make your contribution more informative than is required") and manner ("Be brief (avoid unnecessary wordiness)"). However, as Clark and Wilkes-Gibbs (1986) point out, it also takes into account three additional aspects of production which Grice is unable to account for: speakers' time pressures, errors, and ignorance. Firstly, speakers appear to limit the time and effort they are willing to spend on planning and producing an utterance. Secondly, speakers often make errors (for example, producing incomplete utterances, or sentences which are not grammatically correct) when they could easily spend a little more time and produce perfect utterances. Finally, sometimes speakers do not know enough to formulate complete descriptions or utterances.

For example, a speaker may not be able to remember the correct term for an item, and so may produce a sentence fragment in the hope that his addressee will finish it, as in the oft-quoted example below:

A. That tree has, uh, uh . . .
B. tentworms.
A. Yeah.
B. Yeah.

(from Wilkes-Gibbs, unpublished, reported in Clark and Wilkes-Gibbs, 1986).

The aspects of production mentioned above seem to suggest that speakers *don't* always provide the optimum amount of information in their utterances, either because they can't, or because they simply don't want to use as much as effort as

that would require. As Clark and Brennan point out, "Speakers often realize that it will take more collaborative effort to design a proper utterance than to design an improper utterance and enlist their addressees' help" (1991; p135), therefore the principle of least collaborative effort seems to be a more accurate explanation for language production than Grice's maxims alone.

2.6 Common ground and grounding

Clark and Schaefer (1989) proposed that successful coordination between speakers requires not only collaboration (which they refer to as 'content specification'), but also *grounding*. Achieving grounding involves speakers ensuring that their addressees have the same conceptual representations of certain entities that they themselves do (at which point the partners are said to be grounded). Grounding, according to the Collaborative theory, is the ultimate aim of collaboration, and tends to be most evident in the acceptance phase of Clark and Brennan's presentation-acceptance phase coupling.

Grounding, as Clark (1996) points out, can never be fully achieved between partners, because no-one can ever be quite sure that someone else understands *exactly* what they mean, in the same way as no-one can ever be sure that their concept of any given colour or sound is the same as the next person's. Every detail in any explanation will involve ambiguous terms, and these can only be explained further with more ambiguous terms, so two people can never be entirely grounded. The criterion for sufficient grounding that Clark and Schaefer (1989) suggest is that "The contributor and the partners mutually believe that the partners have understood what the contributor meant to a criterion sufficient for current purposes" (p262). It is clear, then, that interlocutors do not require absolutely

accurate grounding for successful communication, and that their criterion for grounding may vary from situation to situation.

Schober and Clark (1989) suggest that the mechanism by which partners typically attain grounding involves finding the first perspective on an item or concept that they both agree on, and then latching onto it; as they say, "Grounding is really an opportunistic process. It succeeds in part by exploiting adventitious commonalities between speakers and addressees" (p229). For example, imagine Alice and Peter can both see a particular tangram and have to decide on a term of reference for it. Alice thinks it looks like "An ice skater, or perhaps a chicken", while Peter disagrees, and would rather refer to it as "A ballerina doing an arabesque, or a hen". The partners' uses of the terms chicken and hen here represent some commonality in their perceptions; one that they are likely to latch onto, and so they are more likely to end up being grounded on the terms chicken or hen than ice skater or a ballerina.

Clark and Wilkes-Gibbs' study, as well as demonstrating collaboration and grounding between partners, also served to illustrate the building up of mutual knowledge or *common ground* (Stalnaker, 1978) between partners. Stalnaker describes common ground as:

> "Roughly speaking, the presuppositions of a speaker are the propositions whose truth he takes for granted as part of the background of the conversation... Presuppositions are what is taken by the speaker to be the *common ground* of the participants in the conversation, what is treated as their *common knowledge* or *mutual knowledge"* (1978, p320, Stalnaker's emphases).

Common ground is built up through the process of grounding; for example, as partners become grounded in their names for different items, these names, as they

are now mutual knowledge (that is, known by both partners, and known by each partner to also be known to the other partner), enter their common ground. Common ground is necessarily recursive in nature. Not only must each partner know a particular piece of information (for example, that Paris is the capital of France), but he must also know that his partner knows that Paris is the capital of France, and know that his partner knows that he knows that Paris is the capital of France, and so on. There must come a point, then, when the partners simply accept that the information they hold forms part of the common ground *enough to serve the current purpose*, similarly to their acceptance of grounding. In Clark and Wilkes-Gibbs' study, after each tangram had been described once, the directors tended to use more definite references than indefinite (for example referring to "the man who's praying" rather than "a man who's praying"). This demonstrated that both interlocutors presumed that, since they had previously discussed that term, they were grounded in it, and hence it became part of the common ground.

The idea of common ground relates not just to surface meanings of words and phrases like those described above, however, but also to referential aspects of this information. For example, when a speaker refers to "John" or "That holiday we took", he needs to know that his partner will know which John or which holiday is being referred to, and this confidence is determined by his knowledge of the common ground he shares with his partner. Clark and Marshall (1978) propose that common ground (which they refer to as *mutual knowledge*) between interlocutors arises in three main ways:

1) *Community membership*; when I say, "He threw the whisky down him like a true Scot", since my addressee and I both live in Scotland, we will both know of the Scots drinking stereotype.

2) *Physical co-presence*; when I say, "That blue car over there", my addressee can see the car I am referring to.

3) *Linguistic co-presence*; when I refer to, "The problem with my fridge" (or some referential term that we have collaborated upon), my addressee will understand it because we spoke of it earlier.

Reliance upon the common ground accrued by these means allows speakers to tailor their utterances to what they believe their partners know (although the extent to which they actually do this is debatable; see Horton and Keysar, 1996, and Wilkes-Gibbs and Clark, 1992 for two opposing examples), and, when used optimally, minimises misunderstandings between interlocutors.

A second important finding from Clark and Wilkes-Gibbs' study is that when speakers referred to particular tangrams several times, their descriptions tended to become shorter and less elaborate on subsequent references, which the authors considered to be a natural by-product of grounding. Since with repetition the partners became surer of the perspective they attributed to certain tangrams, and of their mutual understanding of that perspective, the number of words needed to describe those tangrams reduced. Research from Krauss and Glucksberg (1977) has also shown a similar effect of shortening on repeated references, and Krauss and Glucksberg (1969) demonstrated that young children are unable to produce useful descriptions in this type of context, but that they acquire this ability with age.

This shortening of descriptions is related to the formation of implicit agreements between partners (or *conceptual pacts*, in Brennan and Clark's (1996) terms) on how each item should be referred to. Conceptual pacts are marked by the occurrence of *lexical entrainment* (more of which later), which simply means that partners in the

dialogue tend to consistently use the same term each time they refer to the same item. This lexical entrainment ensures comprehension by both partners in later references to the same object. This may occur because since the first successful use of that term, the partners will presumably have associated the term with the object it was intended to refer to, and subsequent uses of the same term will then allow them to relate it to that item.

The idea of conceptual pacts is corroborated by findings from Wilkes-Gibbs and Clark (1992), who replicated the Clark and Wilkes-Gibbs (1986) study, but this time made the speakers change partners midway through the task, after each tangram had been referred to several times (and after the partners had typically decided upon a suitable description for each tangram). They found that when the speakers changed partners, even switching to people who the speakers *knew* had listened in on the first trial (bystanders), their tangram references became longer and less definite again. This is presumably because even though the new partners had overheard the initial collaboration, the speakers did not consider the common ground they had with the previous partners, or the conceptual pacts[7], to still be valid. Yet in those cases where the new partner had been a side participant[8] in the first trial (rather than simply a bystander), the speakers continued using shortened references, suggesting that they must have been perceived to have partner status.

---

[7] Wilkes-Gibbs and Clark didn't refer to conceptual pacts in this paper, but looking at later papers which introduce the term (e.g. Brennan and Clark, 1996), it seems to be applicable here.

[8] A side participant is someone who is part of a conversation, but is not being addressed at the present time. A bystander, in contrast, is someone who can hear what is being said, but does not have the opportunity to contribute, and as such is not part of the conversation.

2.7 Factors influencing production in dialogue: Cooperation and coordination

We have seen how the Collaborative theory proposes that speakers are affected by their addressees' behaviour in dialogue. This idea is not a new one (stemming back to Clark and Murphy, 1982), but the literature on this topic is far from clear. There are two main ways in which speakers' speech seems to be affected by their partners, besides in terms of content. *Audience design* refers to a speaker's designing of his utterances expressly in order to facilitate comprehension by his addressee.

The second way in which addressees affect speakers' productions is in terms of *priming* (or alignment), where the speaker tends to re-use those terms employed by the addressee, and vice-versa, often causing lexical, syntactic and even conceptual consistency between partners in dialogue. This is an apparently automatic process, in contrast with audience design, which is assumed to be intentional on the part of the speaker, at least to some extent. The occurrence of audience design and priming will on occasions result in the same effect (when, for example, re-use of your partner's lexical terms facilitates her comprehension), but as we shall see later, they can also exert independent effects (Haywood, Pickering and Branigan, 2005). I shall consider both effects in turn.

2.8 Cooperation in dialogue: Audience design

Clark, Schreuder and Buttrick (1983) proposed that speakers typically employ audience design by following the *Principle of Optimal Design*:

> "The speaker designs his utterance in such a way that he has good reason to believe that the addressees can readily and uniquely compute what he meant on the basis of the utterance along with the rest of their common ground" (p246).

In practice, optimal design involves the speaker taking account of the common ground that exists between himself and his addressee, and adjusting his speech accordingly. There is evidence in the literature that this occurs in some ways (e.g. Clark & Wilkes-Gibbs, 1986; Fussell & Krauss, 1992; Isaacs & Clark, 1987; Wilkes-Gibbs & Clark, 1992). But there are many types of common ground that can potentially come under the consideration of the speaker; Schober and Brennan (2003) detail the many levels of audience design that exist, from the particular to the general: specific-partner adjustments (where I know what my friend knows), cultural/community/group-based adjustments (where I know what a member of my community or group is likely to know) and generic partner adjustments (where I know what information any other person is likely to need in order to understand me). At the very end of this spectrum lies egocentric processing, in which the speaker makes no adjustments for his addressee at all. This section will consider only the first two types of audience design mentioned above: specific-partner and group-based adjustments.

2.8.1 Adapting to addressees' group-based status

We saw earlier how partners in dialogue come to agree upon reference terms for new items, through a process which Clark and Wilkes-Gibbs (1986) referred to as collaboration. But what happens when two partners differ in the amount of knowledge they bring to a discussion simply by virtue of their cultural or group status; how does this affect the speakers' descriptions of items?

Isaacs and Clark (1987) ran a referential communication task where pairs of participants worked together to arrange pictures of famous landmarks in New York City. The pairs of participants varied according to whether they were familiar with the landmarks (experts) or not (novices), and Isaacs and Clark found that the task was completed most quickly when both partners were experts. More crucially, it was found that those participants who were familiar with the landmarks adjusted their descriptions according to their partners' level of expertise; they used significantly fewer proper names of landmarks when the addressee was a novice than when she was an expert. What is of particular interest here is that the speakers appeared to use their classifications of the addressee's expertise on landmarks which were seen *early* in the experiment to generate expectations about the likely familiarity of *later* landmarks, and so Isaacs and Clark concluded that even the terms employed during initial references to objects in dialogue can be affected by how much the speaker thinks his addressee knows (rather than relying upon feedback from the addressee on his description of that particular landmark).

In a similar vein, Krauss and Fussell (1996) report that Kingsbury (1968, unpublished) asked pedestrians on the street in Boston for directions to a nearby department store. He varied both the accent he spoke with (either Boston (local) or Missouri) and whether he said that he was from out of town or not. Not surprisingly, when he mentioned to his directors that he was from out of town,

the directions given were both longer and more detailed than when he did not mention this. However longer directions were also given when he did *not* mention being from out of town, but simply spoke in the Missouri dialect. It seems that respondents in this case presumed that his level of local expertise was low, and adjusted their directions accordingly, to include the additional information that they deemed necessary in this case.

2.8.2 Adapting to addressees' specific knowledge about the situation

As well as making adjustments to addressees' familiarity with the topic under discussion, there is evidence to suggest that speakers also take account of what information their addressees have access to in the current task. Horton and Gerrig (2002) ran a picture-matching task on pairs of participants, where some items were visible to both partners (in common ground) and some were visible only to the speaker (in privileged ground). The speakers changed addressees twice during the task, and it was found that after they had changed addressees for the *second* time, they better tailored their utterances to the needs and knowledge of their addressees (that is, employed better audience design), by not referring to the items that weren't visible to them. Horton and Gerrig proposed that the speakers learnt how to adapt to their addressees during their first partner change. They concluded that it is important to consider speakers' experience (for example, if they have previously discussed the same topic with the same partner) when it comes to assessing audience design, because, they argued, there are two things that must occur before speakers can implement this process: Firstly, they must realise that audience design would be helpful in the present circumstances, and secondly, they must overcome their natural tendency towards consistency of expression. Partner feedback should

ideally alert speakers to the necessity of these tasks[9]. It seems reasonable to presume that it was the speakers' tendency towards brevity that made them *not* employ audience design in the early stages of this experiment, and that a failure in communication led them to realise it was necessary, and eventually use it. That is, there was a delay in employing audience design, but it was finally implemented in response to feedback.

Brown and Dell (1987), rather than investigating how speakers took into account their addressees' knowledge of physically present objects, studied their knowledge of implicitly present objects; that is, the information that is often inferred in a situation. This study involved a story-telling task: speakers silently read a story and saw the accompanying picture, and then re-told the story to another person in their own words. The main manipulation in this experiment related to the instruments used in the stories. These were either 'typical' instruments for the action portrayed (for example, to stab with a knife) or 'atypical' ones (to stab with an ice pick). It was found that those speakers whose stories included atypical instruments specified the instrument used more often than those whose stories included typical instruments. Brown and Dell suggested that one possible reason for this was that in the typical instrument conditions, speakers could have presumed that a certain level of inferred knowledge was held by the addressee, that is, the act of stabbing is very likely to involve a knife, so in this case the instrument didn't need to be mentioned.

Although this explanation would certainly be consistent with theories of audience design in production, a second experiment showed it to be unlikely. Brown and Dell suggested that if speakers are truly taking account of their addressee's knowledge then their storytelling should change to reflect what they think their

---

[9] This is an idea we will return to in Section 2.12.4.

addressee knows. To test this, they varied whether or not the addressee could see the speaker's picture, and studied how this affected the speaker's descriptions. This manipulation did make some difference to the explicitness of the story-telling; for example the speakers tended to mention the instruments in separate clauses after the verb (rather than in the same clause) when the addressees couldn't see the picture. However it did not explain why typicality still had an effect; there was a bias towards speakers mentioning the instrument in a separate clause when the instrument was atypical rather than typical, even when the addressee could see the picture. Brown and Dell therefore concluded that the results from the first experiment were *not* a result of the speakers' thoughts about their partners' knowledge, but were in fact determined by the way typical and atypical information is represented in the speakers' conceptual structuring of the story. The crucial point of this paper is that what looks from the outside to be partner adaptation might not be any such thing; it could simply be reflecting the easiest production strategies for the speaker, and so it is important for researchers on this topic to find methods of distinguishing between these two accounts.

Brown and Dell's paper was recently strongly criticised on methodological grounds by Lockridge and Brennan (2002), who objected to their use of confederates as addressees[10]. They suggested that this might have significantly affected the results obtained, on the basis that confederates, when they had heard the story repeated many times (up to 40 times in this case), would not have produced typical feedback, since they in fact would have known the story better than the speaker by the end. To test this theory, Lockridge and Brennan replicated this experiment, but used naïve participants as the addressees, rather than confederates. Similarly to Brown and Dell's study, they found that speakers mentioned atypical instruments

---

[10] See Section 3.2.2 for further discussion on the necessity of using naïve participants as addressees.

significantly more often than typical instruments, and mentioned both typical and atypical instruments more often when the addressee didn't have pictures. However, contrary to Brown and Dell, they found that speakers were most likely to mention instruments *early* in their utterances when the instruments were atypical and when the addressee didn't have a picture. That is, the instruments were usually mentioned in the same clause as the verb (as opposed to in separate, later clauses) when the speakers knew the addressees couldn't see the instrument. This may suggest that common ground was taken into account at a relatively early stage of planning. Also, atypical instruments were more likely to be marked as *new* (in terms of using indefinite articles, such as 'a', rather than definite articles such as 'the') than typical instruments, indicating that some calculations regarding what the addressees were likely to have inferred had taken place. In contrast to Brown and Dell's experiment, then, the results from Lockridge and Brennan's study supported the use of audience design by speakers.

It is possible that even some aspects of speech that look like shortcuts for the speaker are still produced with a listener in mind. Jucker, Smith and Ludge (2003) suggested that even something as apparently egocentric as vagueness (that is, non-specificity) may actually be produced with the listener in mind, despite simply looking like a reflection of uncertainty on the part of the speaker. They proposed that the use of vagueness allows speakers to indicate to their listeners how much processing effort they should allocate to a particular idea or entity, and as such can be an intentional strategy on their part.

2.8.3 When is common ground used?

If common ground is used to help speakers design their utterances, at what point in the production process does this come into play? Although a number of researchers (e.g. Lockridge & Brennan, 2002) support the idea of audience design occurring

early in the production process, an opposing standpoint would claim that it only occurs as a repair, after the main body of the utterance has been planned. This opinion is held by Horton and Keysar (1996; Keysar, Barr, & Horton, 1998; also Brown & Dell, 1987), who found that when speakers had no time constraints on their productions, they took common ground into consideration, but when under time pressure they tended to ignore it. Horton and Keysar's explanation for this was that common ground isn't taken into account during the initial planning of utterances, but that violations of common ground (where, for example, a person might refer to an object that their partner can't see) are weeded out by the self-monitor, whose action is restricted by lack of time.

It is possible that speakers' tendencies *not* to employ audience design when under time restriction may be representative of the wider effect of cognitive load on audience design. One way of investigating if this is true would be by altering the cognitive load in some other way and testing for the use of audience design. Buhl (2001) assessed how participants' use of self- or other-perspective varied with the difficulty of a particular task (also see Roβnagel (2000) for a similar paradigm and concurrent finding). Speakers had to describe routes in a video simulation of real streets, and Buhl found that they were more likely to give other-oriented (i.e. considerate) descriptions to their partners in easy tasks than they were in more difficult, attention-consuming tasks. He suggested that this difference in perspective use was due to a reduction in conceptual abilities when cognitive resources are limited. Additionally, Schober (1993, discussed in detail later) found that speakers took their partner's perspective in a spatial orientation task less often in dialogue than in monologue. This clearly concurs with the above findings, since it could be imagined that the concurrent need to comprehend a partner's speech

would increase the cognitive load on a speaker, possibly leaving fewer resources for audience design[11].

Bard, Anderson, Sotillo, Aylett, Doherty-Sneddon and Newlands (2000) and Bard and Aylett (2004) reached a conclusion that would show similar experimental results to Horton and Keysar's (1996) theory, but with a different underlying mechanism. They propose that language production involves two parallel processes: a fast, automatic element which doesn't take account of common ground, and a slow, more controlled element, which does consider it. If the primary, automatic element production is slow enough for the second, slower system to adjust the utterance, then speakers will be seen to employ audience design; otherwise, they will produce egocentric utterances.

As we have seen above, one way in which speakers employ audience design is by adjusting their utterances according to what they think their partners know on a general level. However they must also adjust them according to the response from the addressees. When this happens, it allows the addressees to control the utterance in a way, by using what we might refer to as 'addressee control'. Take as an example a tangram-matching task, where the speaker is describing a particular tangram to his addressee, to enable her to identify it from a selection. Part-way through the speaker's sentence, the addressee interrupts with, "Wait! I've got it. Right, go on to the next one". If the speaker obeys, then in a sense he is employing audience design, in that he is adjusting his description according to what his addressee knows. In another sense, though, it is the *addressee* who is imposing this

---

[11] Alternatively, this result could be a result of the speaker's diminished responsibility in dialogue where he knows his partner can query anything she doesn't understand, and so he does not need to design optimal utterances from the outset.

upon him, by directly controlling the length of the description. Another situation might involve the speaker finishing a description, and then the addressee asking for more information on a particular aspect of it, resulting in him expanding the description to meet her needs. Again this allows the addressee to exert control over the description. These coordinating aspects of *audience design* and *addressee control* highlight the extent to which interlocutors work together in dialogue. Not only that, but they also demonstrate that the speaker (who would typically have the more dominant role in this situation) doesn't necessarily control every aspect of a successful dialogue, but must allow the addressee to influence both the timing and the content of his productions in a potentially very significant manner. Any researcher wishing to examine the implementation of speakers' audience design must also take into account the addressees' control, and successfully separate these two factors, perhaps by curtailing the communicative ability of the addressee in some circumstances[12].

2.9 Coordination in dialogue: Priming

Distinct from the process of audience design is the occurrence of *coordination* in dialogue[13]. Many studies have noted speakers apparently unconsciously altering their manner of speaking in order to become more similar to their partners' speech, a process which is sometimes known as *convergence* or *accommodation*. These results

---

[12] Although the resulting lack of feedback may also affect the speaker's speech in unnatural ways; see Section 2.12.2 for details.

[13] Pickering and Garrod (2004) noted that in the literature, the idea of coordination has been used to refer not only to partners sharing the same representations at a particular level, but also to partners working together to achieve a goal; we use it in reference to the first notion only here, in keeping with Haywood, Pickering and Branigan (2005).

lie in the mould of Giles' (1973) *Speech Accommodation Theory*, which focussed on the seemingly unconscious convergence and divergence of accents for social reasons. It seems that this theory may also be applicable to other aspects of speech; for example, fundamental frequencies are affected by both social status (Gregory & Webster, 1996) and gender (Bilous & Krauss, 1988), and the convergence of vowel placements in particular shows gender and task role effects (Pardo, 2006). This social convergence is not specific to speech; Chartrand and Lakin (2003) have demonstrated that people mimic the behaviour of their partners (the 'chameleon effect': in this situation, face-touching and feet-shaking) to a greater extent when they are attempting to build up rapport with their partners than when they are not.

Another way of looking at this type of convergence is as *priming*. The term priming (in this context) refers to a speaker's apparently automatic re-use of linguistic elements that they have previously heard, produced, written or read (e.g. Bock, 1986; Potter and Lombardi, 1998; Branigan, Pickering, & Cleland, 2000; Cleland & Pickering, 2006). This may involve the repetition of syntactic forms, lexical items or conceptual structures, and seems to improve speaker efficiency and facilitate faster language production (Hartsuiker & Kolk, 1998; Pickering & Branigan, 1999; Smith & Wheeldon, 2001). Priming has been found to occur with syntactic structures (Branigan, Pickering & Cleland, 2000) lexical terms (Wheeldon & Monsell, 1992), conceptual representations (Garrod & Anderson, 1987), spatial reference frames (Watson, Pickering & Branigan, in press), articulatory structure in terms of speech rate (Bard & Aylett, 2004), and possibly even inter-turn pauses (ten Bosch, Oostdijk & Boves, 2005[14]). This section will mention a handful of the most notable findings.

---

[14] The authors suggest that another possible explanation for accommodation in inter-turn pauses might be the overall character of a dialogue (rather than priming between partners); they do not distinguish between these two possibilities.

2.9.1 Syntactic priming

One of the earliest indications of priming came from a 1982 study, where Levelt and Kelter telephoned shopkeepers in the Netherlands and asked them (in Dutch), either:

a) What time does your shop close?
or
b) At what time does your shop close?

They found that the answers given tended to follow the form of the questions; that is, the response to a) was typically something like 'Five o'clock', whereas b) would have been answered 'At five o'clock'. This certainly appears to be priming of some kind, although whether it is syntactic or lexical is not quite clear; Levelt and Kelter simply referred to it as repetition of the 'surface form'.

Following this, the first *explicit* description of syntactic priming in production came from Bock (1986), whose seminal paper drew attention to the occurrence of syntactic priming in monologue in active/passive and prepositional object/double object contexts. She found that these sentence structures tended to be produced more reliably following the production of sentences of similar structure by the same speaker, and subsequently showed that this was not reliant upon the repetition of lexical items (Bock, 1989), or the metrical form (Bock & Loebell, 1990). Bock and Griffin (2000) showed that similar syntactic priming of passives and datives can persist over as many as 10 intervening sentences.

Based upon those initial studies in monological production, the scope of this topic has only relatively recently broadened to include the analysis of priming between partners in dialogue. Branigan, Pickering and colleagues (Branigan, Pickering & Cleland, 2000; Branigan, Pickering, McLean and Cleland, in press; Cleland &

Pickering, 2003) have run a series of experiments using an innovative *confederate priming* paradigm to investigate between-partner priming. This type of experiment involves a pair of participants, one of whom is secretly a confederate (or 'stooge') of the experimenter. In one variant, the two participants are seated at different computers, out of view of each other, and take it in turns to describe cartoon pictures to each other. Unbeknownst to the real participant, the confederate does not see pictures on her screen; instead, she reads out a sentence given to her by the experimenter, apparently describing a picture. This gives the experimenter control over what the confederate says, whether it be making her produce a double object or prepositional object structure (to study syntactic priming), or use a particular lexical item (lexical priming). The real participant then describes a picture that could potentially be described using the same linguistic expressions (words or structures) as the confederate's description. He must use the given verb, and in the experimental trials he will have to choose what words and structures to use; for example, he could call the 'cook' either a cook or a chef, he could say 'the soldier threw the ball to the boy' (PO) or 'the soldier threw the boy the ball' (DO) and so on. The key finding from this paradigm is that the confederate influences the participant's speech in terms of syntactic structures and lexical items, thus demonstrating priming from comprehension to production between partners (e.g. Branigan, Pickering & Cleland, 2000; Branigan, Pickering, McLean & Cleland, in press).

2.9.2 Lexical priming

Not only do speakers repeat syntactic structures, but they also re-use lexical items that were introduced by either themselves (Brennan & Clark, 1996) or their partners (Cleland & Pickering, 2003). Brennan and Clark (1996) demonstrated the occurrence of lexical priming, or *lexical entrainment*, as they termed it. They carried out a referential communication study where a speaker had to describe to his addressee

how to arrange a number of picture cards (a combination of tangrams and pictures of everyday objects) into the right order. The experimenters found that the speakers tended to re-use the same names for the pictures again and again, indicating that they had made an association between the picture and the name, and so had become *entrained.* Brennan and Clark further investigated the cause of this entrainment by analysing speakers' references to pictures when the context of those pictures changed from one reference to the next. After the first experimental trial, when all the cards had been described several times, the experimenters changed the speaker's set of cards to another set, which contained only some of the same cards as the first trial; the rest were different. In one set of cards, the 'target cards' (the cards which the experimenters were focussing upon) were 'nonunique'; that is, there were other cards which looked similar in the same set, for example two or three items that all belonged to the category of shoes. In the second card set, these same target cards were 'unique', meaning that there were no similar cards in the same set. Brennan and Clark compared the names used by each speaker for these target cards when they saw them in first the non-unique and then and the unique context. They found that speakers tended to use the previous, now overinformative descriptions even when that level of detail was no longer necessary (that is, when the cards were seen in the unique context as opposed to the previous non-unique context), demonstrating lexical entrainment and the subsequent formation of conceptual pacts with addressees. They also found that the strength of the entrainment or conceptual pact varied with the previous frequency of use of that name by the speaker.

The entrainment that was demonstrated here appeared to be somewhat specific to the current partners in dialogue. Brennan and Clark found that descriptions of pictures were typically shortened with repetition, but that a given referring expression was more likely to be lengthened again when the speaker was addressing a new partner than when they continued addressing their current partner, in keeping with Wilkes-Gibbs and Clark (1992). They proposed that the

social interaction between partners had a significant effect on the formation of conceptual pacts, and that these pacts were specific to the people they were formed with, hence when a new partner was addressed, the pact was no longer valid. More recently, Horton and Gerrig (2005) also suggested that this type of partner-specific effect could be considered as simply a feature of normal memory processes, where descriptions are associated with particular partners, so that the presence of a particular partner cues the speaker's production of the description spoken to that same partner previously. However, Pickering (2005) and Barr and Keysar (2002) pointed out that in Brennan and Clark's experiment, the speakers often initially used the previous term with their new interlocutors, only altering this term upon receiving feedback from their partners. It is plausible, then, that they took less consideration of their addressees' needs, and were more influenced by their prior entrainment, than was initially implicated by these results.

Barr and Keysar (2002) examined the response of addressees to speakers' references, and took the opposite viewpoint to Brennan and Clark, in suggesting that descriptions are independent from partners in conversation. They used a referential communication task where the speaker was a confederate, and he used particular reference terms to refer to items during the task, resulting in an entrainment effect. Barr and Keysar proposed that when addressees entrain on reference terms with a speaker, and then a *new* speaker uses the same reference term to refer to the same object, there should be inhibition of the addressee's response to this term (for example, they might look at or reach for the object more slowly). They found no such inhibition, and concluded that the entrainment does not rely upon partner-specific representations.

Evidence to challenge this proposition comes from Metzing and Brennan (2003), who suggested that partner-specific pacts should not cause difficulty when a new speaker used an old term (as Barr and Keysar tested), but only when an *old*

speaker used a *new* term, as this would represent the speaker breaking a pact. Metzing and Brennan demonstrated, using an eye-tracking referential communication paradigm, that the comprehension of object names was more difficult for addressees when they heard a speaker change their manner of referring to an object (for example, calling an object a 'shiny cylinder' rather than the previous 'silver pipe') than it was when they heard a *new* speaker using the new referring expression (or the old expression, supporting Barr and Keysar's findings), or the old speaker using the old expression. That is, it seems that referring expressions are specifically associated with speakers, and when the speaker changes, the present conceptual pact is in effect wiped clean by *both* the speaker and the addressee, ready for a new pact. Additionally, evidence that conceptual pacts may be specific to *people*, rather than to conversations comes from Markman and Makin (1998), who found that pacts between partners can persist for as long as 5 days. Malt and Sloman (2004) also found that the pacts formed between partners affect long-term naming strategies, where the names that had been decided on between partners were more likely than not to be chosen as the individuals' 'preferred names' in a later typicality assessment.

Despite the apparent automaticity of this entrainment effect, sometimes a determination *not* to align with a partner on lexical terms can be used to make a social point; Danet (1980) discussed the differential use of terms in the legal prosecution of a doctor who had carried out a late abortion, reporting the use of the terms *baby* and *foetus* by prosecution and defence respectively. These two terms differ notably in their implication of life (or lack of), a factor which was important to the trial, and Danet concluded that entrainment on reference terms can be consciously overridden when the speaker is determined to do so for social reasons.

2.9.3 Conceptual priming

Garrod and Anderson (1987) also investigated how partners end up using consistent methods of description, but they focussed on the priming of conceptual representations along with lexical terms. They employed a maze game, where participants had to describe the positions of various objects in a maze in order to complete a task. In their game, it was found that after a certain number of trials, pairs of speakers and addressees tended to use the same types of descriptions for points in the maze ('rows', 'lines' and so on), and the authors proposed that they did this by invoking an 'input-output co-ordination strategy'; that is, by making each of their outputs use the same referring terms as the most recent input (similarly to Brennan and Clark's historical explanation for lexical entrainment). So if one person referred to an object as being 'on the third row down, two cells across', the next reference from their partner was also likely to refer to the horizontal lines as rows, and the vertical ones as cells. This involves agreement on not only the terminology used, but also on the general schemas employed by the partners, and their interpretations.

Garrod and Doherty (1994) carried out a similar maze task between pairs of participants, but in addition, asked them to change partners between games. They found the same initial settling on terms between partners that Garrod and Anderson found, but also found that when people began a game with a new partner, they tended to carry over the same reference terms, until, after a number of partner changes, the majority of the 'community' (in this case, 8 people) were more likely to employ the same kind of references. This wasn't just a coincidence, where people just used the most obvious set of terms; the isolated pairs who didn't swap partners used a larger range of descriptions, and didn't end up using the same descriptions as their neighbouring pairs. Garrod and Doherty concluded that their study

showed that interaction between people results in coordination convergence on a common description schema.

Fay, Garrod and Carletta (2000) showed that convergence between partners (of opinions here rather than reference terms) is affected by the size of the group; small groups (<7) behave more like dialogues, with individual participants being influenced most by those they interact with, whereas in larger groups (>7), participants are influenced more by dominant speakers in the group, making it more like a monologue situation.

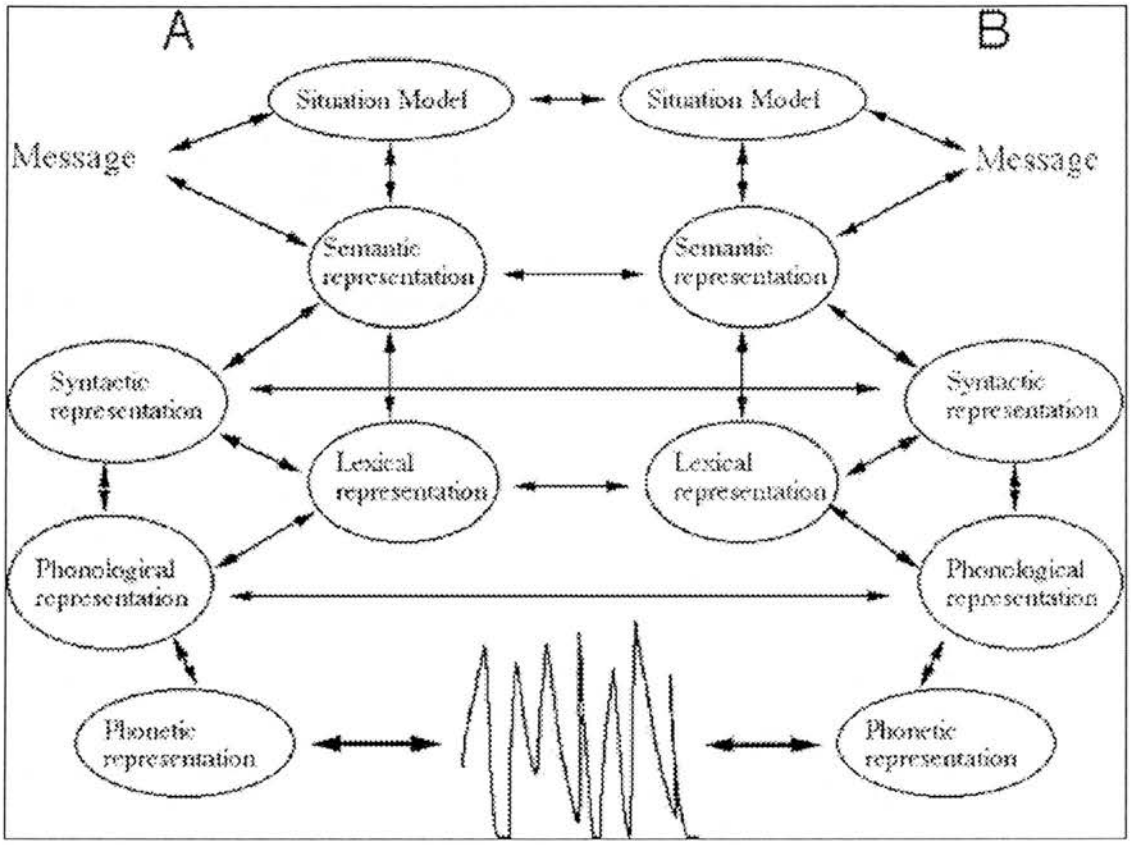
2.10 Pickering and Garrod's Interactive Alignment Model

Where the idea of cooperation in dialogue is linked with Clark's Collaborative model, the occurrence of priming is straightforwardly linked with Pickering and Garrod's (2004) Interactive Alignment model. This was developed from the input-output coordination theory of Garrod and Anderson (1987) and Garrod and Doherty (1994), although it uses the term *alignment* rather than coordination. This theory differs markedly from the Collaborative model: Clark and colleagues would attribute the above effects of addressee speech on speakers to the use of explicit common ground, where each partner holds a mental model of their partner's knowledge and beliefs; a model which is constantly updated upon the introduction of new information, and which the speaker refers to at every point before planning his speech. In contrast with this, Pickering and Garrod propose that speakers do not hold representations of mutual beliefs and knowledge, and that some of what we view as their audience design arises from the automatic alignment of representations between partners. This negates the need for complex, resource-expensive partner models, and relies instead upon automatic processes which should be less cognitively demanding upon the interlocutors. However, the

outcome would hypothetically be the same as if common ground were being consulted, as both partners' productions will employ the same representations. Pickering and Garrod propose that production and comprehension are linked at each level within the language system, and that once two interlocutors have become aligned at one particular level (via Garrod and Anderson's input-output coordination), this alignment will automatically percolate upwards to other levels in the system, for example, from lexical to syntactic, and syntactic to conceptual.

Evidence to support this theory comes from the finding that priming between different levels of production is linked: syntactic priming between speakers exerts a stronger effect when the verb in question is repeated between prime and target (Branigan et al, 2000), and syntactic priming in noun phrases is also increased when the nouns in the prime and target are either semantically or phonologically related to each other (Cleland & Pickering, 2003). Both these findings suggest that alignment at one level of the language production model causes (or at least facilitates) alignment at other levels. Figure 2 demonstrates the structure of the model.

Figure 2: Interactive Alignment Model (Pickering and Garrod, 2004)



According to this theory, the ultimate goal of cooperation in conversation is for the interlocutors' alignment to percolate up to the uppermost stage – that of the situation or mental model (cf. Johnson-Laird, 1983; Zwaan & Radvansky, 1998). Pickering and Garrod describe a situation model as 'a multi-dimensional representation of the situation under discussion' (p4), and this encompasses all the features of the current conversational environment, such as time, space, causality, intentionality and the contents of the current dialogue. That is, it takes account not only of what is said during the discourse, but also of the entire social context. Once interlocutors' situation models are aligned by this mechanism of upward-percolating alignment, they can be sure that each of them will understand the other's references to the necessary degree; in Clark's terms, they will be grounded. This means that it is not necessary for each speaker to model his interlocutor's

situation model, because his and his interlocutor's model are essentially the same. Thus the Interactive Alignment model, unlike the Collaborative model, does not propose that speakers intentionally design their utterances for their addressees' benefit. Rather they suggest that a speaker's aim is to align his situation model with his partner's, and the fortunate consequence of this is that his utterances will be easy for that partner to comprehend. What appears to be audience design, then, may just be egocentric processing on the part of the speaker, harking back to discussion by Brown and Dell (1987).

Pickering and Garrod stress that alignment between partners is a base level strategy; that is, it is the simplest and most common tactic, and that there will be obvious exceptions to this case, such as when one person is deceiving the other (in which case they will not be aligning at some level), or when a person refuses to align for some reason (for example, if they disagree strongly with the content of what their partner is saying, as in Danet, 1980, although in this type of case, *some* alignment of concepts will still be necessary to ensure that both partners are speaking about the same thing). Pickering and Garrod also note that the default performance can be adjusted by social factors that enhance or reduce the tendency towards alignment, which explains, for example, the audience design reported in the studies above (e.g. Kingsbury, 1968, in Krauss and Fussell, 1996). They point out that, "In general, we suspect that speakers make a one-off decision based on such issues as the perceived expertise of their addressees about how to frame their contributions...[but] Such decisions need not remain fixed for the whole conversation...such a change is very different from a continuous, dynamic process of utterance accommodation based on full common-ground inference" (2004, p48). This obviously differs substantially from Clark and colleagues' employment of audience design, which arises as a result of a listener model which is continually referred to during the course of a conversation. Additionally, Pickering and Garrod propose that performing inferences about common ground is, "...an optional strategy that interlocutors

employ only when resources allow" (p11, 2004), which seems to very broadly concur with Horton and Keysar's (1996) Monitoring and Adjustment theory.

It certainly seems that alignment may be more of an automatic effect than audience design, and this is the viewpoint put forward by Pickering and Garrod (2004), Bard et al (2000), and Bard and Aylett (2004). Bard et al found that speakers' speech (in this case, relating to the articulatory length of their noun phrases) is unaffected by social factors such as what their addressees have heard or seen, even when the addressees give explicit feedback confirming their state of knowledge. However some recent findings, reported below, suggest that more partner-specific effects might be found in the context of priming and alignment than was first assumed.

Branigan, Pickering, McLean and Cleland (in press) demonstrated that on occasion priming *can be* influenced by social factors; they found that people were less syntactically primed by a previous speaker when they were 'side participants' in a referential communication task (that is, they were considered to be part of the task, but were not contributing at that particular time) than when they were addressees of the speaker. Branigan et al concluded that syntactic priming was *not* confined to solely linguistic influences, but that context (in this case role) could exert a real effect[15]. A further demonstration of social factors in priming comes from Branigan, Pickering, Pearson, McLean, Nass, and Hu (2004), who found that speakers tend to demonstrate more syntactic priming when they believe they are interacting with a computer than with a human being who is communicating via a computer. Similarly, Pearson, Hu, Branigan, Pickering and Nass (2006) found a related result in the lexical domain; participants were more inclined to align on lexical terms with

---

[15] In this situation, the authors surmised, the role differences occurred because addressees of an utterance have to process that utterance more deeply than side-participants, causing greater priming.

a computer which appeared to be unsophisticated and old-fashioned than one which appeared to be sophisticated and capable (note: the computer behaved in exactly the same manner in both cases; only the interface was different). Pearson et al (2006) attributed the results of both these studies to a belief of the participants that they had to cater to the computers' restricted language ability. It seems, therefore, that it is possible for priming to be affected by certain characteristics of an interlocutor, although the extent to which this happens in face-to-face (as opposed to computer-based) interaction is as yet unclear.

2.11 Coordination and cooperation; Competitive effects?

Sections 2.8 and 2.9 have outlined research on two ways in which speakers' speech can be affected by their partners' speech. Cooperation (or audience design) involves the speaker designing his utterance in a way that he believes will facilitate his addressee's comprehension. Coordination, in contrast, involves the speaker's speech becoming more like that of his addressee. In a large proportion of cases, these effects will overlap; for example, people find it easier to comprehend linguistic elements similar to those they have just produced (Branigan, Pickering and McLean, 2005), and so coordination will also ensure cooperation. However in some circumstances, these effects will pull in opposing directions. For example, people might at times need to resist the temptation to align with their partners, if communication is to be successful. Garrod and Clark (1993) found that seven year old children found it difficult to introduce new maze description schemes when they needed to, because they couldn't resist aligning with the previous descriptions. On this occasion, then, cooperation and coordination had differing effects, and a more cognitively-advanced speaker would be expected to override the pressure to align in order to implement audience design.

In a direct comparison between the effects of audience design and priming, Haywood, Pickering and Branigan (2005) carried out a referential communication task which involved participants alternating in giving instructions for their partners to move objects around a grid. Haywood et al took advantage of the ambiguity of sentences such as 'Put the penguin in the cup on the star', which, in an array containing a penguin standing in a cup, a second penguin standing alone, and an empty cup, could have two possible interpretations (either put the penguin which is standing in the cup onto the star, or put the penguin who is standing alone into the cup, and then move both objects onto the star). The presence of *that's* in such a sentence has a disambiguating effect ('Put the penguin *that's* in the cup on the star'), and on half the occasions (crossed with the ambiguity of the display), one of the participants (a confederate) used *that's* to disambiguate his utterances. The experimenters analysed whether the real participants tended, firstly, to be primed by the confederates (resulting in their use of *that's* following a similar description from their partners) or secondly, to use audience design (resulting in their use of *that's* in ambiguous contexts, that is, where there were two possible interpretations, as opposed to contexts where there was only one).

Haywood et al concluded that both effects occurred independently, although the priming effect was more reliable than the audience design effect. These results on audience design do however contrast with other studies that fail to show speakers employing optional disambiguating words in ambiguous contexts (e.g. Kraljic and Brennan, 2005, in a similar task; Ferreira and Dell, 2000, in a sentence recall task).

The type of situation described in the above experiment is somewhat unusual. It is likely that in the majority of cases, audience design and alignment would predict the same effect, as mentioned previously. Perhaps this is the key to understanding this area of speech production; it is probable that the architecture of the language system is designed so that what is easiest for speakers to produce is also easiest for

their addressees to comprehend in the majority of cases, and that it is only in unusual situations that these two factors will diverge.

2.12 Feedback

2.12.1 Production of feedback

If a speaker is to take account of his addressee's needs, he must first be aware of those needs. The most obvious way for an addressee to impart her state of knowledge or understanding to a speaker is by simply telling him, by means of producing backchannel responses or feedback (which may refer to the same thing). Backchannel responses are defined as, "...brief vocal responses ('uh-huh', 'yes', 'I see', etc.) by the nominal listener, which do not constitute an attempt to take the conversational floor"[16] (p186, Bilous & Krauss, 1988). Heinz (2003) gives a comprehensive review of the descriptive literature on backchannel responses, listing common types of verbal feedback, such as sentence completions, requests for clarification, exclamations, attempted interruptions and so on, along with such non-verbal responses as head nods and shakes, smiles, shrugs, gazes and many more.

Heinz's own study demonstrated that backchannel responses differ from language to language; she found that they are more prevalent in American English than in German. Bilous and Krauss (1988) found that female Columbia University students

---

[16] Some researchers also include short questions and requests for clarification in their definition of backchannels (e.g. Kendon, 1967, quoted in Heinz, 2003); for the purposes of the current review, we will examine the role of 'feedback' as a whole, which includes, but is not restricted to, backchannel responses.

produced more backchannel responses than their male counterparts, and Feke (2003) also found gender differences in the production of backchannel responses (in both English and Spanish), with males producing more backchannel responses than females in same-sex dialogues, but females using more than males in mixed-sex dialogues. It is clear, then, that the distribution and use of backchannel responses is in no way universal. Heinz mentions the many purposes of backchannel responses reported in the literature: amongst them, signalling attention, showing involvement, acknowledging ongoing telling, indicating agreement or expressing disagreement or lack of understanding.

The definition of feedback in the literature is less clear than that of backchannel responses, in that it is not apparent whether this term should also encompass questions from the addressee (which in some circumstances may involve them taking the floor temporarily). For the purposes of this thesis my definition of feedback will be more inclusive than that of backchannel responses. Feedback will be considered to include *any audible* noise produced by the addressee and heard by the speaker [17]. This incorporates backchannel responses, and also includes asking questions, making comments, and producing indistinct sounds (for example 'uh', or the audible movement of experimental items). These noises can be intentional (like a comment) or unintentional (like a sneeze). It will *not* include visual feedback, like gestures, facial expressions and so on. Hence a dialogue, for my purposes, will involve full audible feedback, with no restrictions on what either participant can say, what kinds of sounds they produce by other means, or what they can hear, but

---

[17] Visual feedback was eliminated completely because while assessing the amount of auditory feedback is relatively straightforward (in terms of number of words produced etc), visual feedback is much more complicated to measure, and would be better assessed by a study which set out to investigate it specifically. Additionally, this thesis purports to study only language, rather than dialogue behaviour as a whole.

will include no visual feedback[18]. A monologue will involve none of these, so in an ideal monologue situation, the speaker would not be able to hear *any* sounds produced by the addressee, nor receive any visual feedback[19].

The feedback described above contributes in part to the to-and-fro, turn-taking aspect of dialogue that distinguishes it from monologue (studied in detail in Sacks, Schegloff and Jefferson, 1974; Schegloff, 2000), but its use does not stop there. Krauss and Fussell (1996) propose that there are two main benefits to feedback in dialogue. The first is that it allows the speaker to constantly measure his addressee's state of understanding and knowledge. Similarly, Bangerter and Clark (2003) propose that the main intention behind an addressee's production of feedback is that he should become grounded with his partner. It is likely that addressees produce feedback to firstly, query the speakers' productions, and secondly, inform the speakers of their state of knowledge. Together these processes will result in the grounding of certain concepts, which the speaker may then refer to in subsequent utterances. This may lessen his need to rely on prior assumptions in creating a model of the addressee's knowledge (although even this model's existence is debatable, according to Pickering and Garrod (2004)), potentially easing his cognitive burden.

The second benefit to feedback in dialogue, according to Krauss and Fussell, is that feedback reduces the pressure on the speaker to create a fully comprehensible

---

[18] The 'minimal feedback' and 'restricted feedback' conditions in Experiments 1 and 5 would be considered to be more like dialogue than monologue, because some interaction was permitted.

[19] In reality, this definition of monologue was not used in Experiment 1, because it was developed after that experiment had been conducted. In later experiments, sound-muffling headphones were used to prevent the speaker from hearing sounds produced intentionally or unintentionally by the addressee.

message at the outset, since he can clarify his addressee's misunderstandings in response to her feedback. When the speaker *does* produce an inadequate description, the addressee can use feedback to signal a lack of comprehension, and the speaker's reply to this should ease her difficulty in making sense of the description. Indeed, it seems that speakers often design their utterances with the expectancy of receiving feedback, or even to intentionally elicit its production. Clark and Wilkes-Gibbs (1986) note two speech patterns which are seemingly employed for this purpose: 'try markers', where the speaker proposes a description with a rising intonation to query if the description is acceptable to his addressee, and also 'installment phrases', in which the speaker provides information incrementally, leaving gaps for the addressee to signal her understanding or lack thereof.

2.12.2 Effect of feedback on addressees' performances

It seems that addressees, for the most part, find it beneficial to be able to give feedback to speakers. Schober and Clark (1989) used a similar tangram-matching task to that of Clark and Wilkes-Gibbs (1986), but they also allowed a second, naïve, participant to overhear the resulting exchange along with the addressee. It was found that addressees tended to be better at arranging their figures than overhearers, presumably because they had the advantage of asking the speaker questions about their task. This study demonstrated that it is not just hearing information which is useful in a dialogue situation (otherwise the overhearer and addressee would have performed similarly, as they heard the same information); the actual *interaction* seems to be crucial.

A similar addressee/overhearer comparison was also carried out by Kraut, Lewis and Swezey (1982), involving descriptions of movie plots, which is likely to have allowed freer speech than descriptions of tangrams. The task of the speaker was

to watch a film and then describe the plot to his addressee, with another person overhearing this description. There were three levels of addressee feedback: full feedback (where the addressee could respond freely) partial feedback (one-word responses only) and no feedback. Kraut et al analysed the effect of addressee feedback on speakers' film descriptions, assessing the descriptions on how many of the plot details were mentioned, and the accuracy and clarity of these mentions. It was found that the quality of the speakers' descriptions increased in proportion with the amount of feedback they received. Kraut et al also found that the ability to give feedback appeared to aid the addressees in their task; they had to re-tell the story to another person and then answer questions on it, and those in the full feedback condition completed this task more successfully than those in the lesser feedback conditions. Again, this result seems to fit with our expectations, because the ability to ask questions will have increased the addressees' knowledge and understanding of the story, helping them with the secondary task. A key point of this study is that similarly to in Schober and Clark (1989), the addressees in this study tended to complete their tasks more successfully than the overhearers[20], presumably since the speaker was coordinating his speech specifically to the addressee's knowledge (or lack of thereof), rather than to the overhearer's knowledge.

However there appear to be methodological problems with the no feedback condition of this design; crucially, the speakers were unaware of the feedback restrictions placed on the addressees, and believed they were always getting full feedback, which could potentially have had a significant effect on their utterances[21].

---

[20] It is not clear whether in Kraut et al's study even those addressees in the no feedback condition performed better than overhearers.

[21] The same concern applies to Bavelas et al (2000), and *may* also apply to Krauss and Weinheimer (1966); it is unclear from the paper.

This might have caused them to over-iterate points (if they thought the addressees' silence reflected non-understanding) or cover them too briefly (if they thought the silence reflected full comprehension), and therefore it may be that the situation was too unnatural to represent normal speech. This experimental design could also have adversely affected the performance of the addressees, because in the no feedback conditions, they were actually giving feedback, but their microphones were disconnected so the speaker couldn't hear them. So if, for example, the addressee had asked a one-word question at some point, they would have received no response to this, possibly leading them to become confused and wonder if the speaker had heard them, thus also affecting their concentration.

The apparently beneficial effect of feedback on the addressees' performance might also have been partly due to the artificiality of the task. In the full feedback condition, the addressees were strongly encouraged to give as much feedback as they could: the instruction was 'You should be as responsive as possible', which might not equate to a normal conversation, and would have required a higher level of concentration in terms of tailoring their feedback to the story (see Bavelas, Coates and Johnson, 2000, described below, for evidence of this) than in the limited or no feedback conditions, possibly increasing their comprehension of the story. In this situation, then, although there was an effect of feedback, it is difficult to draw any definite conclusions regarding the extent to which this was a result of the speakers' lack of knowledge of the situation.

2.12.3 Effect of feedback on overhearers' performances

The trend of the above experiments suggests that it is more beneficial to be an addressee than an overhearer. Fox Tree (1999) investigated this idea further and tested if, since collaboration is beneficial for interlocutors, there is any benefit in *overhearing* the collaboration of others. That is, are monologues or dialogues more

informative for overhearers? This experiment again involved a tangram-matching task, and in two conditions, the addressees were either allowed to converse freely with the speakers, or had to remain silent. The sessions were recorded and later played to overhearers, who attempted to complete the same task. The results suggest that the overhearers' ability to match the tangrams was better in the dialogue condition than in the monologue condition. Two possible explanations for this result suggested by Fox Tree were that, firstly, the additional discourse markers (for example, "I mean" and "You know") she found in the dialogue condition may have benefited the overhearer's comprehension, by allowing him to mentally structure the descriptions better, or secondly, that overhearing two people's perspectives may be more informative than one. However if it were only the additional perspectives that increased performance, then overhearers of dialogues would be just as good at task completion as participants in the dialogue, which doesn't seem to be the case (Schober & Clark, 1989). Perhaps another potential explanation is that the questions asked by the addressee might have mirrored the (unvoiced) queries of the overhearer, and so the speaker's answer to these questions would benefit the overhearer as well as the addressee. It is also possible that the apparent advantage of dialogue for overhearers is a result of the improved contributions apparently produced by speakers as a result of feedback (Bavelas, Coates & Johnson, 2000). It is difficult to determine, then, if the advantage to overhearers is in hearing the listening partners' contributions, or in hearing the speakers' improved descriptions as a result of receiving feedback.

2.12.4 Effect of feedback on speakers' performances

In the same way that it helps addressees to be able to give feedback, and overhearers to hear it, it also helps speakers to receive it.

As a conversation proceeds, the amount of common ground held between the partners will increase, and this may affect their subsequent productions. But how useful to speakers is the knowledge that addressees provide by way of feedback?

There are a substantial number of studies reporting that feedback affects speakers' speech in different ways. According to these papers, feedback tends to increase speakers' efficiency and clarity in giving instructions, makes their story-telling more animated and enables them to communicate humour more effectively (Krauss & Weinheimer, 1966; Bavelas, Coates & Johnson, 2000; Smith, Noda, Andrews & Jucker, 2005). Krauss and Weinheimer (1966) found, using a picture-describing task, that speakers reduced the number of words they used to describe each picture over successive repetitions (similarly to Clark and Wilkes-Gibbs, 1986), and that, more saliently, this occurred most notably when they were given both partner-feedback and positive reinforcement of their partners' success. It is likely that in these cases, as Clark and Schaefer (1987) suggested, the speaker became aware of their common ground through receiving the addressee's feedback, and then tailored his utterances to the addressee's knowledge, this allowing him to shorten them to a greater extent. Additionally, in the full feedback condition, the addressee was able to ask for the information she required and indicate when she had understood, and therefore the speaker wouldn't have had to be over-cautious and re-iterate key points. Speakers may also have employed Clark and Wilkes-Gibbs' installment phrases, allowing the addressees to interrupt and confirm their comprehension even before the speakers had finished their descriptions. In the same vein, Fussell (1990, unpublished, quoted by Krauss, 1996) found that referring expressions for public figures were significantly longer when the speaker couldn't receive feedback from his partner than when he could, suggesting that speakers in monologue might have over-iterated points or given more information than was actually required by their addressees.

The effect of the quality and quantity of feedback on speakers' productions was demonstrated by Bavelas, Coates and Johnson (2000) in a story-telling task. Speakers were invited to tell addressees a real-life story of a close-call incident they had been involved in (for example nearly being involved in a car-crash or nearly being trampled on by a cow). The experimenters manipulated the type of feedback provided by the addressees, by giving some of them a distracting task to complete, for example counting the number of 't's in the speakers' speech. This caused them to make fewer specific responses (for example, 'wow', 'oh no!') and more generic ones ('um', 'uh' etc) in response to the speakers' stories. When addressees provided specific responses, the speakers tended to tell their stories 'better' (as rated by independent addressees, based on the number of negative characteristics) than when the addressees used generic responses. This finding suggests that it is not just the presence of feedback that affects speakers' productions; the actual content of the feedback is crucial too. However what constitutes coherence or quality in discourse production is unfortunately not always consistent between these types of papers; in Bavelas et al's paper, the stories were rated on the number of *negative* features present, for example an abrupt ending, inappropriate story extension and so on, whereas in Kraut et al's study, the number of *positive* features were totalled, and therefore there could have been any number of negative features that went unmarked. Presumably both positive and negative features would need to be taken into account for a truer assessment. It is also worth noting that similarly to Kraut et al (1982), the speakers in this experiment were not aware of their addressees' distractor tasks, and so may have been disconcerted by the lack of specific response from their partners, impairing their performance in this condition.

It seems that that the presence or absence of feedback is not the only crucial factor relating to its use in communication; its *timing* is important too. Krauss, Garlock, Bricker and McMahon (1977) used a task that was similar to tangram-matching, but employed novel (non-tangram) pictures in a referential communication task. Using an audio channel to enable communication between participants, in half the trials

they placed a timing restriction on the audio channel, which meant that after every utterance, one second elapsed before the other partner was able to access the channel to respond. They found that in this delayed feedback condition, speakers used more words to describe the pictures than in the undelayed condition. (Although it would have been interesting to see how the results of the delayed feedback condition compared to a no-feedback condition, this manipulation was unfortunately not included.) Interestingly, when the participants were able to see each other over a monitor, the disadvantage of having delayed feedback disappeared, whereas this additional visibility offered no extra benefits when the feedback was undelayed, visual feedback in this case acting as an adequate substitute for audible feedback[22]. It appears, then, that the timing of audible feedback appears to be crucial, as well as the content. As Oviatt and Cohen (1989) put it, "Even minimal delays can disrupt the organisation and efficiency of spoken discourse" (p130). Feedback certainly seems to provide a benefit for speakers in terms of reducing the number of words needed to express concepts, but perhaps only when it occurs within what might be considered a 'normal' conversational framework.

Schober (1993) found that speakers describing locations of objects tend to be more egocentric (that is, employ less audience design) when they are in a dialogue than when they are in monologue; that is, when they receive feedback as opposed to not receiving any. He suggests this egocentricity could be because, "In a sense, a speaker in conversation can get away with more than a speaker giving a monologue, because she can rely on her addressee to point out any lapses in clarity" (p5), similar to Krauss and Fussell's (1996) earlier suggestion. Nevertheless, he points out that the opposite conclusion can also be logically reached, as he

---

[22] This correlates with findings from Hadelich, Branigan, Pickering and Crocker (2004), who found that visual feedback was an adequate substitute for verbal feedback in a tangram task.

continues, "In another sense, a speaker in conversation must perform to a higher standard, because if her addressee is willing to put in the effort to understand, no lapses in clarity will be tolerated" (p5, Schober 1993). However his findings appear to support the former, rather than latter, point of view. Despite this overall preference for egocentricity in dialogue, coordination still occurred in this spatial perspective-taking task: pairs tended to find a perspective that they both implicitly agreed upon (for example, the director's perspective, the partner's perspective, or a neutral perspective), and continued to use that perspective during a high percentage of the following exchanges.

Fussell and Krauss (1992) downplayed the importance of feedback in dialogue, claiming that "Although interactional feedback is one important source of information about others' perspectives, it is neither necessary nor sufficient for audience design" (p379). Their rationale for this standpoint is that feedback cannot in itself represent the totality of audience design, because speakers need to produce initial references before they receive feedback, so some proportion of their audience design in these early references must be based on their initial perceptions about their addressees. Krauss and Fussell (1996) also propose that an addressee's feedback alone is not sufficient for a speaker to coordinate his speech with her knowledge, but that it must be used in conjunction with the assumptions the speaker has about his addressee, based on her perceived category membership and so on. In a monologue situation, however, no online feedback will be available, and so the speakers will be forced to rely more on their prior assumptions, with the result that utterances in monologue will be less tailored to the needs of addressees than those in monologue (which is demonstrated by the typically poorer performance of addressees in monologue than in dialogue in the above studies).

Feedback from partners is particularly crucial for speakers because often they are not as good at assessing their addressees' knowledge as they think they are;

Fussell and Krauss (1992) found that in judging how much they expect a partner to know, people tend to be biased in the direction of their own knowledge. When describing famous American people and objects for their partners to guess, the amount of detail the speakers gave was directly proportional to how recognisable *they* judged the people to be[23]. Additionally, Keysar and Henly (2002) found a similar biasing effect in the area of prosody; speakers who said ambiguous sentences while attempting to disambiguate them with prosody tended to overestimate the degree to which they were successful in doing so, and expected their addressees to understand the correct meaning more often than they did. Kraljic and Brennan (2005) found that although speakers *did* use disambiguating prosody, they did this regardless of whether the utterance was initially ambiguous or not (ignoring, for example, whether they produced optional disambiguating words such as 'that's' in the phrase 'put the dog food *that's* in the bowl on the floor'), and so they surmised that audience design was not employed in this use of prosody.

This chapter has given an overview of the two main models of dialogue in psycholinguistics, the occurrence of audience design and priming, and finally, the effect of giving, receiving and hearing feedback on speakers, addressees and overhearers. Now that I have set the scene, the following 4 chapters will present a series of studies that are intended to contribute to the body of knowledge described here.

---

[23] However the speakers also seemed to adjust their descriptions in response to the feedback they received from their partners.

Chapter 3: Experiment 1: Lego referential-communication task

3.1 Chapter overview

This chapter will look in detail at the specific experimental techniques that have previously been used to compare the production of monologues and dialogues, and what their findings have told us of the differences between these two modalities. Following this, Experiment 1 will assess how monologues and dialogues that are produced by means of a referential communication task differ on detailed aspects of their production.

3.2 Introduction

Whilst producing a monologue can be very much a solo activity, dialogue may be best considered as a joint construct (Clark, 1996). The interactive nature of dialogue allows participants to provide continuous feedback to each other, feedback which affects both the content of the current discourse and the path it might take as the conversation proceeds. In an ideal situation each speaker will use this feedback to be aware of what his partner does and does not know, and will take account of this in his choice of words and syntax[24]. In some senses the speaker in monologue can be more egotistical in his approach (see Schober, 1993, for a discussion of this), since he will not know if his addressee understands him at all, and so has no obligation to

---

[24] For example, by using the same word in referring to the same object he had mentioned previously, or by mentioning objects earlier in sentences to indicate their new-ness.

adjust his speech in light of this feedback (or lack thereof). Dialogue also differs from monologue in that its structure is more defined; set social rules determine when interlocutors may speak[25] (Clark, 2002; Sacks, Schegloff & Jefferson, 1974), where these are not necessary in monologue.

Along with these very obvious differences, many studies have shown that dialogue and monologue vary in more subtle aspects of language production: they tend to differ in the number of words used to describe objects (Krauss & Weinheimer, 1966), the length of sentences (Oviatt, 1995), and the perspectives taken by the speaker (Schober, 1993). This chapter aims to add to these findings in assessing the influence of an addressee's feedback on their partner's speech, with particular emphasis on the number and consistency of words used by the speaker. It will also look at how feedback influences a pair's joint success at the given task.

3.2.1 Making a reasonable assessment of feedback effects

Why should we study the differences between monologues and dialogues? The main reason is that the bulk of previous psycholinguistic research has focussed on the production and comprehension of monologue, and yet surely the primary goal of psycholinguistics is to detail how we produce and comprehend language in our everyday lives, which in the main involves dialogical interaction. It is clear that monologues and dialogues differ even in the contexts in which they are usually

---

[25] Although these rules are often violated in practice, for example with interruptions.

produced, and in the level of advance planning involved[26]. Perhaps the most common types of monologues are used in activities like lecturing and public speaking, and these planned monologues are typically fairly efficient ways of imparting information. In comparison, those monologues that are *not* planned, for example messages left on the phone, often sound unstructured and incoherent, despite the speaker's best attempts at fluency (this may be partly because he is used to receiving feedback from a partner, and here doesn't receive it, which puts him in an unusual situation). Experimental evidence demonstrates that the amount of prior planning involved in monologues seems to have a notable effect upon language production: Oviatt (1995) found that fewer disfluencies are produced in 'structured' monologues (where the speakers were told what to mention at what point) than in spontaneous ones. This may be a direct result of the amount of advance planning carried out.

In comparison with monologues, the majority of everyday dialogues (for example, greeting someone in the street or answering the telephone) are unplanned, and as a result of this, we are probably all more practiced at producing unplanned dialogues than unplanned monologues. This may explain why spontaneous dialogues have been found to demonstrate more communicative efficiency than spontaneous monologues (Kraut, Lewis & Swezey, 1982). Kraut et al demonstrated that people listening to a story in which they were allowed to give feedback (i.e. in a dialogue situation) remembered and retold the story better than those who weren't allowed to give feedback (in a monologue situation), suggesting that some aspect of the interactive nature of dialogue makes it helpful to addressees in a way that

---

[26] By advance planning, I mean planning which takes place intentionally before the whole monologue begins; such planning may comprise, for example, making a written record of what is to be said, or having a spoken rehearsal.

monologue is not[27]. Of course the advantage of dialogue demonstrated in this study may have been due to the benefit for the *addressees* of being able to give feedback, rather than the effect of feedback on the speakers (or plausibly both). Regardless of the underlying explanation, though, speakers in dialogue may communicate information to their addressees (and even to overhearers, according to Fox Tree, 1999) more effectively than speakers in monologue, at least in as far as laboratory-based production tasks are concerned.

3.2.2 Techniques for studying dialogue and monologue and their implications

In some ways the study of dialogue is more complex than that of monologue, not least because of the difficulty of orchestrating the production of conversations that are natural, yet controlled to some extent; a juxtaposition which has challenged dialogue researchers for some years. Whilst the most natural study of dialogue would involve simply recording people's everyday conversations, this tends to produce wildly varying discourses, which may be very uncontrolled in their content, style and intentions. Even experimentally generated 'spontaneous' conversations will display the same variability[28], producing somewhat unclear data that may be more fit for qualitative examination than quantitative. One method of evading this problem, while still allowing for some degree of spontaneity, is to employ a referential communication task (based on Krauss and Weinheimer, 1966), which involves one person (the speaker, describer, instructor etc) giving information to another person (the listener, matcher, follower etc), in order to help her complete a given task. The main advantage of this type of paradigm is that it

---

[27] But see Section 3.2.2 for a critique of the method used here.

[28] See Kent, Davis and Shapiro (1978) for an example, where pairs of participants were simply instructed to 'get to know each other' in a dialogue task.

constrains the topic of conversation to that determined by the experimenter, whilst not restricting the speaker's actual choice of words or syntactic structures. The most common variant of the referential communication task involves a speaker telling an addressee how to re-order a number of items (for example cards with pictures on them) into a display similar to the one seen by the speaker. There have also been tasks involving the construction of items from their constituent parts (Clark and Krych, 2004, using Lego models) or other similarly practical tasks (e.g. Kraut, Fussell and Siegel, 2003, repairing bicycles).

Clark and Krych's (2004) study was typical of the genre. One partner (the Director) told the other partner (the Builder) how to construct Lego models, under a variety of feedback conditions. In a third of the pairs, the partners could speak freely and the Director could see the Builder and his workspace, containing the model. In a second third they could speak freely but the Director could not see the workspace (and in half of these he could not see the Builder's face either), and in the final group, the Director and Builder had no direct interaction, as the Director's instructions were simply recorded on audio tape for the Builder to listen to at a later point. Clark and Krych found that the more interaction was allowed between the partners, the better the Builders were at creating the models. They also noted that in the full interaction condition, the Builders communicated constantly with the Directors in a variety of ways, and this communication seemed to lead the Directors to adjust their utterances on a minute-by-minute basis according to the Builder's needs, demonstrating a high level of audience design.

If the use of referential communication tasks such as that above seems to be relatively successful for the production of dialogues, we must consider how to generate monologues within a similar context, to allow a direct comparison of these two modalities. Although again it is difficult to orchestrate the production of monologues in the referential communication task in any *entirely* natural sense,

given that the speaker must be aware of his partner and yet cannot receive any feedback from her, numerous attempts have been made to do this in previous studies. The techniques employed have varied widely, such as making the speaker speak into a tape player in lieu of talking to his partner (as above; also Schober, 1993), letting the addressee speak but disconnecting her microphone (Kraut, Lewis & Swezey, 1982), and distracting the addressee so that she barely speaks at all (Bavelas et al, 2000).

Schober (1993), like Clark and Krych, also had speakers describe pictures into a tape recorder in lieu of a real partner, in order to produce a 'monologue' condition in his study on spatial-perspective taking. He found that speakers in this 'monologue' condition tended to use fewer egocentric perspectives (that is, descriptions from their own point of view, rather than their partner's) than speakers in dialogue, and attributed this to the difference between having an 'imaginary' partner (that is, a partner who would listen to the tape recording) and a 'real' partner. However it is unclear how much of this difference was actually due to the lack of interaction (that is, the lack of feedback from a partner), and how much was due to the lack of an actual partner, regardless of whether that person spoke or not. That is, there was a confounding in this study between the amount of interaction allowed and the presence of a real addressee, so it is difficult to draw any definite conclusions on this methodology.

There is evidence, however, that using a tape recorder instead of an addressee in monologue may have some disadvantages. Smith, Noda, Andrews and Jucker (2005) had speakers describe film plots into tape recorders, pretending that the tape recorder was the answer phone of another person who would listen to it later. Upon analysing the monologues produced, they found that the speakers had varied wildly in their impressions of who their intended addressee was, and had adjusted their reference strategies accordingly. For example, a significant proportion of the

monologues appeared to have been produced with the experimenters in mind, perhaps because the speakers believed that they were participating in a memory task (or some other similar test). The monologue speakers tended to introduce New characters in a Given style (for example referring to '*the* waiter' when that character had not been previously mentioned, rather than '*a* waiter'), which is an unusual strategy when you are speaking to someone who hasn't seen the film (rather than the experimenter, who participants knew had). Smith et al concluded that it is necessary to use a real person as the addressee in the production of monologues, rather than a tape player, to ensure that the speakers are designing their utterances for naïve addressees. It is also possible that even those speakers who *do* believe their speech will be played to an actual addressee may not behave entirely 'normally'; many speakers will be self-conscious about speaking into a tape recorder, and may be particularly careful or specific about what they say for two reasons: partly because of the unnatural situation, and partly because they may be embarrassed that their recordings will be kept and replayed to others.

Kraut, Lewis and Swezey (1982) generated monologues using a different method from that above. Their task involved having a speaker describe the plot of the film they had just watched to an addressee who was in one of three possible feedback conditions: full feedback, limited feedback and no feedback. The full feedback condition (equating with dialogue) involved free speech between the partners, and the limited feedback allowed the addressees to give one-word replies. The no feedback condition (equating to monologue) involved seating the partners in separate soundproof booths, and disconnecting the addressees' microphones from 30 seconds into the task, so that the speaker received no feedback. Whilst it is possible that this strategy might have fooled the speaker on some occasions, it is more likely that, similarly to my interpretation of Bavelas et al (2000; see below), the lack of any feedback in the no feedback condition confused him, particularly when he asked questions of his addressee and didn't appear to receive a reply. Moreover, the addressee may have been even more confused, since she will have received no

reaction at all to her queries and comments, and this may have dramatically affected her success at the task (in this case, her understanding of a film plot).

Bavelas et al (2000) distracted addressees with a second task while they were supposed to be attending to speakers. This restricted the addressees' feedback and caused them to produce near-monologues. The paradigm used was as follows: the experimenters asked speakers to describe 'close call' scenarios that they had experienced to an addressee. The addressee was in one of two conditions: either she listened carefully to the story, or else she was performing a distraction task (like counting the number of times the speaker used the letter 'T'), unbeknown to the speaker. This study therefore restricted, but did not eliminate, the amount of feedback produced by the addressee, and so the second condition would be more accurately assessed as a 'limited feedback' condition than a monologue. The experimenters analysed the amount of feedback that was provided by the addressees, and the effect that this had on the speakers. They found that the speakers told their stories significantly less 'well' in the distraction condition, becoming more repetitive and less well-structured, and that the addressees also understood the descriptions less well. However, it is possible here that the task design – in particular, the speakers' lack of awareness of the distractor task - may have adversely affected the results, because the speakers were expecting a dialogue and in fact were only partaking in a monologue. Although this task would have been more realistic for speakers than Kraut et al's (in that they were still receiving some feedback), it is still likely that they will have been disconcerted when they were expecting to receive specific feedback relating to the story and did not receive much from their partners. This confusion may have affected their production, and caused the general disjointedness and poor standard of storytelling that Bavelas et al attribute solely to the lack of feedback. It is possible that the speakers' performances might have been very different had they known that they wouldn't receive feedback.

Our report of the two studies reported above makes the point that, when experimenters are orchestrating the production of a monologue, the speaker should be made aware of the listener's inability to give feedback. After all, the majority of monologue speakers (radio show presenters, for example) do not expect to receive a direct response from their listeners; in some sense, this lack of expectation is a characteristic of monologue speech.

Unnatural experimental set-ups such as those above are not specific to the elicitation of monologues. Some studies investigating *dialogue* have also neglected to use 'real' participants in tasks, most notably Brown and Dell (1987; also Keysar et al, 2000). As mentioned in the previous chapter, Brown and Dell found that people describing stories to addressees tended to mention the instrument in a separate clause after the verb, no matter how much they thought their addressee knew about the instrument, apparently showing a lack of audience design. However, crucially, the addressees used in this experiment were confederates, who by the end of the experiment had heard the story up to 40 times, and so presumably would not have given the same response as a first-time hearer. Lockridge and Brennan (2002) criticised the paper on this basis, and replicated Brown and Dell's study using naïve addressees rather than confederates. Their results differed significantly from Brown and Dell's, and demonstrated much more audience design than had first been assumed. They surmised that this difference was due to the response of the confederate 'addressee' in the original experiments; where a real addressee would give appropriate and natural responses in the right places, they suggested that confederates, having heard the story being told a number of times, would produce, at best, unconvincing feedback, and at worst, no feedback at all. The findings they obtained using real addressees certainly seemed to be consistent with this conclusion.

3.2.3 Rationale for Experiment 1

Given the potential difference in communicative effectiveness between monologues and dialogues, it is important that in carrying out a direct comparison between them we orchestrate the production of monologues and dialogues which are equivalent to each other, at least with regard to the task goal. Otherwise we are in danger of comparing two situations in which even the communicative intentions differ wildly, which could potentially have a substantial effect on the more form-based aspects that interest us. It appears that the most experimentally useful situation for my purposes might involve a natural comparison between *instructive* monologues and dialogues (where one person gives another person task instructions to follow), since instructive situations occur both in monologues and dialogues. Although uninstructive dialogues are commonplace (the main example being informal conversations), there are few situations in which uninstructive monologues are produced in everyday life, except by speakers who are experienced at producing monologues, such as radio show presenters, lecturers or sports commentators. For this reason I decided to use a referential communication task in this study.

The present experiment involved building Lego models. One participant (the 'Describer') was given an abstract Lego model, and had to describe it to his partner (the 'Builder') in such a way that the Builder could build an identical model from his own, loose Lego pieces. (Note: Although both males and females played both roles in all experiments in this thesis, for ease of reference the Describers (or speakers) will be referred to as male, and the Builders (or addressees), female).

Mindful of the problems mentioned above, I employed naïve participants, not confederates, as addressees in all three conditions. I decided to restrict the amount of feedback that the Builder was allowed to give, and, crucially, to make the speaker fully aware of this restriction. The amount of feedback that the Builder was

allowed to give was constrained to produce three conditions: 'no feedback' (Builder remained silent), which was considered to equate with monologue, 'full feedback' (Builder interacted normally), which was comparable to dialogue, and a third condition of 'minimal feedback' (Builder was only allowed to give very brief feedback, with no task-specific questions or comments). This final condition was designed to investigate just what it is about feedback that makes it beneficial (if it turns out to be so); whether it be the actual words the listeners use, and the amount of detail they go into, or the mere fact of having feedback at all. One way of assessing this was to see which condition, out of monologue and dialogue, the minimal feedback behaved most like[29]. The intention was to analyse the resulting recordings, to present an overall picture of the defining characteristics of monologues, dialogues, and dialogues with minimal feedback, in particular focusing on the differences in the number of words the Describers used and how repetitive their speech was throughout the three conditions. It was decided to use only pairs of friends as participants simply so that the relationship between the participant pairs in all conditions was similar.

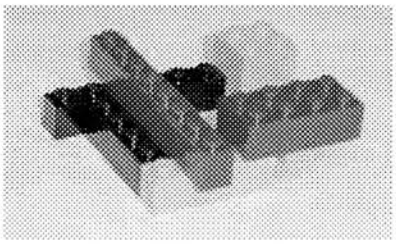
---

[29] Kraut et al (1982) created a similar 'limited feedback' condition, in which the listener could say only one-word utterances, but, as discussed earlier, I have concerns with the methodology of their monologue condition, and so a comparison of the monologue and limited feedback conditions in their study may not show us an accurate picture.

3.3 Experiment 1

Materials

Nine abstract Lego models, created by the experimenter, were used as the construction prototypes. Each model was formed from 7 pieces of Lego. The colours varied across models, but each model had only 3 colours. Each model used a random selection of shapes (some colours and shapes were repeated within models but each model contained at least 5 different shapes) from the basic Lego kit. An example of one of the models is shown below (Note: although the picture is in black and white, the models were in colour).



One of the models was given to the Describer at the beginning of each trial. The Builder was given only the exact Lego pieces necessary to build each model. There were nine trials, and so each Describer saw, and every Builder tried to create, every one of the nine models.

Procedure

The pairs of participants were told that the experiment was investigating 'how good people are at building Lego models when they can't ask questions'. They were encouraged to complete the task to the best of their abilities, and a financial bonus was promised to the pair who attained the highest score. Each participant was randomly designated the role of either 'Describer' or 'Builder', and was given written instructions (see Appendix A) pertaining to this role. They were allowed to ask the experimenter any questions relating to their task before the experiment began. The participants were seated at tables in the same room, but were separated by an opaque screen, so although they could hear each other clearly, they could not see each other, nor see each other's models.

Each participant pair took part in all three feedback conditions. Along with the general instructions, which were the same in each condition, at the beginning of each new condition the Builder was given instructions referring to the amount of feedback she was allowed to give in the following trials. The Describer was given corresponding instructions informing him of the precise limitations on the Builder. The instructions given to the Builder in the three conditions were as follows:

> *Full feedback:* You may talk to your partner normally, and ask as many questions as you like in an effort to complete your task.
>
> *Minimal feedback:* You are restricted in the types of things you can say to your partner. You can only use simple one-word terms like these:
>
> 'Wait', 'Repeat', 'Expand'.
>
> You may also say 'okay', 'yes' and 'no'.
>
> You may not ask any specific questions or make any comments pertaining to the Lego models, such as 'Where?', or 'It doesn't fit!'. You are not restricted to the exact phrases above, but you must not say anything more specific than these. Your experimenter will tell you if you say more than is permitted.

When you believe you have completed a model, you may say 'Finished' to your partner.

No feedback: Your partner will talk to you, but you must not communicate with him in any way. This includes making noises (such as 'uh' or 'em'). You must listen carefully to your partner and try to complete your task. When you believe you have completed a model, you may say 'Finished' to your partner.

Both participants were made aware of the feedback restrictions on the Builder, in order that the instructor would not misunderstand his or her silence for assent or confusion. On the few occasions when the Builder attempted to break the rules, the Experimenter interrupted after only one word and issued a warning, with the understanding that if this happened again, the experiment would be terminated.

The experiment lasted for 9 trials, each one involving a different Lego model. The models were presented consecutively in a random order determined by the experimenter, which differed for each pair of participants. The feedback conditions were run within-participants, with each pair undergoing 3 consecutive trials in each feedback condition (for example, Pair 1 might have seen models 1-3 in full feedback, 4-6 in minimal feedback and 7-9 in no feedback). The overall order of models and the order of models within each condition varied systematically, such that every model appeared in every condition 6 times, twice in each position in the condition. The order of conditions was balanced over participant pairs, with 3 pairs in each of 6 order groups, which represented all possible orders of the three feedback conditions.

Each participant was given a lapel microphone, and these were linked together by a coupler which fed into a DAT recorder, to provide a stereo sound file where the two participants' voices were on separate channels. This was in order to facilitate the transcription of overlapping speech, should it occur. The time limit for each trial

was 5 minutes, after which time an alarm would sound. This was imposed in order to stop the experiments from running on extensively. If the participants had finished before the time was up, they were allowed to stop and move on to the next trial if they wished (in all conditions the Builder was allowed to say 'finished' when they had completed the model). At the end of each trial, the experimenter collected the Builder's model and compared it to the Describer's model. A score was calculated for each model, and the participants were only told their total score at the conclusion of the experiment. The raw audio data from the experiment was transferred from DAT tapes to WAV files on computer, and was then analysed manually using the phonetic software package PRAAT.

The disfluencies and filler items (e.g. 'er', 'um') produced by the Describer were transcribed and coded separately, because the following analyses were primarily concerned with the number of informative words produced, in terms of both the amount of information the Describer provided and the amount the Builder received. Once these had been removed, the remaining speech was transcribed separately, forming a written account of the fluent content produced by each Describer. The speech of the Builder was not transcribed, because the full-feedback condition was the only one in which they could speak freely, and so no comparison could be made across conditions.

Participants

18 pairs of friends participated in the experiment, in exchange for payment. All were undergraduate and postgraduate students at the University of Edinburgh. They each indicated the closeness of their relationship with their partner on a six-point scale after the experiment, and all participants, bar one, judged their relationship with their partner as being point 4, 5 or 6 on this scale, indicating that

they knew each other fairly well (the one exception chose point 3). The mean rating was 5.7. The experiment took approximately one hour.

3.4 Results

The instructions produced by the speaker given varied widely between speakers and between conditions. There were two main sources of variance. The descriptions of the individual Lego blocks differed substantially in their level of detail, from *"the long yellow"* to *"the six-node yellow, the really long one"*. In addition to this, the actual instructions for creating the model also varied in length, from succinct descriptions like *"fix the third and fourth on the yellow to the top two of the red"*, to the lengthier *"take the yellow in your right hand, and the red in your left. Now, you want to attach the third and fourth nodules on the left hand side of the yellow to the top two nodules on the red, so that they lie perpendicularly"*. Example transcripts for one of the models in each of three conditions (by three different participant pairs) are reproduced in Appendix B.

The analyses to follow were all carried out with respect to two main variables: Condition and Trial Order. Condition refers to the amount of feedback that was allowed from the Builder (with 3 levels: Full Feedback (FF), Minimal Feedback (MF) and No Feedback (NF)), and Trial Order refers to the temporal position of the trial in each overall experiment, from Trial 1 to Trial 9. It was decided to carry out analyses by-participants only because the Lego models used were not distinct items as much as, say, particular picture-cards would be; the decisions to use seven Lego blocks in each model, and three models in each feedback condition, were somewhat arbitrary, and so the results from these models may not be generalisable to a difference set of Lego models.

3.4.1 Score

The first obvious aspect of the data to look at is how successfully the participants completed their task. Each pair was given a score for each trial, which defined how similar the Builder's created model was to the Describer's prototype model. The pair was awarded one point for each connection correctly made between pieces. Although there were 7 pieces, there were only 6 connections made, and so the maximum score available was 6. There were nine trials in the experiment, and Table 1 (below) demonstrates how the mean scores in each condition (NF= No feedback, MF= Minimal feedback, FF= Full feedback) varied as the experiment proceeded. The mean scores are shown below in Table 1.

Table 1: Mean scores attained per trial, per condition

	Mean Score		
Order	NF	MF	FF
1.00	1.7	3.0	3.0
2.00	3.2	2.7	4.5
3.00	3.7	4.2	4.7
4.00	4.2	4.8	5.0
5.00	4.5	5.0	5.7
6.00	4.7	5.2	5.3
7.00	4.3	5.0	4.8
8.00	3.8	4.7	5.3
9.00	5.2	4.3	4.8
Mean	3.9	4.3	4.8

The distribution of scores over all trials for the three conditions is shown below in Table 2.

Table 2: Frequencies and percentages of scores attained in each feedback condition.

	Frequency			Percentage		
Score	NF	MF	FF	NF	MF	FF
0	5	4	0	9.3	7.4	0
1	7	7	3	13.0	13.0	5.6
2	4	3	2	7.4	5.6	3.7
3	4	3	5	7.4	5.6	9.3
4	4	3	9	7.4	5.6	16.7
5	12	5	9	22.2	9.3	16.7
6	18	29	26	33.3	53.7	48.1

It is clear from the table that the task had a high overall success rate, with more than a third of models being completed perfectly. The frequency of perfectly-completed models did not differ significantly over the three conditions ($X^2(2) = 4.84$, NS). 3 (Feedback condition, within-participants) X 9 (Trial order, within-participants) repeated measures ANOVAs demonstrated that Trial Order had a significant effect on scores by participants, with scores tending to increase significantly in later trials ($F(8,40) = 5.78$, $p<.001$). There was no effect of Condition ($F(2,10) = 2.69$, NS), nor was there an interaction between the effects of Condition and Trial Order ($F(16,80) = .313$, NS). However, 2X9 ANOVAs on just the FF and NF conditions produced a slightly different pattern of results; there was a main effect of Trial Order ($F(8,40) = 3.03$, $p<.01$), with later trials producing higher scores, and also of Condition ($F(1,5) = 6.53$, $p=.05$), with the FF condition showing higher scores than the NF condition. There was still no interaction, suggesting that feedback had an overall beneficial effect and influenced task performance fairly evenly throughout the experiment. A planned comparison on a one-way ANOVA showed that polynomial linear trends were

significant between score and trial order in the NF condition ($F(1,7) = 7.3$, $p=.01$) and the FF condition ($F(1,7) = 5.48$, $p<.05$).

The finding that Builders completed their task more successfully in the FF condition than in the NF condition (as illustrated by the main effect of Condition) is not surprising, given that they had the opportunity to ask questions of their Describers, which should have eliminated any misunderstandings. What was less predictable is that there was no interaction between Trial Order and Condition in these results. It could have been supposed that in the FF condition, performance would improve more markedly than in the NF condition, because the Describer would find out quickly what kind of descriptions were most useful to the Builder, and would tend to produce more of these descriptions. But this seems not to have been the case. It is possible that a ceiling effect in the FF condition was responsible for this lack of improvement; the FF scores were higher to begin with, so there may have been less room for improvement than there was in the NF condition. However, the fact that only 48% of the FF models were completed correctly (thereby suggesting that there was still room for improvement in the majority of FF trials) makes this an unlikely explanation.

Why did the scores increase over time in all conditions? This could be a result either of the Describer describing the models more effectively with practice, or the Builder interpreting the Describer's descriptions more competently with practice, or both. It is possible, of course, that the explanation will be different for the different conditions; perhaps the Describer's descriptions became more effective in the FF condition as a result of his partner's feedback, while in the NF condition, the Builder became more competent at interpreting her partner's descriptions with practice.

3.4.2 Time

The amount of time taken per trial and the number of words used by the Describer were highly correlated ($r(162) = .583$, $p<.001$), and so these measures will be considered in adjacent sections, to allow direct comparisons to be made.

The time taken by each pair of participants to complete each trial (that is, to build one model) was measured. Since there was a time restriction of 5 minutes (300 seconds) on each trial, this was the maximum time that was ever recorded. More often, though, the subject pairs would finish their model construction before the end of the given time, and on these occasions a note was made of the time taken for the trial. This occurred 135 times out of 162 (79.67% of the time overall, occurring in 16.7% of NF trials, 18.5% of MF and 14.8% of FF; these frequencies did not differ significantly from each other ($X^2(2) = .27$, NS)). The distribution of timings over the course of the experiment, broken down by condition, is shown below in Table 3.

Table 3: Mean times taken per trial, per condition.

	Mean time (seconds)		
Order	NF	MF	FF
1.00	232	208	246
2.00	231	238	243
3.00	228	247	250
4.00	247	240	260
5.00	237	223	192
6.00	245	240	240
7.00	229	189	183
8.00	223	200	168
9.00	218	232	194
Mean	232	224	220

3X9 ANOVAs demonstrated that there was a significant main effect of Trial Order on mean timings ($F(8,40) = 3.72$, $p<.01$), where the later trials were completed more quickly than the earlier ones. There was no effect of Condition ($F(2,10) = .95$, NS), and no interaction ($F(16,80) = .56$, NS). 2X9 ANOVAs on only the NF and FF conditions also showed an effect of Order ($F(1,5) = 2.48$, $p<.05$) but still no effect of Condition ($F(1,5) = 1.50$, NS) and no interaction ($F(8,40) = .93$, NS). Polynomial linear trends were significant between time and trial order only in the FF condition ($F(1,7) = 11.51$, $p<.01$). Thus it seems that later trials were completed more quickly than early ones, but that the amount of feedback permitted did not influence this significantly.

3.4.3 Total word count

Next the total number of full words (not including disfluencies) spoken by the Describer in each trial was counted. The mean words per trial (that is, per model), per condition are shown below in Table 4.

Table 4: Mean number of words spoken by the describer per trial, per condition

	Mean Number of Words		
Order	NF	MF	FF
1.00	369	407	397
2.00	386	517	445
3.00	393	504	449
4.00	430	431	450
5.00	389	434	403
6.00	400	465	443
7.00	419	463	427
8.00	338	432	362
9.00	344	439	367
Mean	385	455	416

3X9 ANOVAs revealed a main effect of Trial Order ($F(8,40) = 2.98$, $p=.01$), with fewer words being used in later trials. There was also an effect of Condition ($F(2,10) =9.51$, $p<.01$), but no interaction. 2X9 ANOVAs between just the NF and FF conditions also demonstrated a main effect of Trial Order ($F(8,40) = 5.35$, $p<.001$), but not of Condition ($F(1,5) = 1.29$, NS), suggesting that the effect of Condition in the previous analysis was caused by the results from the MF condition. There was still no interaction ($F(8,40) = .834$, NS). Polynomial linear trends were significant between total number of words and trial order only in the FF condition ($F(1,7) = 15.37$, $p<.01$). As above, it seems that the number of words used by the Describer reduced over trials, but was not significantly influenced by feedback.

3.4.4 Length of first references to Lego blocks

A large portion of each transcript comprised references to Lego blocks; from these, the number of words used to refer to each Lego block *for the first time* was counted. References varied wildly in their length and level of detail, changing from lengthy descriptions like 'a small brown one with just four nodules on, the square one', to simply 'the brown square'. For these analyses, the relevant NPs were regarded as beginning with, and including, their determiner, and ending just before a shift in focus (typically towards a movement; something like, 'and attach it to...'). For example, an NP might have been the emboldened parts of the extracts below:

"...then take **a long four red one** and place that over..."

"...if you take **the two by four black one** and place that..."

"...get **a blue thin one that's one across in width and three along** and attach that..."

As mentioned previously, disfluencies were not included in these word counts. The mean lengths of NPs are shown below in Table 5.

Table 5: Mean length of first block descriptions, per condition, per trial

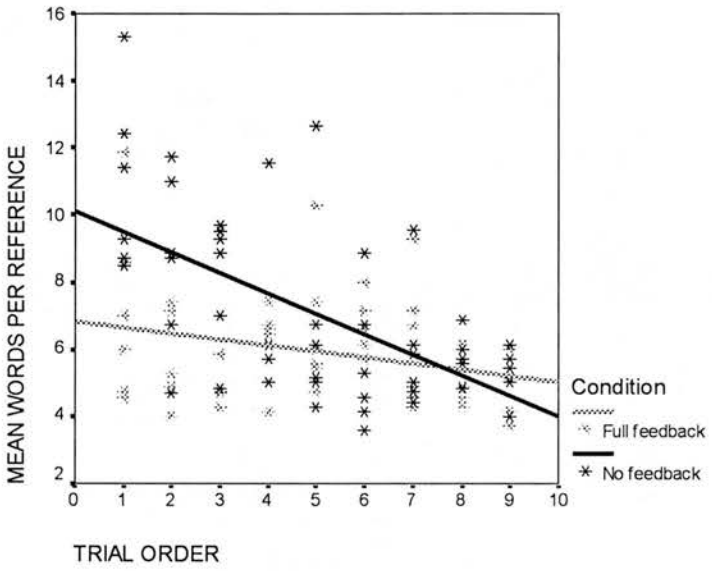
	Mean words per description		
Order	NF	MF	FF
1.00	10.94	5.82	7.12
2.00	8.61	5.70	6.26
3.00	8.20	5.36	4.94
4.00	6.52	6.74	6.05
5.00	6.66	6.44	6.57
6.00	5.52	5.74	6.62
7.00	5.95	5.62	6.14
8.00	5.83	6.21	5.05
9.00	5.24	5.14	4.55
Mean	7.05	5.86	5.92

It was expected that the length of NPs would decrease over trials (cf. Clark & Wilkes-Gibbs, 1986; Krauss & Weinheimer, 1966[30]), and this is exactly what happened: 3X9 ANOVAs demonstrated a main effect of Trial Order ($F(8,40) = 9.78$, $p<.001$), where the number of words per NP typically decreased in later trials. There was also an effect of Condition ($F(2,10) = 4.32$, $p<.05$), and an interaction ($F(16,80) = 1.93$, $p<.03$). 2X9 ANOVAs on only the NF and FF conditions revealed a similar effect of Trial Order ($F(8,40) = 6.70$, $p<.001$). There was also a marginal effect of Condition ($F(1,5) = 4.8$, $p=.08$), demonstrating that the NF condition used more words than the FF condition. There was still an interaction ($F(8,40) = 2.38$, $p<.05$), showing that the number of words reduced more over the course of the experiment in NF than FF. A polynomial linear trend was significant between NP length and trial order in both the NF ($F(1,7) = 28.9$, $p<.001$) and FF ($F(1,7) = 4.72$, $p<.05$) conditions. Figure 3 (below) demonstrates the interaction, with the mean number of

[30] Note, though, that these papers both report reductions of repeated references to the same objects, rather than shortening references to *new* objects.

words per NP along the vertical axis, and the trial order (indicating temporal progression through the experiment) along the horizontal axis.

Figure 3: Mean number of words per noun phrase per trial.



The graph demonstrates that references in the NF condition tended to start off significantly longer on Trial 1 (mean length 10.9 words) than those in the FF condition (mean words 7.1; significantly different from NF: ($t(10) = 2.44$, $p<.05$)). They then proceeded to reduce much more in the NF condition than in the FF condition, and by the last reference, on trial 9, there was no significant difference in length between the descriptions in the NF and the FF conditions ($t(10) = 1.42$, NS).

Comparison with Krauss and Weinheimer (1966)

These results initially seem to contrast with Krauss and Weinheimer's (1966) finding that NPs reduce in length more radically in dialogue than in monologue.

However my experiment is not directly comparable to theirs; they investigated repeated references to the same object, whereas the present study looks at first references to new (previously undescribed) objects. Perhaps, then, the shortening they reported was a result of the accumulation of common ground in dialogue but not in monologue. This is less likely to occur in the present experiment, since only new references were studied[31], and thus the current findings may represent the natural shortening of references *in the absence of common ground.* Since here *neither* condition was able to use common ground to facilitate shortening, it may be that all descriptions naturally tend to become shorter with repetition regardless of this, and that the monologue descriptions in this situation only shortened more because they began longer. There may however also be alternative reasons why the current findings depart from theirs, and these are worth exploring. It is possible that the contrasting results of these two studies can be attributed to task differences; it is unclear from Krauss and Weinheimer's paper whether, in the 'no verbal feedback' condition, the speaker was aware that the listener couldn't give feedback. If he was not aware, it is possible that the lack of shortening in this condition was actually a reflection of the speaker's confusion. Since he would not have received the feedback he expected, he might have been reluctant to take the risk of shortening his noun phrases, for fear that his partner would not understand him.

A second important point relates to the actual shapes used as experimental stimuli in these experiments. Where the current task used Lego pieces, Krauss and Weinheimer used abstract geometrical shapes. At first glance there seems to be little difference between the stimuli sets, given that both of them are fairly abstract. However, it could be argued that Krauss and Weinheimer's shapes were likely to be

---

[31] It is however possible that particular features of Lego blocks from previous descriptions (e.g. 'long', 'blue') may become part of the pair's common ground, despite no repeated references to *exactly* the same object.

viewed more figuratively than the Lego pieces were. Krauss and Weinheimer's shapes could potentially have been described in many different ways. It is likely that the speakers in the dialogue condition of their study, after producing a certain number of detailed descriptions of the shapes, began to describe them less in detailed geometrical terms and more figuratively, like, for example, "the alien with the spiky bits". (While Krauss and Weinheimer don't give any examples of the descriptions given in their study, this style would certainly be in keeping with those found in Clark and Wilkes-Gibbs (1986), which used similarly ambiguous Tangrams). Producing a shortened figurative description would constitute taking a risk for a speaker, because he couldn't be sure that his partner would understand it. He may have felt even less confident about shortening it when he couldn't receive feedback from his partner. This could lead to less reduction occurring in the no feedback condition than in the full feedback condition.

In contrast with the scenario above, any reductions of Lego descriptions, as in the current experiment, are likely to remain quite literal and transparent, because there is a very limited number of ways to describe Lego pieces. Even a short description like "the long blue" does not require much interpretation by the listener; it has an obvious meaning that will not be easily mistaken, even by a partner who is not allowed to ask questions. This could mean that the descriptions in the NF condition are likely to end up just as short as those in the FF condition (as demonstrated statistically above), because the Describer is less wary of being misunderstood. The fact that the descriptions reduced more over the course of the experiment in the NF condition is most probably a result of their being longer to begin with, which in turn may be due to the Describer's understanding that his partner cannot give feedback, and so he should give particularly explicit instructions at the beginning of the experiment.

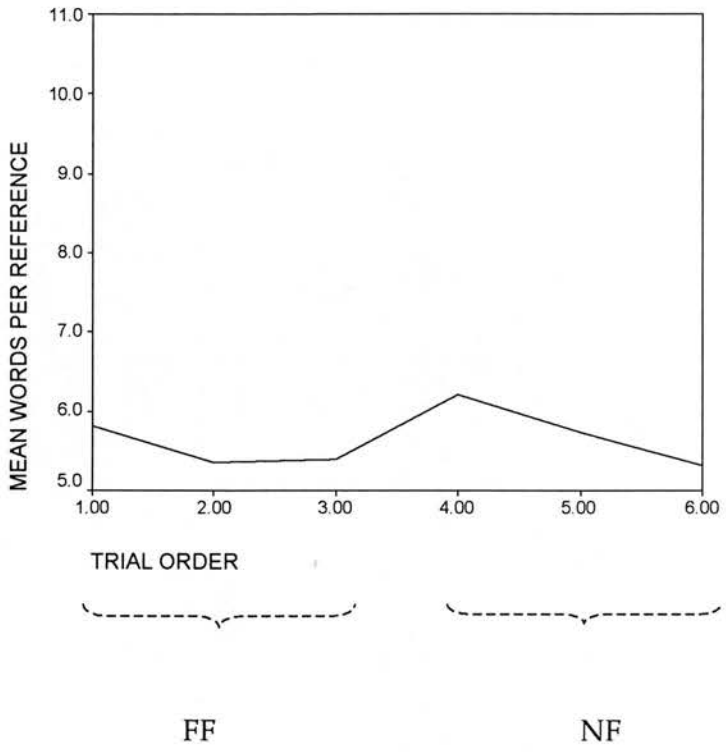
3.4.5 How length is affected by condition change

It seems that although people reduced their NPs more without feedback than with it, the NPs still ended up being roughly the same length in both the NF and FF conditions. But what happened when the conditions were changed during the experiment? Since the feedback conditions were run within-participants, these data can be split up further, to investigate what happened in the two most extreme conditions: when a pair in the FF condition changed into the NF condition, and vice-versa.

Full Feedback – No Feedback

The graph below illustrates the length of NPs for those participants who completed three FF trials followed immediately by three NF trials, with no intervening MF trials. The trials were re-numbered for the sake of the graph, so that for half of the participants used in these analyses, the data below represents the actual trials 1-6, and for the other half, the data represents trials 4-9. Figure 4 (below) shows that in this situation, the NPs lengthened on the 4th trial, which represented the first trial in the NF condition, however paired t-tests comparing the 3rd and 4th trials show that this increase was not significant ($t(5) = -1.17$, NS).

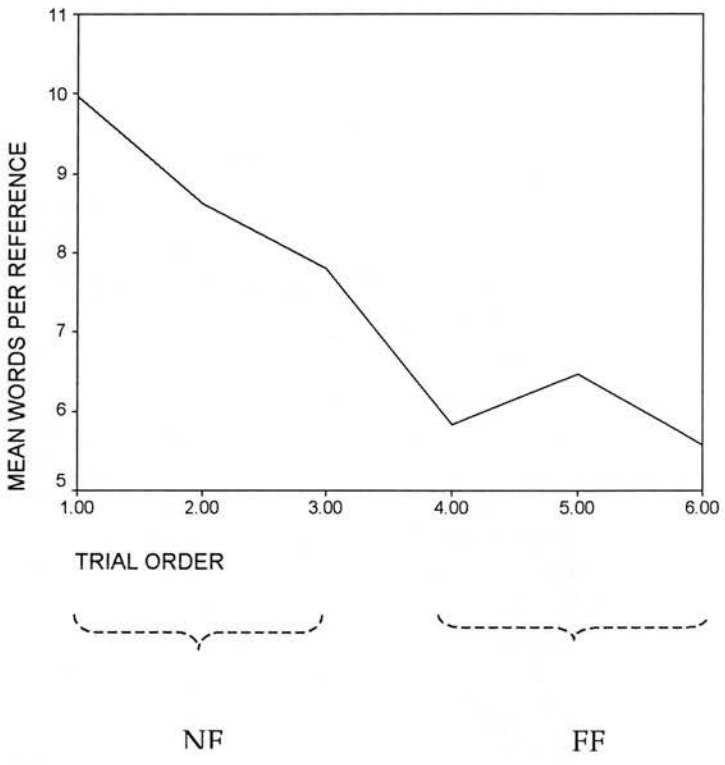
Figure 4: Pattern of Noun Phrase length with decrease in feedback



No Feedback-Full Feedback

In the opposite situation, when a participant went from the NF condition to the FF condition at trial 4, there was an immediate increase in the negative gradient between trials 3 and 4, which is where the change of condition took place. Figure 5 (below) demonstrates that this is a more obvious effect than that seen in Figure 3 above, and the decrease in NP length between trials 3 and 4 is significant here ($t(5) = 2.97$, $p<.05$). None of the other adjacent trials were significantly different ($ps<0.05$).

Figure 5: Shortening of Noun Phrases with increase in feedback



3.4.6 Number of features in noun phrases

One way of assessing how much information is included in descriptions is by counting the number of 'features' (that is, distinct conceptual fields) contained within them. If we compare these numbers over conditions and trials, we can see if the amount of detail provided by the Describer about each block is affected by, firstly, practice at the task and, secondly, the amount of feedback from the Builder. Ten features were identified, and these are listed below (Table 6) with examples.

Table 6: Noted Features of Noun Phrases

Feature	Examples (feature in italics)
Noun	*Bit, piece*
Colour	*Red, yellow*
Size	*Big, small*
Shape	*Square, rectangle*
Relative width	*Wide, narrow*
Actual width	*Two nodes wide, three wide*
Relative length	*Long, short*
Actual length	*Four nodes long, six long*
Total nodes	*Four-node* square

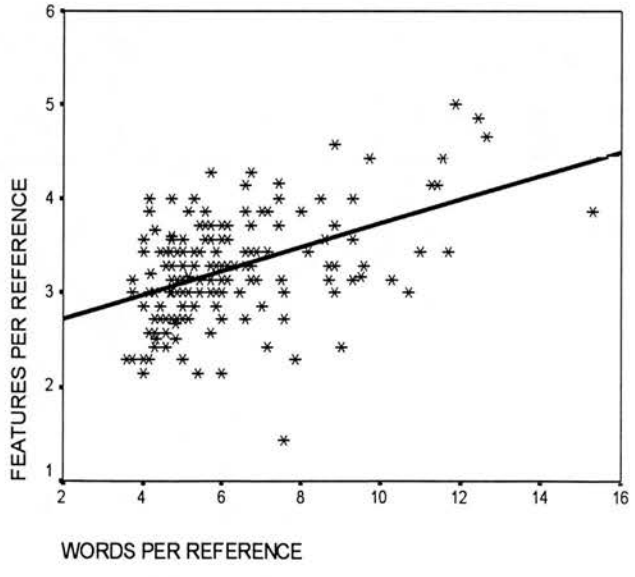
The mean number of features per description for each trial and condition is shown below in Table 7.

Table 7: Mean number of features in each description, per condition, per trial

	Number of features per description		
Order	NF	MF	FF
1.00	4.00	3.34	3.62
2.00	3.48	3.16	3.17
3.00	3.45	3.14	3.13
4.00	3.36	3.20	3.45
5.00	3.44	2.86	3.41
6.00	3.29	3.05	3.64
7.00	3.26	3.21	3.21
8.00	3.32	3.05	3.10
9.00	3.48	2.76	2.62
Mean	3.45	3.09	3.26

3X9 ANOVAs revealed a main effect of Trial Order ($F(8,40) = 2.87$, $p=.01$), where there were fewer features in later descriptions. There was also an effect of Condition ($F(2,10) = 4.46$, $p<.05$). There was no interaction ($F(16,80) = .787$, NS). 2X9 ANOVAs on just the NF and FF conditions showed the same effect of Trial Order ($F(8,40) = 2.23$, $p<.05$) but no effect of Condition ($F(1,5) = 2.07$, NS), suggesting that the effect shown in the 3X9 ANOVAs was due to the results of the MF condition. There was still no interaction. Only the FF condition showed a polynomial trend with Trial Order ($F(1,7) = 2.25$, $p=.01$). The reduction in number of features across trials is presumably partly due to the participants' efforts to maximize the efficiency of their references as they became used to the task. As would be expected, the values for words per NP and features per NP were positively correlated ($r(162) =.461$, $p<.001$, 1-tailed), and Figure 6 (below) demonstrates this relationship.

Figure 6: Correlation between number of features and number of words in Noun Phrases



3.4.7 Definiteness of nouns

Next I analysed whether each block was referred to with a definite ('the') or indefinite ('a') determiner. Blocks that have been previously referred to within a particular trial are very likely to be referred to with a definite determiner (because of their Given status), and so I studied only the first reference to each block, because I wanted to investigate how often New blocks were referred to with Given determiners. Five of the models contained a pair of identical blocks, and reference to these blocks were discounted for the purposes of this analysis, because with these blocks, *indefinite* determiners like 'a' were likely to be used to refer to the first block simply because it wouldn't matter which block the Builder chose, for example "Pick up *a* white square". Similarly, the second identical object would be more likely to be referred to with a *definite* determiner, because it was in a sense Given, as a result of the reference to the previous block, for example, "Now pick up *the* other white square".

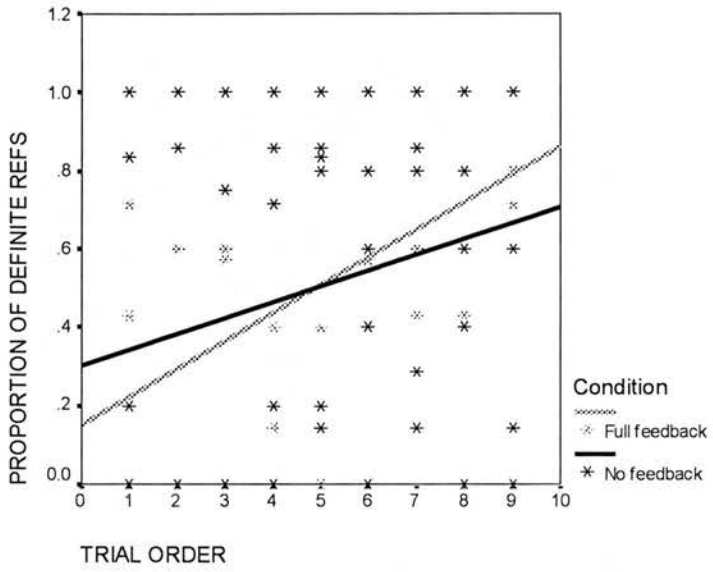
After the references to identical items were discounted, I calculated the number of definite determiners as a proportion of all the remaining determiners, per participant, per trial. The mean proportions of definite determiners are shown below in Table 8.

Table 8: Mean definite determiners as a proportion of all determiners

	Proportion of Definite Determiners		
Order	NF	MF	FF
1.00	.34	.26	.22
2.00	.31	.56	.41
3.00	.29	.40	.65
4.00	.60	.71	.26
5.00	.64	.81	.23
6.00	.63	.73	.26
7.00	.54	.46	.77
8.00	.57	.32	.87
9.00	.62	.41	.89
Mean	.51	.52	.51

3X9 ANOVAs showed that neither Trial Order ($F(8,40) = 1.79$, NS) nor Condition ($F(2,10) = .012$, NS) had a main effect, but that there was an interaction ($F(16,80) = 2.53$, $p<.01$). 2X9 ANOVAs on just the NF and FF conditions showed an effect of Trial Order ($F(8,40) = 2.34$, $p<.05$), where the proportion of definite determiners increased over trials. There was no effect of Condition ($F(1,5) = .000$, NS), but there was still an interaction ($F(8,40) = 2.55$, $p<.05$), demonstrating that the proportion of definite determiners tended to increase more in later trials in the FF condition than the NF condition. A polynomial linear trend was significant between proportion of definite references and trial order in the FF condition ($F(1,7) = 16.12$, $p<.01$) and marginal in the NF condition ($F(1,7) = 3.36$, $p=.07$). The graph below (Figure 7) demonstrates the interaction.

Figure 7: Mean number of definite references as a proportion of the total number of references per trial.



The finding that definite articles were more commonly used for first references in later trials in the Full Feedback condition is not surprising, as it demonstrates an increasing familiarity with the Lego pieces involved. However this is unlikely to simply be an accurate reflection of whether the pieces themselves were Given (i.e. had been seen before) to the Describers. If this had been the case, then we would expect to see the same trend of increasing definiteness occurring in the NF condition too, since the Describers held the same knowledge there. So why does this increasing use of definite determiners happen most notably in the condition with most feedback?

Since all the Lego pieces in front of the Describers will have the status of being Given (by virtue of being physically present; see Bard and Anderson, 1994), it would be the most natural inclination for the Describers to refer to them in a way that is congruent with their Given status, that is, with definite determiners.

However it seems that the Describer will only choose this method of description if he is quite sure that the Builder will understand it. Taking the two opposing poles of feedback, in the NF condition, the Describer cannot ascertain how much the Builder knows or will understand, so he must err on the side of caution in an effort to be as helpful as possible. In the FF condition, in contrast, the Describer knows how much the Builder understands (or at least can find out, as a result of their feedback), and so he can tailor his references to suit the Builder's knowledge. Items must be Given to both the Describer and the Builder before they become part of common ground, and so if the Builder consistently gives informative feedback that demonstrates that he knows the pieces that are being referred to, then the Describer will, over time, begin to refer to these pieces as Given.

These data show that describers in the NF condition frequently used indefinite determiners even in the later trials; 38% of NF last-trial determiners were indefinite (in comparison with 11% in the FF condition). Since the items were just as Given for the Describers in NF as they were in FF at that point, this continuing use of indefinite determiners in the NF condition is likely to be the result of audience design. This is stronger evidence for audience design than many previous studies, because there has often been a confound between what was beneficial for the speaker and what was beneficial for the listener; that is, what seemed to have been produced intentionally for the listener might have just happened to be the easiest thing for the speaker to produce[32].

---

[32] While there is no obvious reason why New and Given references to objects should differ in their intrinsic production difficulty, it is probable that a speaker's computation of his addressee's knowledge of an object will require some effort. This may mean that making a New reference to an item that the speaker considers to be Given will be taxing on his part, at

This is particularly applicable in studies that have demonstrated shortening of noun phrases, both in number of words and in articulatory length (e.g. Clark and Wilkes-Gibbs, 1986; Fowler and Housom, 1987), where brevity in descriptions will have benefited the speaker at least as much as the addressee. In contrast, the present study demonstrates a situation where the Describer departed from his natural method of referring with the apparent intention of producing something that his Builder would understand, thus demonstrating the allocentric type of audience design that was also reported in Wilkes-Gibbs and Clark (1992).

3.4.8 Overall repetitiveness

Next, the overall repetitiveness of the Describers' dialogues was assessed by means of a token/type calculation, which took into account every word (reduced to stems, rather than exact word-forms) in the whole fluent transcription[33]. These means represent the *total* number of words spoken divided by the number of *different* words spoken, calculated per trial, per speaker, and are shown in Table 9 below. Higher values here represent a higher degree of repetitiveness.

---

least more demanding than producing a reference according to his own knowledge would be.

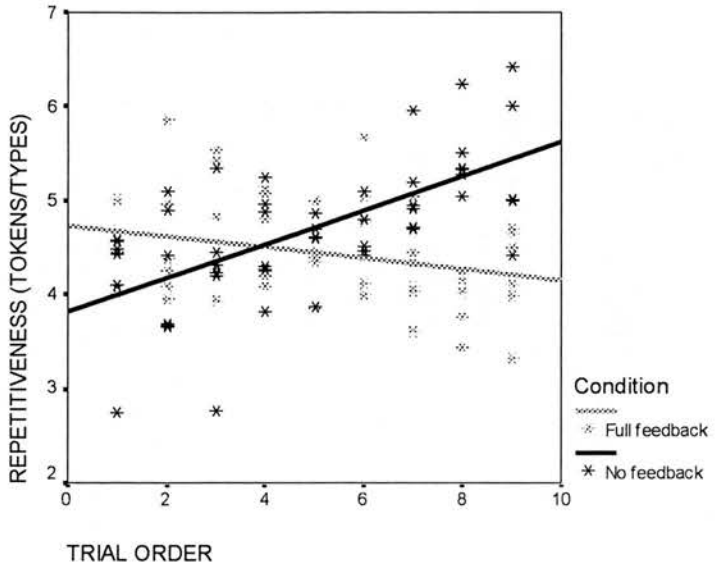
[33] Thanks to Martin Corley for writing and running a Perl program which determined repetitiveness in these transcripts.

Table 9: Repetitiveness of describers' speech

	Repetitiveness (tokens/types)		
ORDER	NF	MF	FF
1.00	4.15	4.50	4.45
2.00	4.24	4.87	4.58
3.00	4.22	4.92	4.70
4.00	4.58	4.85	4.60
5.00	4.56	4.74	4.42
6.00	4.73	5.15	4.78
7.00	5.07	5.03	4.26
8.00	5.46	4.82	3.98
9.00	5.48	8.64	4.21
Mean	4.72	5.28	4.44

3X9 ANOVAs demonstrated a main effect of Trial Order ($F(8,40) = 2.85$, $p<.05$), where the repetitiveness increased in later trials. There was also an effect of Condition ($F(2,10) = 6.31$, $p<.05$), and an interaction ($F(16,80) = 2.25$, $p=.01$). 2X9 ANOVAs on only the NF and FF conditions showed no effect of Order ($F(8,40) = 1.52$, NS), and only a marginal effect of Condition ($F(1,5) = 4.14$, $p=.097$), where NF was slightly more repetitive than FF. The lack of full significance suggests that the MF condition was the locus of the Condition effect in the 3X9 analysis. There was also an interaction ($F(8,40) = 4.77$, $p<.001$), where the NF condition increased more in repetitiveness over trials than the FF condition. Both the NF and FF conditions showed significant linear trends (NF: $F(1,7) = 33.49$, $p<.001$); FF: $F(1,7) = 4.40$, $p<.05$). Figure 8 (below) demonstrates the shape of the interaction.

Figure 8: Mean repetitiveness values (token/type) of the Describers' speech per condition and trial.



The significant increase in repetition that occurred in the NF condition could be a result of the Describer falling into more routinised methods of description, where, with each trial, his method of describing models became more like that of the previous trial. Alternatively, a more useful way of looking at these data might involve assessing why the FF condition *didn't* become more repetitive over time. It is likely that this continuing variability in descriptions was due to the Builder's influence; in the NF condition, the Describer was free to repeat his description style again and again without interference, and without knowing how helpful (or unhelpful, alternatively) it was to the Builder. In the FF condition, however, there were two potential sources of variation: firstly, the Builder might have suggested alternative descriptions, which would constitute an interruption to the Describer's natural tendency to repeat himself from phrase to phrase. Secondly, the Builder may have given other kinds of explicit feedback, for example saying 'I don't understand, can you rephrase that?'. This could have affected the style of phrases

employed by the Describer, making them evolve over time in a manner similar to that displayed in Clark and Wilkes-Gibbs (1986).

3.4.9 Repetitiveness of individual nouns

There are many possible ways of referring to Lego pieces, ranging from the very general 'bit' or 'block', to more specific nouns such as 'square' or 'rectangle'. Describers tended to be somewhat inconsistent in their choice of nouns throughout the experiment, usually preferring to employ many different terms rather than sticking to one or two favourites. The mean number of different nouns used overall was 3.2 per trial in the NF condition, 3.3 in MF and 3.5 in the FF condition. There were no main effects of Trial Order ($F(8,40) = .961$, NS) or Condition ($F(2,10) = 2.46$, NS), nor was there an interaction ($F(16,80) = .728$, NS).

3.4.10 Noun consistency and type

Looking further at the nouns used by Describers, we might find there are differences in the particular *types* of nouns that were used over trials. Since the length of NPs reduced over time, particularly in the FF condition, it seems that the Describers somehow reduced the number of surplus words they used in later trials. This might have involved a reduction in the number of 'empty' nouns (generic nouns which do not denote any physical characteristics, and therefore could apply to any block, for example 'piece' or 'bit') and an increase in more informative nouns (which do denote physical characteristics). The informative nouns used were divided into three common types: those which relate to aspects of the pieces' 'shape' (e.g. square, rectangle, long, short), 'colour' (e.g. red, yellow) and 'specifications', which refer to the actual dimensions of the piece (e.g. four by three, one by six etc).

Table 10 (below) gives the overall proportions of these in the transcripts.

Table 10: Noun types used in NPs

Category	Type	Percentage	Examples
Informative	Shape	13.1	Square
	Colour	7.3	Blue, red
	Specifications	7.7	Three by two
Uninformative	Empty	71.9	Block, Piece

Empty nouns were by far the most common type, accounting for over 70% of the nouns used in descriptions. It is possible that the proportion of these might have decreased over trials, and been replaced with more useful words, so that instead of 'the six-dot red piece', the Describers would have produced the shorter NP, 'the six-dot red'. This would have allowed the Describers to be more efficient in their descriptions. Analyses were carried out to test this hypothesis. The number of empty nouns used was calculated as a proportion of all the different nouns used, per Describer, per trial. Note: Here the *total number* of empty nouns was counted; that is, if 'bit' was used 4 times in one trial, and 'piece' 3 times, this would be counted as 7 empty nouns. Later analyses will consider the number of *different* nouns used, which in this example would be two.

3X9 ANOVAs demonstrated a marginal effect of Trial Order ($F(8,40) = 1.91$, $p=.085$), with fewer 'empty' nouns being used in later trials. There was no main effect of Condition ($F(2,10) = .130$, NS) and no interaction ($F(16,80) = 1.22$, NS). 2X9 ANOVAs on just the NF and FF conditions revealed no significant effects or interactions, suggesting that the number of empty nouns used was relatively constant over feedback conditions and trial order.

3.4.11 Consistency of noun use

Empty nouns (e.g. block) can potentially be applied to any Lego piece, regardless of its shape, colour or size, and so could have been repeated frequently by the speakers if they had decided to do so. Bearing this in mind, I analysed if speakers tended to consistently use one particular empty noun, for example calling everything a 'piece', or whether they used more different nouns in certain conditions or in particular trials. 3X9 ANOVAs showed that the number of different empty nouns (as a proportion of all different nouns) used did not vary significantly with Trial Order ($F(8,40) = 1.23$, NS) or Condition ($F(2,10) = .775$, NS), and there was no interaction ($F(16,80) = .770$, NS). 2X9 ANOVAs on just NF and FF also showed no effects. This demonstrates that Describers were relatively consistent in their choice of nouns over the whole experiment, regardless of feedback.

3.4.12 Order of features in noun phrases

Next I examined the order in which Describers referred to certain features of each Lego block in their descriptions. Each NP had a 'code' allocated to it, which specified which characteristics of the Lego block were mentioned, and in what order. The features coded were the same as those shown in Table 10 above: noun, colour, shape, size, actual specifications (further split into actual length and actual width) and relative specifications (split as before).

Individual codes for the features present were then combined, in the order they were mentioned by the Describer, to produce a code for the whole NP.

For example, the three descriptions below would all have been attributed the code RL/RW/C/N:

RL (relative length)	RW (relative width)	C (colour)	N (noun)
Long	wide	blue	piece
Short	narrow	yellow	bit
Longest	thin	green	block

These codes were then compared to look at consistency of structure. When two of them matched, that is, they contained the same features in the same order, then the noun phrases were deemed to be the same in a structural sense. This matching process allowed us to investigate the routinisation of entire NPs, irrespective of the actual colour, size and shape of the block, just pertaining to the style and particularly the *order* in which their characteristics were presented.

The number of *different* codes used by each Describer in each trial was counted, and then the proportion of different codes, relative to the whole number of NPs produced by that Describer in that trial, was calculated. Again for these analyses I only counted the first reference to each block. No main effect of Trial Order ($F(8,40) = 1.66$, NS) or Condition ($F(2,10) = 1.60$, NS) existed, nor was there an interaction ($F(16,80) = 1.28$, NS). The results from 2X9 ANOVAs on NF and FF were similarly non-significant. It is possible that any effect of Trial Order here was moderated by the reduction in number of features used in later trials (Section 3.4.4.2), which would necessarily reduce the amount of direct code repetition.

3.4.13 Disfluencies

The disfluencies uttered by the Describer were then counted. Five main types of disfluencies were considered in this process, in accordance with the HCRC disfluency coding handbook (definitions and examples for the first four items below taken from Lickley, 1998):

Repetitions: Strings are repeated verbatim with no addition or deletion. The repeated strings may be word-fragments, whole words or sequences of words. e.g. "no [s-] straight southwest"

Substitutions: A word, fragment or string is directly replaced by a word or string of words: some words may be repeated: the string replaced and its replacement share syntactic features. e.g. "and you [s-] finish your curve at..."

Insertions: Some string is repeated with a word or words inserted before or within the repetition. e.g. "[just] I'm just to the east of it"

Deletions: The speaker interrupts their utterance and restarts without repeating or directly substituting any word or structural unit. e.g. "[s-] you then come up north"

Filler Terms: Nonword items used to mark suspensions. e.g. 'um', 'er'.

The number of disfluencies (all types) per 100 fluent words was calculated. These mean values are shown below in Table 11.

Table 11: Mean number of disfluencies produced by the describer per 100 fluent words, per condition, per trial.

	Mean disfluencies per 100 words		
Order	NF	MF	FF
1.00	9.57	8.12	9.91
2.00	8.59	8.07	8.35
3.00	8.13	8.00	7.93
4.00	7.44	9.17	8.13
5.00	8.23	9.31	6.66
6.00	8.15	8.72	6.74
7.00	9.60	6.33	7.04
8.00	10.77	6.04	6.25
9.00	10.17	6.01	7.17
Mean	8.96	7.75	7.58

3X9 ANOVAs revealed that there was no effect of Trial Order ($F(8,40) = .877$, NS). There was a main effect of Condition ($F(2,10) = 4.01$, $p=.05$), but no interaction ($F(16,80) = .534$, NS). 2X9 ANOVAs on only the FF and NF conditions showed a marginal effect of Condition ($F(1,5) = 4.18$, $p=.09$), where there were more disfluencies in NF than in FF, but there was no effect of Trial Order and no interaction.

3.5 Discussion

The results reported above demonstrate that the simple practice of a task can influence its completion in many ways. Task practice enabled participants to complete trials more accurately and quickly, using fewer words and less detailed descriptions. As mentioned earlier, though, some of these effects are likely to be a reflection of the rapidly-reducing number of blocks in the Builder's pile as the experiment proceeds; the fewer Lego blocks the Builder has to choose from, the less information the Describer needs to provide. The amount of feedback received by the Describer played little or no role in these effects.

Other effects were radically influenced by the amount of feedback received by the Describer. The participant pairs' overall competence at the task (as measured by score) and the proportion of definite determiners increased more over the course of the experiment in the FF condition than in the NF condition. The first descriptions of objects produced by Describers were initially longest in the NF condition, and they decreased in length over trials, more notably in NF than in FF, until by the end of the experiment, references were approximately the same length in both conditions. The repetitiveness of the Describers' speech increased during the course of the experiment in NF but not in FF. There were also more disfluencies produced in the NF condition than in FF.

It seems sensible to focus first on those cases where the NF and FF conditions (which equate to monologue and dialogue) differed significantly, before considering how the MF condition fits in.

3.5.1 Less information, but better results

The finding that participant pairs completed their tasks more successfully in the dialogue condition than in the monologue condition corroborates the findings from Kraut et al (1982) and Clark and Krych (2004) that the task performance of addressees increases with their ability to give feedback to the speaker. This occurred despite the speakers using fewer words to describe each Lego piece in the dialogue condition than in the monologue condition. These two factors of score and NP length could plausibly be related; it seems that it may be the speaker's tailoring of speech to his addressee that allows him to be more succinct, and that also aids the addressee in her task.

It seems to be the addressee's *feedback* that allows the speaker to tailor his speech to her requirements. Since the listener is capable of telling the speaker both *what* she needs to know, and *when* she has been given enough information, the speaker can then restrict his descriptions to the key information required. This happened from the very beginning of the current study; there was less shortening of NPs over trials in the dialogue condition than in the monologue condition. This suggests that the speakers in dialogue employed brevity from the very first trial, apparently presuming that their listeners would ask them for more information as required. This was further demonstrated in the finding that description length was affected by the amount of feedback received on almost a moment-by-moment basis, with an increase in feedback from the addressees causing the speakers to shorten descriptions significantly even on the very trial that the additional feedback was introduced (that is, when the conditions were changed).

The idea that utterances were more tailored to addressees in dialogue than in monologue is also supported by the finding that speakers increased their use of definite determiners more over time in dialogue than in monologue. This

suggests that the dialogue speakers were aware of which items were Given and New to their addressees, and took account of this in their speech, whereas monologue speakers continued to be more careful and tentative even towards the end of the experiment.

As mentioned earlier, the finding that new NPs in monologue shortened more than those in dialogue conflicts with findings from Krauss and Weinheimer (1966) on repeated references to the same items. This may result from a difference between the NPs in question; they studied repeated references to the same item, whereas I looked at references to new items. The lesser amount of shortening in the dialogue condition of my experiment may be a consequence of the lack of common ground built up; common ground which may have caused the shortening in Krauss and Weinheimer's study. However there are also two major differences between my task and theirs, either of which (or perhaps both) can account for the difference in results. It seems that the extent to which speakers shorten their descriptions with feedback may be strongly influenced by the type of item they are describing. This may be because when a speaker is describing ambiguous figures (such as tangrams), his descriptions start out being longer than when the figures are unambiguous (such as Lego blocks), and so there is more room for shortening in that situation. Alternatively, the speaker's awareness of the addressee's ability to give feedback (or not) may also be crucially important. Where Krauss and Weinheimers' speakers may have been unaware of the restrictions placed upon their partners[34], ours were fully aware, as both the Describers and Builders were told of the communicative restrictions placed upon the Builders.

---

[34] It is unclear from the paper whether speakers were aware of their addressees' feedback restrictions or not.

Speakers in the dialogue condition here seemed to produce fewer disfluencies than those in monologue. In general, prepared monologues are recorded as being more fluent than spontaneous dialogue (see Schachter et al, 1991, for an example of lectures versus informal discussions); however, it seems likely that the key difference is the level of preparedness. The monologue speakers in this task did not get a chance to prepare what they would say (any more so than the dialogue speakers did), and it seems that, according to this study, unprepared monologues were more difficult for speakers to produce than unprepared dialogues, resulting in their production of more filler terms and disfluencies. One possible reason for this could be that, in a dialogue, a speaker has time to plan while his addressee speaks. In contrast, monological speech is more constant and contains fewer enforced breaks, and this will be particularly salient in a time-restricted task like this one. The time constraint imposed here may have influenced the amount of self-monitoring being carried out to some extent, potentially affecting fluency[35]. Although the time restriction here will have affected the monologue and dialogue conditions equally, the speakers in dialogue will have benefited from their partners' speaking time, whereas the monologue speakers will have had no such opportunity. Perhaps a task without time restraint would produce differing results from the current study. Additionally, it could be proposed that the burden of responsibility on the speaker was less in dialogue here than in monologue, because the addressee was a joint partner, and carried out some of the work. This could have occurred in two main ways; firstly, the addressee's contributions may have provided linguistic material (for example, partial or whole descriptions) which could then be re-used by the speaker, and secondly, her contributions could have helped the speaker to know which part of his descriptions were most helpful to her, and so were worth

---

[35] Horton and Keysar (1996) found that less common ground self-monitoring occurs under time pressure, and this may also be applicable to articulation processes. However, Oomen and Postma (2001) proposed that the self-monitor increases speed to keep up with production.

repeating in future descriptions. These two aspects of the addressee's speech may have reduced the cognitive load on the speaker, and increased his fluency. Alternatively, the finding of more disfluencies in monologue than in dialogue could be a result of more practical influences. Oviatt (1995) found that longer utterances in dialogue tend to contain more disfluencies than shorter utterances (although length of utterance was not analysed in the current study). If the utterances in monologue were typically longer than those in dialogue, then this alone might account for the difference in fluency, regardless of the amount of feedback received. From yet another point of view, the added disfluency in monologue might be simply due to the speakers being in the somewhat unusual situation of producing a spontaneous monologue, and feeling embarrassed or nervous about what they were saying, whereas participating in a spontaneous dialogue is certainly a situation with which we are all more familiar.

A brief look at those factors which were *not* affected by the presence or absence of feedback can also tell us a little about the production system, although since these are effectively null results, any conclusions drawn must be tentative. It appears that feedback does not affect which particular nouns are employed in referring to blocks or the exact structure (in terms of order of features) of NPs. In conjunction with the effects reported above, these findings suggest that feedback does not have a large influence on word choice; it may affect quantitative aspects of speech more than qualitative, and the number of words more than their actual identity.

3.5.2 Does quantity of feedback matter?

So what of the MF condition? Did it behave more like monologue or dialogue? Is it simply the *presence* of feedback that is key to these effects, or does the *quantity* matter too? The results here seem to be quite mixed. The mean task score attained by addressees in this condition was roughly halfway between the NF and FF

conditions, but did not differ significantly from either of them. This suggests that while the minimal feedback that was allowed in this condition was slightly beneficial, it wasn't as useful as full interaction. This is the type of result that would have been expected in most of the measurements, but in fact some of them differed greatly from my expectations.

It is particularly intriguing that when the total words per trial were measured, the MF condition came out to have significantly more words than either the NF or FF conditions. Why should this be the case? Perhaps because the addressee in these circumstances could let the speaker know that she needed more information (for example by saying 'expand'), but she wasn't able to say exactly what she needed to know. The speaker in this situation may have simply said everything he could think of in an attempt to help her. In contrast, in the FF condition the addressee could ask specifically for only the information he needed, which would have helped the speaker to be succinct, and in the NF condition, the speaker would not even have known that the addressee needed more information, and so would not have elaborated on his initial descriptions.

In terms of the length of NPs, the MF condition had a mean value that was similar to that of the FF condition, and as in the FF condition, the NPs did not reduce in length over trials as much as in the NF condition. In this case it seems that the minimal amount of feedback was as good as complete interaction, presumably because the speaker had a rough idea of whether the addressee understood him or not, and this understanding constituted permission to the speaker to use more brief NPs. With regard to the number of features included in NPs, the MF condition also acted like the FF condition, but did not reduce over time as FF did. This is somewhat surprising, since the same logic applied above – regarding speakers shortening NPs when they receive affirmation – should still apply here.

The mean repetitiveness of the MF condition was similar to that of the NF condition, and also increased just like in NF. This is perhaps because there was not enough content in the addressee's utterances to actively adjust the speaker's descriptions by suggesting alternatives. Moreover, the restricted content of the MF utterances will not have interrupted the speaker's repetition of his descriptions to as great an extent as in the dialogue condition. One way in which the MF condition did meet my expectations is that speakers in the MF condition used progressively more informative nouns as the trials proceeded (unlike in the other two conditions). This is an effect I would have expected to see in the dialogue condition as well, an indication of exactly how the speakers are reducing the length of their NPs.

The MF condition was comparable to the dialogue condition in one more sense – the number of disfluencies present. It seems that the speakers simply found the MF situation easier to speak in than the apparently difficult NF condition. It seems from these results that the small amount of feedback allowed in the MF condition was *almost* sufficient to deem the speakers' contributions part of a dialogue, rather than a monologue. That is, the MF condition reflected dialogue more than monologue, although crucially, it didn't produce scores as high as those in dialogue. So it seems that the tentative conclusion we can reach here is that the presence of feedback is more important than its quantity, but that the quantity still has an effect on its benefit.

3.5.3 Designed for dialogue?

The high scores in dialogue, as compared to monologue, say a lot about the benefit of feedback for interlocutors. Garrod and Pickering (2004) postulate that dialogue is the form of language use for which we are 'designed', and the results of the current experiment certainly seem to corroborate this, demonstrating that communication is more effective with feedback, at least in terms of the task scores attained. So how

should these results influence our interpretation of the previous findings on monologue? The main thing to consider is that there are many levels on which dialogue and monologue are not comparable, and as such, the findings on monologue which have comprised the bulk of psycholinguistic research cannot be naturally extended to encompass dialogue. Any interpretation of those results must take into account the fact that asking a participant to listen to, or produce, a monologue is not comparable to their taking part in a 'normal' dialogue, and that those factors which seem to taint our study of language in real-life scenarios may well play a key role in our use of language in everyday contexts.

Chapter 4: Experiment 2: Tangram referential-communication task

4.1 Chapter Overview

The previous chapter demonstrated some of the ways in which feedback from a partner can influence the form of language produced by a speaker. This chapter presents a second experiment in the same vein, that is, a referential communication task comparing conditions with differing amounts of feedback, but this time using Tangram shapes instead of Lego blocks. I look in detail at the shortening of noun phrases with repetition, and the consistency of tangram descriptions by the same speaker. I then attempt to determine how these two processes relate to each other and if either or both of them are affected by feedback from an interlocutor. Additionally, the oral descriptions of tangrams recorded in this experiment will provide materials for Experiments 3 and 4, in which new groups of participants will overhear experimentally manipulated versions of these descriptions and will attempt to select which tangram was being described.

4.2 Introduction

Shortening

When an object is referred to more than once by a speaker, their manner of referring often changes with repetition, sometimes altering quite subtly and at other times in obvious ways. For example, Clark and Wilkes-Gibbs (1986) found that when tangrams are described several times in a referential communication task, the

number of words used to describe them reduces with repetition. Additionally, in cases where exactly the same descriptions are repeated, the articulatory lengths of these noun phrases are shortened with repetition, meaning that they tend to be said more quickly on the second pronunciation than the first, independently of the overall speaking rate (Fowler, 1988; Fowler and Housum, 1987). It appears at first glance that these two repetition effects might result from the same underlying process. However it seems that while articulatory shortening occurs even when speaking to a 'new' partner who has not heard the previous references (Bard and Aylett, 2003), lexical shortening of descriptions does not occur in these circumstances (Wilkes-Gibbs and Clark, 1992), suggesting that it may be more under the speakers' control than articulatory shortening. Clark and Wilkes-Gibbs suggest that lexical shortening is employed as a type of audience design, where the speaker adjusts his speech according to what he believes his addressee needs to know, and rids repeated descriptions of unnecessary information as the common ground which is accepted between the partners grows.

Consistency

A second characteristic of item descriptions is that speakers tend to re-use some of the same words when describing the same object again. That is, "The man who's ice-skating" is more likely to become "The ice-skater" on the second description than, for example, "The man who's dancing" (Clark and Wilkes-Gibbs, 1986). This may reflect to some extent the manner in which objects are conceptualised by speakers; a concept, once formulated, will be relatively static, and ensuing references to the same object will tend to employ the same concept again and again. Repetition of this type could be referred to as 'consistency', where each description can be judged as more or less consistent with (that is, similar to) the previous description.

One of the main purposes of this chapter is to investigate how feedback from an interlocutor affects repeated noun phrases in terms of, firstly, their shortening, and secondly, their consistency. Lexical shortening of descriptions can apparently be affected by feedback from an interlocutor; Krauss and Weinheimer (1966) found that descriptions of abstract objects were shortened more on repetition when the speaker received verbal feedback from a partner than when they didn't. Hadelich and colleagues (Hadelich, Branigan, Pickering and Crocker, 2004) found that a similar effect occurred with visual feedback (that is, being able to see the partner's facial expressions) as well as verbal feedback[36].

Turning our attention to the effect of feedback on consistency, there is abundant evidence that lexical priming, which is essentially a measure of consistency, occurs in both within-speaker (from production to production: Wheeldon and Monsell, 1992) and between-speaker (from comprehension to production: Brennan and Clark, 1996; Garrod and Anderson, 1987) contexts. This proves that feedback from a partner certainly does not eliminate speakers' consistency altogether. However Hadelich et al (2004) reported that there was less lexical overlap of between-partner descriptions in their verbal-feedback condition than in the no-feedback condition, suggesting that feedback may have reduced speakers' consistency somewhat. This finding contrasts with Pickering's (2005) theory of how feedback might influence consistency between partners, which here is referred to as alignment (cf. Pickering

---

[36] Although they studied the length of first intonational phrases rather than whole descriptions, and their analyses were of between-partner, rather than within-partner descriptions.

and Garrod, 2004, summarised in Section 2.10). He suggests that alignment between partners occurs primarily by a priming mechanism, and that feedback contributes independently to this, providing a 'meta-commentary' by which the addressee tells the speaker how successful the alignment is being (and in a sense, how useful his descriptions are to the addressee). According to this account, facilitation occurs in three ways. Feedback allows the addressee to, firstly, tell the speaker that she is aligned with him, secondly, tell him she is not aligned, and thirdly, in the most extreme situation, tell him that his situation model is wrong. Thus feedback benefits the interaction as a whole, but nonetheless alignment between interlocutors can still occur to some extent without it. Although Pickering does not directly analyse or comment upon the effect of feedback on *within-partner* consistency, if the mechanisms determining alignment between interlocutors and consistency within speakers are related, then it is very plausible that the effect of feedback on both these processes will be similar.

In contrast with Pickering, Clark and colleagues propose that the consistency of descriptions over repetition (in their terms, 'lexical entrainment') is a result of the formation of a 'conceptual pact' between interlocutors, and emphasise that this is a joint construct, in which both partners' contributions are necessary (see Chapter 2 for more details). Brennan and Clark (1996) proposed that when conceptual pacts are created between partners, "One consequence is lexical entrainment, the repeated use of the same or closely related terms in referring to an object on successive occasions" (p1491). Although again they did not make any specific predictions about *within-partner* consistency, one can infer that, if this involves a similar mechanism to between-partner entrainment, then speakers' descriptions should be less consistent in the absence of feedback from an interlocutor, because no conceptual pact can be formed. Thus these two accounts - Clark's and Pickering's - might have differing predictions about the role of feedback in the present experiment.

Interestingly, neither of the theories above support the idea that speakers will be more consistent in monologue conditions than in dialogue, which was the result found by Hadelich et al (2004) in a between-partners context. This could happen because speakers might 'play it safe' to try and minimize the risk of misunderstandings by their partners. Another possibility is that the presence of feedback may disrupt the self-priming mechanism in the speaker, stopping them from simply repeating the same descriptions again and again[37], or alternatively it might be a result of the speaker having split attention in dialogue (relating to his partner's and his own productions) but not in monologue. One more potential reason why consistency might be greater without feedback than with is that a speaker in dialogue may not try so hard to remember his previous description of a given tangram because he knows it is not so crucial for the addressee to understand him first time, as she has the ability to ask questions if she doesn't understand.

4.2.1 Rationale for Experiment 2

The previous experiment examined the role of feedback on many detailed aspects of language production and on the participants' overall success at the task. In those analyses, I found that first descriptions of Lego blocks in the No Feedback condition were significantly longer than those in Full Feedback. However, the length of first descriptions reduced more in NF than in FF over the course of the experiment, resulting in first descriptions that were approximately the same length in both these feedback conditions by the end of the experiment. These results initially appeared to conflict with the finding of Krauss and Weinheimer (1966) that descriptions produced in dialogue shorten more than those in monologue. However the

---

[37] However Wheeldon and Monsell (1992) found that lexical repetition priming in a picture-describing context persisted even after 100 intervening word productions.

analyses of these experiments were not directly comparable, since in Experiment 1, only the first description of each Lego block was analysed, in contrast with Krauss and Weinheimer's task, where they analysed a number of descriptions of the same items. That is, my analyses demonstrated increasing brevity in the speakers' descriptions of items for the first time, whereas theirs showed brevity occurring as a result of having described the same item previously.

The present experiment will again involve feedback conditions which are similar to Krauss and Weinheimer (1966), but this time, will analyse repeated references to objects. In Krauss and Weinheimer's No Feedback condition, the speakers may not have known that their addressees could not give them any feedback (and additionally, the addressees may not have known that the speakers could not hear their feedback; this was certainly the case in Kraut et al, 1982), which may have added an element of confusion into both roles. Using similar feedback conditions, but where both partners are aware of the restriction on feedback in the NF condition, may produce a different pattern of results. Nevertheless, my hypothesis on description length agrees with their finding; that is, that the descriptions in the FF condition will reduce in length more than those in the NF condition, perhaps simply because the speaker in the FF condition can receive confirmation from his partner that she understands his descriptions, and so he might feel more confident about shortening these descriptions.

It is important to note, though, that there are several other potential reasons besides that mentioned above why more shortening might occur in FF than in NF. Firstly, a speaker in a dialogue might tend to produce overly succinct first descriptions with the intention that the addressee will ask for more information if she needs it. Secondly, in dialogue the addressee can interrupt the speaker's descriptions at any point, cutting off his description before the end of it, and these interruptions could feasibly occur at earlier and earlier points as the experiment proceeds and the

addressee gains confidence in her choices. This would give a false indication of shortening, which would be nothing to do with the speaker's plans and more a result of the addressee's behaviour. For this reason my aim will be simply to find out what the pattern of shortening is in this kind of paradigm, and how it is affected by the presence or absence of feedback, rather than attempting to discern its exact cause.

Another purpose of the current study is to analyse what role, if any, feedback plays in a speaker's consistency of referring expressions with repetition. I expect consistency of descriptions to increase during the course of the experiment, as the speakers become more familiar with their task, and with the tangram figures they are describing. If my inferences based on the theory by Pickering (2005) are correct, then feedback from an interlocutor will only play a minimal role in consistency (as he expects it to do in between-partner alignment). However if Clark and colleagues are correct, then feedback will be crucial for consistency, as according to this theory, consistency depends upon the formation of conceptual pacts between interlocutors.

The hypotheses for Experiment 2 are as follows:

1. Descriptions of tangrams will be shortened more during the course of the experiment by speakers in the dialogue (FF) condition than in the monologue (NF) condition, as found by Krauss and Weinheimer (1966).

2. The overall consistency of tangram descriptions will increase during the course of the experiment.

3. The consistency of tangram descriptions will not differ significantly between the monologue and dialogue conditions.

4.3 Experiment 2

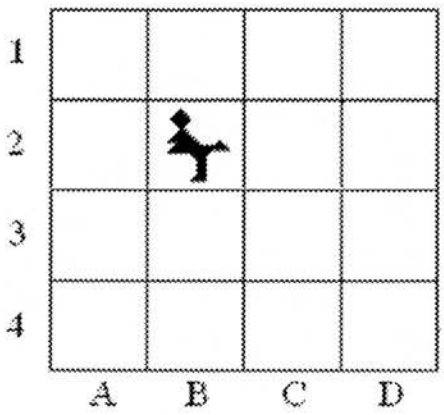
Participants

16 pairs of participants took part in the experiment, in exchange for payment. All were undergraduate and postgraduate students at the University of Edinburgh, and none of the pairs knew each other prior to the experiment. One member of each pair was randomly designated the role of 'Speaker', and the other, 'Addressee'.

Materials

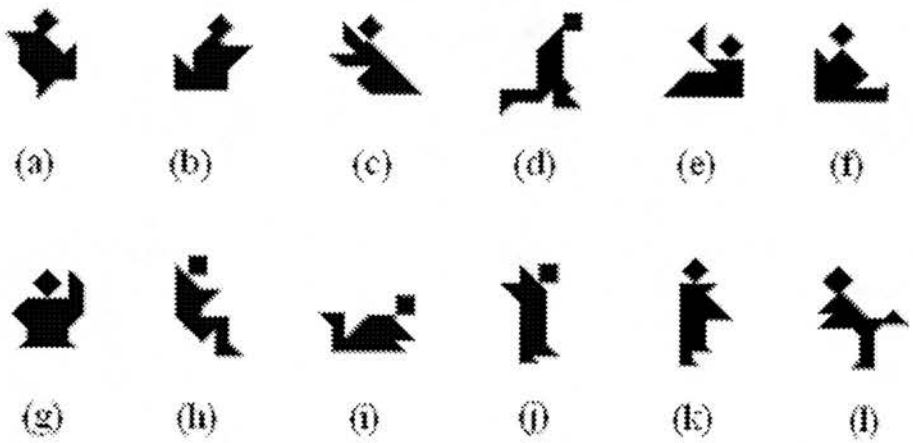
The speaker was given an A4 file with 48 pages in it, 24 for each game. Each page contained a grid showing the position of a particular tangram in one of the squares; all the other grid squares were empty. The grid squares were labelled A-D along the bottom and 1-4 up the side, as shown below in Figure 9.

Figure 9: Addressee grid used in Experiment 2



The addressee was given a set of 12 different tangram cards (Elffers, 1976), shown below in Figure 10, and a larger, empty version of the speaker's grid, approximately 1 metre squared.

Figure 10: Tangram figures used in Experiment 2



The participants were seated at separate tables in the experimental room, facing opposite directions, so that they could hear each other's speech, but could not see each other or each other's materials.

The participants were told that the experiment was designed 'to see how people communicate when they can't see each other'. In half the pairs, the addressee was not allowed to talk, so that the speaker was effectively producing a monologue. In the other half, both participants were allowed to speak with no restrictions, in order to produce a normal dialogue. Therefore this variable of 'feedback level' (No Feedback versus Full Feedback) was run between participants. They were both given instructions (reproduced in Appendix C) and were allowed to ask the experimenter any questions regarding the task.

The speaker was given the experimental folder containing the picture arrays, and the addressee was given 12 tangram cards laid out beside the empty grid. The speaker was also given the full array of tangrams laid out on the table, in order to facilitate descriptions and understand his partner's feedback (for example, to allow him to say "not the man with his arm out to the right, the other one"). In the No Feedback condition, only the speaker was given a lapel microphone, which was attached to a DAT recorder, and he also wore a pair of sound attenuating earmuffs to block out the sound of the addressee placing her cards. This was deemed necessary because Schober (1993) found in pilot tests that experimental participants in a similar condition used the sounds of their partners' pens scratching as a kind of feedback, so I wanted to eliminate all possible sources of feedback; verbal and non-verbal. Unlike in the NF condition in Experiment 1, there was no feedback at all; the addressee could not indicate to the speaker when she had completed a move. In the Full Feedback condition, no earmuffs were worn and both participants wore

lapel microphones; these were coupled before inputting into a DAT recorder, producing a sound file in which each participant was recorded on a separate channel.

When the experiment began, the procedure was as follows: the speaker would look at the first page in the file, and describe the tangram shown and its position on the grid. The addressee would try to find the same tangram in his set (with or without discussion, depending on the condition), and would put it in the correct position on the grid. Then the speaker would turn over to the next page, describe the next tangram's position, and so on until the 12 tangrams were filling 12 spaces on the addressee's grid, leaving just 4 empty spaces on the grid. Then 36 more moves would be described, such that each tangram was moved three times to different positions on the grid. These were always moved to empty spaces, so there was no overlap of tangrams. The order of tangrams to be moved was pseudo-randomised (with the restriction that in every set of 12 moves, each tangram would be moved once), so that the participants couldn't refer to them as 'the first one' or anything similar. Pre-tests showed that Speakers never referred to the positions of the tangrams instead of describing them (e.g. 'move the one that's in A3 to B4), possibly because they could not remember their positions. Speakers were not allowed to turn back to previous pages in the folder to check the previous positions.

Each experimental pair played two separate 'games', using the same tangrams but in a different order. After the first game was over, and 48 movements had been made, the experimenter removed the tangrams from the grid, shuffled them and laid them beside the grid again, in a manner similar to the beginning of the experiment (although in a different order). The speaker was then instructed to continue on to the second half of the file, and 48 more (different) movements of the tangrams were made.

4.4 Results

Qualitative findings

The styles of descriptions used varied substantially between speakers. As an illustration of this, the tangram below was described by two different speakers in the No Feedback condition as:

1) "The next one's kindof like a trapezium shape on the bottom with a diamond shape on top of the trapezium and a triangle to the left of the diamond shape".

And...

2) "The next one looks also like a man popping out of a grave or something, but he's waving to you, and the grave being the dark bit on the ground. Or someone lying on their, their side but they somehow managed to get their, their back at a ninety degree angle to the rest of their spine. But they're still waving".

Description 1) could be described as a feature-focused point of view, with descriptions of the geometrical shapes involved in the whole picture. Description 2), however, suggests that the speaker was viewing the picture as a more figurative image, and was assessing the whole picture together. Quite often types 1) and 2) were blended, so there were descriptions like:

3) "It's maybe someone lying down, the head's like a diamond, at the top right, and there's a triangle at the top left, which sortof seems separated from the rest of the figure. And there's a right angle at the bottom right"

4.4.1 Collaboration in descriptions

There was a lot of collaboration evident in the current Full Feedback condition transcripts, where partners worked together, adjusting each others utterances until a description that satisfied them both was formed, as in Clark and Wilkes-Gibbs (1986). An example is shown in the dialogue below. In this, the '1st reference' refers to the first time the pair have to place the tangram, and the '2nd reference' is the next time they see it, when they are required to move it to a different square on the grid. This dialogue refers to the tangram below:

1st reference

Speaker: Okay the next one, there's two long lines finishing in a big spike at the bottom right, and a diamond at the top middle, and looks like two arms to the left
Addressee: Is it like he's sort of praying to something, kindof thing
S: Sortof
L: Sortof like bowing down
S: Yeah it looks like that, there's an arm pointing to the left and an arm pointing to the upper left
L: Yeah got it

2nd reference

S: Person praying I think I called it
L: Yeah
S: Looking towards the left
L: Yeah

In this description, the speaker's use of the phrase "I think I called it" demonstrates that the pair's first discussion about the tangram laid the foundation for their subsequent descriptions[38].

Quantitative Findings

4.4.2 Number of turns taken

I then analysed the number of turns taken between speaker and addressee on any given reference in the Full Feedback condition. A turn was defined as any contribution of one word or more, with the exception of those which only resulted in concurrent speech between the partners and did not result in the interrupter taking the floor. Lone non-word fillers like 'um', 'er' etc were not considered to be turns.

---

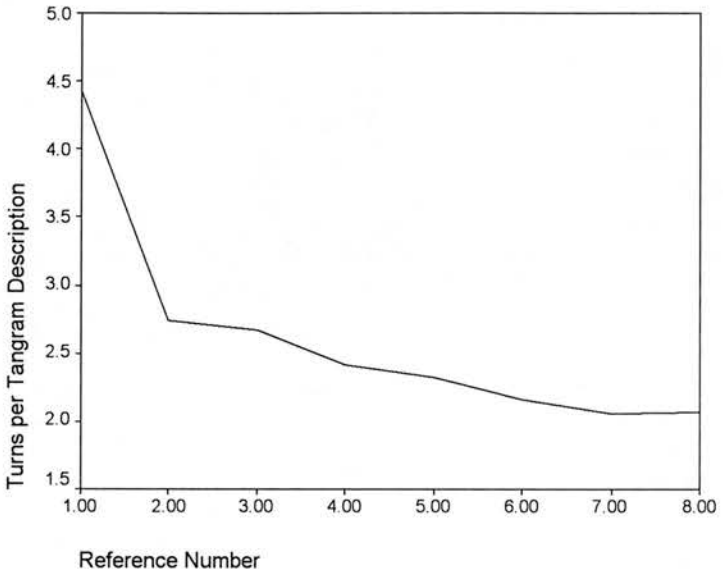
[38] It is particularly interesting that in this case, the speaker imagined that he was the one who proposed that the tangram was like a person praying (when in fact this was the speaker's suggestion) – this suggests that, due to the joint nature of the interaction, some elements of the discussion had become mutual knowledge without any definite record of who contributed them.

Figure 11 below shows how the number of turns taken decreased in later descriptions. The minimum number of 2 turns shown in the graph represents a speaker's description and his addressee's agreement, e.g.

S: Facing towards the left looking upwards but with hands like they're not separated, just all one
L: Okay

Any higher number of turns represents another partner response; for example, a third turn would have consisted of the speaker responding to the addressee's utterance once more. On the graph, Reference Number refers to the number of times that tangram has been described during the experiment, hence Reference Number 1 is the first time it was described, and Reference Number 8 is the eighth (and last time) it was described.

Figure 11: Mean number of turns taken per description by participant pairs in the Full Feedback condition



It is clear from the graph that the number of turns taken in FF decreased radically over the course of the experiment, most notably between the first and second references. A repeated-measures ANOVA demonstrated a main effect of Reference Number ($F1(7, 49) = 11.13$, $p<.01$; $F2(7,77) = 96.62$, $p<.01$). Paired t-tests demonstrated that the difference in number of turns taken between the first and second references was significant ($t1(7) = 4.12$, $p<.01$; $t2(11) = 6.37$, $p<.01$), with more turns being taken in the first references. None of the other adjacent pairs showed significant decreases (all $Fs<1$). This suggests that the majority of interactive communication, in terms of number of turns, occurred the first time the partners referred to each tangram, and this strategy presumably facilitated the production of more efficient later references from the second reference onwards.

4.4.3 Length of descriptions of tangrams

The number of words used by the speaker to describe each tangram each time it was referred to in the NF and FF conditions was counted. The rules used for determining length of descriptions are shown below.

1. Descriptions were considered to begin with the first determiner used in that description (e.g. the, a, this).

2. Descriptions included any reference to the picture or the card, rather than being only focussed on the image itself. For example when a describer began with 'This picture looks like...' or 'The next card is...' then those words were counted as part of the description.

3. Descriptions did not include locative phrases which determined the placement of the card on the grid, connectives which linked these to the description proper, or pronouns or other words which were only made necessary by the intervention of a locative phrase. For example, for an utterance 'The chicken *is in 3A*', only 'The chicken' would be counted as the description, or for 'The next card *is in 3A and it* looks like a chicken', the counted phrase would be 'The next card looks like a chicken').

All turns in the FF condition were included; that is, even if the addressee interrupted the description, all speaker words after that were included too.

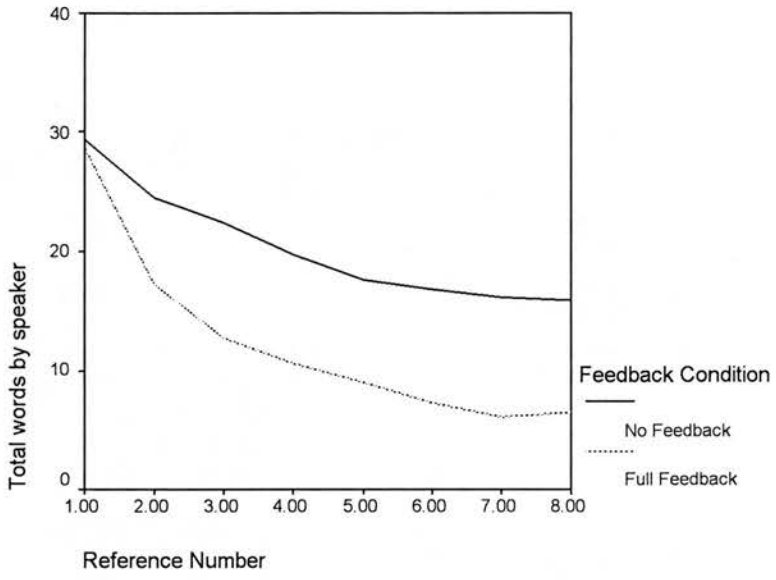
The mean numbers of words used to describe each tangram are shown below in Table 12.

Table 12: Mean number of words used by speaker per description – all turns.

	Mean number of words	
Reference No.	NF	FF
1.00	29.35	28.87
2.00	24.49	17.24
3.00	22.34	12.76
4.00	19.72	10.59
5.00	17.64	9.07
6.00	16.77	7.25
7.00	16.13	6.09
8.00	15.87	6.52
Mean	20.29	12.30

2 (Feedback condition, between-participants, within-items) X 8 (Reference number, within-participants, within-items) repeated-measures ANOVAs demonstrated main effects of Reference Number ($F1(7,98) = 37.78$, $p<.01$, $F2(7,77) = 134.10$, $p<.01$), and Condition ($F1(1,14) = 5.76$, $p<.05$; $F2(1,11) = 70.09$, $p<.01$) and an interaction ($F1(7,98) = 2.45$, $p<.05$; $F2(7,77) = 7.03$, $p<.01$). The interaction is demonstrated by the graph below (Figure 12).

Figure 12: Mean number of words used by speaker per description– all turns



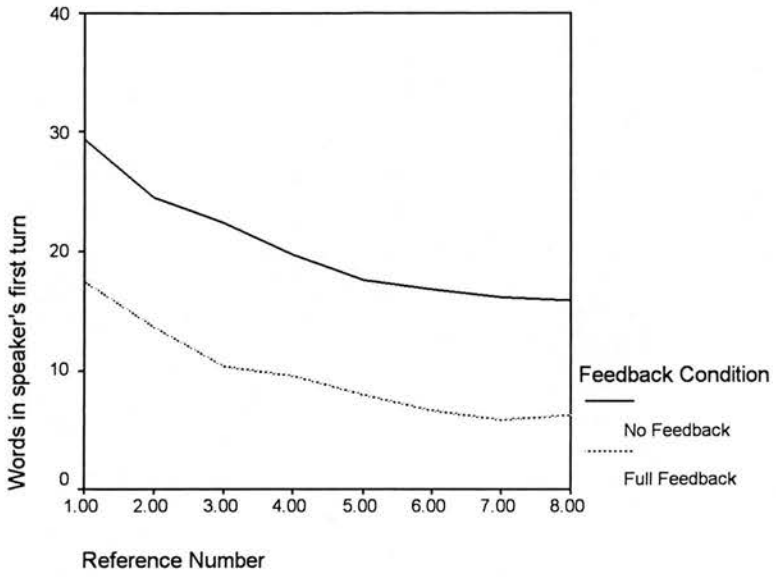
In the *first* reference to each tangram (i.e. reference 1 in the graph), independent t-tests showed that the number of words did not differ significantly between monologue and dialogue ($t1(14) = .126$, NS; $t2(22) = .183$, NS), and it is evident that following this, FF reduced more than NF over the course of the experiment, concurring with Krauss and Weinheimer (1966). However, it is possible that some of the words in the dialogue condition may be a result of the speakers answering the addressees' queries, in which case their content would be more determined by the addressees than by the speakers. One way of evading that problem with these data could be to only look at the number of words in the speaker's first turn in each reference, in this case measured *until the addressee begins to speak.*

For example, take the transcript below:

S: Like the torso of a person with the right arm sticking up (12 words)
L: Well would it be their left arm?
S: If they were facing you I guess (7 words)
L: Facing us yeah, right fair enough.

When we count the total number of words said by the speaker in *all turns* (as in the previous analysis), this would be 19 words. However in the present analyses I measured only the number of words in the speaker's *first turn*, which in this example is 12 words. Figure 13 below demonstrates how the number of speaker words in the first turn of dialogue compares with the number of words in monologue.

Figure 13: Mean number of words used by speaker per description– first turn only



Now we see that there was little difference in the reduction of NPs over repetitions in the two conditions, and in keeping with this, there only an interaction between the main effects of Reference Number and Condition in the by-items analysis ($F1(7,98) = .37$, NS; $F2(7,77) = 2.31$, $p<.05$). There were still main effects of Condition ($F1(1,14) = 10.42$, $p<.01$; $F2(1,11) = 177.2$, $p<.01$) and Reference Number ($F1(7,98) = 22.89$, $p<.01$; $F2(7,77) = 96.35$, $p<.01$).

4.4.4 Consistency of descriptions

To analyse the extent to which Speakers became consistent in their descriptions of tangrams, the number of exact word-types that were common or 'shared' from one Speaker reference to the next was counted[39]. The equation used was:

$$\text{Consistency} = \frac{\text{Number of word-types in N+1 that are also in N}}{\text{Number of word types in N+1}}$$

The steps involved in this were as follows:

1) The number of word-types that were shared between the first reference of a tangram and the second reference to that tangram (by the same speaker) was calculated, then the number of word-types shared between the second and third references was calculated and so on, until all the adjacent pairs of utterances had been compared.

For example, one Speaker referred to the tangram below for the first time as:

Reference 1: "This guy that's like resting on his stomach with his two arms in front of him with his legs lifted up with the feet turned up"

---

[39] Thanks to anonymous programmers on the Excel Forum online (www.mrexcel.com) for writing formulae and macros that enabled me to count strings and compare pairs of strings in Excel.

And the next time...

Reference 2: "It's the guy resting on his tummy with the leg bent up"

The number of word-types shared between these two descriptions would be counted as 8 (underlined: is, the, guy, resting, on, his, with, up). The descriptions used were those reported in the previous section, but rather than using the overall number of words in each description to calculate proportions, I used the number of *word-types* in each description, meaning again that any repeated words would be counted as only one token. This is in keeping with the manner of calculation of shared word-types (but means that the apparent length of utterances here will differ from those in Figure 13).

Note: for the purposes of these analyses, all contractions (e.g. it's, she's, there's, isn't etc) were expanded to their full forms, to prevent situations where a contraction in one description and an expansion in another appeared to produce a different word count (for example, 'it's', as one word, could become 'it is' in a repetition, which is two words, apparently increasing the length of the description, although these may arguably be considered to be the same lexical item. However differing inflections (such as 'chicken' versus 'chickens') were considered to be different items; only lexically identical tokens were considered to be the same type.

In the NF condition, since the speaker only produced one utterance for each description (because there were no interjections from the addressee), the whole utterance was taken into account. In the FF condition, only the first utterance from the speaker was taken into account, until the addressee first spoke. This was in order to reduce the immediate influence of the addressee on the exact words used. For example, the speaker might have omitted something from his initial description, and then the addressee would ask him about it, which would involve the addressee determining the content of the speaker's speech. Instead I wanted to analyse

what details the speaker's *initial* description contained, to determine exactly what kind of description they thought was sufficient to describe each item.

2) The number of shared word-types was then turned into a proportion of the overall number of word-types in the second utterance of each pair.

There were two potential ways of doing this. The first method involved analysing how many word-types were common to both descriptions as a proportion of the number in the *first* description of the pair, and the second method involved analysing how many were common as a proportion of the number in the *second* description.

e.g. Two successive descriptions of the tangram below are given beneath it.

Ref 1: Guy who looks like he is going down on one knee, towards the right hand side
Ref 2: Guy down on one knee

In this example, there were 5 words in common between the descriptions. Method 1 would produce a consistency value of 5 words/16 words = .31. Method 2 would produce 5 words/5 words = 1.0. It was decided to employ Method 2, because as shown earlier, the number of words in descriptions diminished overall with repetition, and employing Method 1 would mean that the greater the shortening, the less consistent the descriptions would appear to be (assuming a constant amount

of repetition), which was not necessarily an accurate indication of the type of consistency I wanted to measure[40].

The table below (Table 13) demonstrates how the amount of consistency increased in later references. Because there was no description for the first reference to be compared with, the second reference is the first one with a value for consistency.

Table 13: Proportions of word-types shared with previous description, per reference number in NF and FF conditions

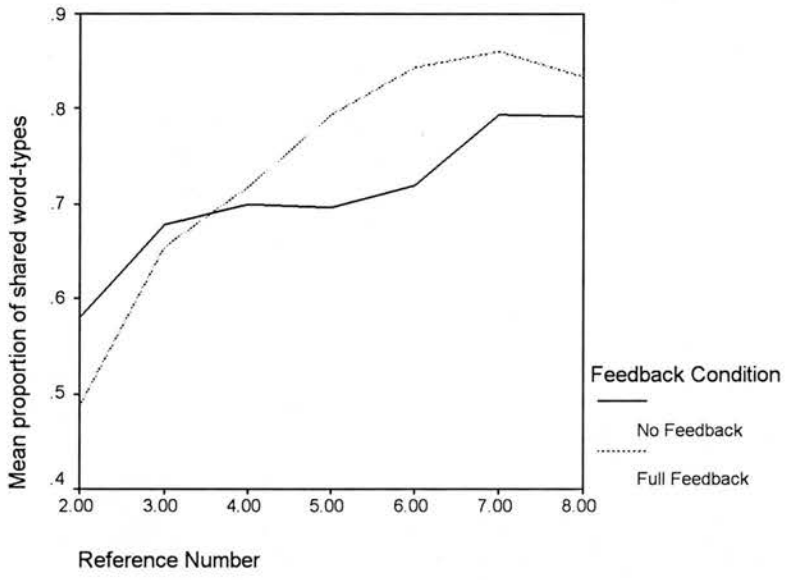
	Prop. of shared word-types	
Reference No.	NF	FF
2.00	.58	.49
3.00	.68	.66
4.00	.70	.72
5.00	.70	.79
6.00	.72	.84
7.00	.79	.86
8.00	.79	.83
Mean	.71	.74

2X7 ANOVAs demonstrated a main effect of order ($F1(6,84) = 3.22$, $p < .01$; $F2(6,66) = 43.47$, $p < .01$), where the proportion of shared word-types increased in later trials. There was an effect of Condition (with the FF condition showing more consistency than the NF condition), but this only held in the by-items analysis ($F1(1,14) = 1.48$,

[40] In a sense, of course, when shortening of descriptions occurs alongside consistency, the consistency can never be perfect or complete, as total consistency would involve the speaker simply re-using *all* of the same words from description to description, without shortening at all.

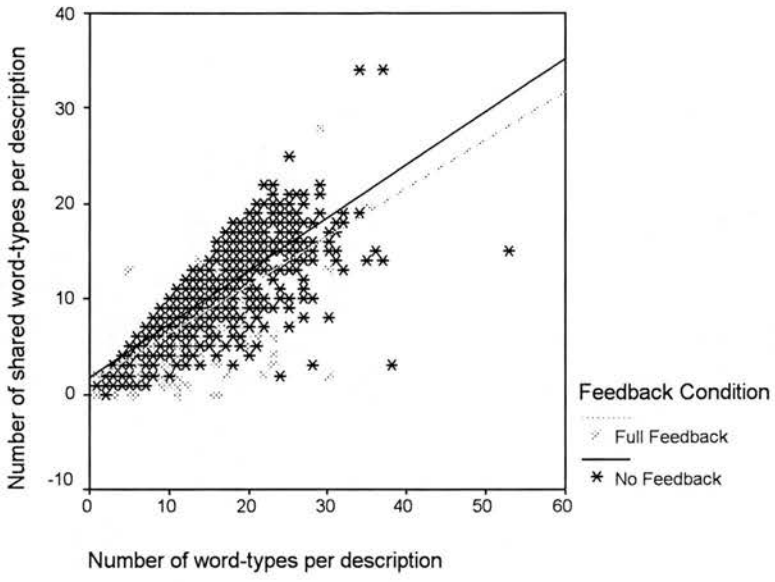
NS; F2(1,11) = 8.49, p<.05). There was an interaction (F1(1,14) = 7.19, p<.05; F2(6,66) = 6.11, p<.01), demonstrated by the graph below (Figure 14).

Figure 14: Proportion of shared word-types with repeated references



The interaction shows that the increase in consistency over repetitions was greater in the FF condition than the NF condition. However, a closer look at the data brings out another interesting point. First turns in the FF condition in these data (that is, all references, but only until the addressee spoke in each case) tended to be shorter overall than NF utterances, as mentioned earlier. Figure 15 shows that the consistency of repeated descriptions correlates with the length of those descriptions.

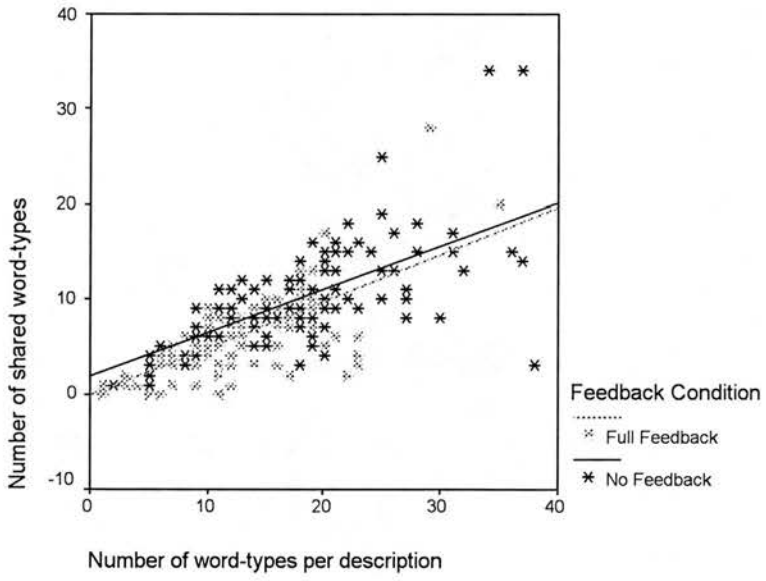
Figure 15: Proportion of shared word-types as a function of description length



It appears that the factors of length and consistency are closely correlated for each tangram ($r(24) = .982$, $p<0.01$). However it is not clear which is the driving force here; perhaps when a Speaker is using fewer words, he is more likely to use the same words again and again, because the words chosen are likely to be those that are more salient and useful to the description. Additionally, he would be likely to be able to remember his previous description more clearly if it was shorter, which would facilitate his reproduction of it. Alternatively, it might be the case that more consistent descriptions tend to be shorter, because when key words are repeated, the need for other 'padding' in the description diminishes, and so the description as a whole is less wordy.

Because NF descriptions were typically longer than FF descriptions (as detailed in section 4.4.3), this means that at any point on the graph above (i.e. for any particular length of description), the FF values will represent descriptions that were nearer to the beginning of the experiment than the NF values. Earlier descriptions, as we have seen from Figure 15, are less consistent than later ones. So, in order to factor out reference number, the relationship between total number of word-types and number of shared word-types between references 1 and 2 only is shown in Figure 16 below[41].

Figure 16: Number of shared word-types between references 1 and 2 as a function of description length



[41] Although there was a main effect of feedback condition on first utterance length, there was no interaction between the effects of feedback condition and reference number on length (Figure 5), so in theory this should not cause an interaction between the effects of reference number and feedback condition on consistency. However it seems to be worth excluding this possibility statistically here.

Both conditions still show significant correlations between number of words and number of shared word-types in each separate reference number (all ps<0.05). Comparing these independent correlations for the separate reference numbers, there were no significant differences between the feedback conditions for any of the references except reference 5 ($z=-2.06$, $p<.05$; all others $z<1.96$ ($p>0.05$)). A Bonferroni adjustment taking account of the number of analyses (7) produced a lower alpha level of .0073 ($z=2.45$), under which this result was not significant. The relationship between description length and consistency requires further investigation, as it cannot be discerned from the present results which factor is driving this effect.

4.4.5 Consistency of content words in descriptions

A second, slightly different analysis involved calculating the number of shared content (or lexical) word-types, that is, those words which are not function words. Those categories of words shown in Table 14, below, were excluded.

Table 14: Function words for exclusion

Word Class	Examples
Prepositions	of, at, in, without, between
Pronouns	he, they, anybody, it
Determiners	the, a, that, my, more, much
Conjunctions	and, that, when, while, or
Modal Verbs	can, must, will, should, need
Auxiliary Verbs	be (is, am, are), have, got, do
Particles	not, nor, as

The function words from the tangram descriptions were excised, and the analyses above were re-run on only the content word-types (again eliminating

repetitions). Some words could function as either content or function words depending on the context; the categories of these words were determined by the context.

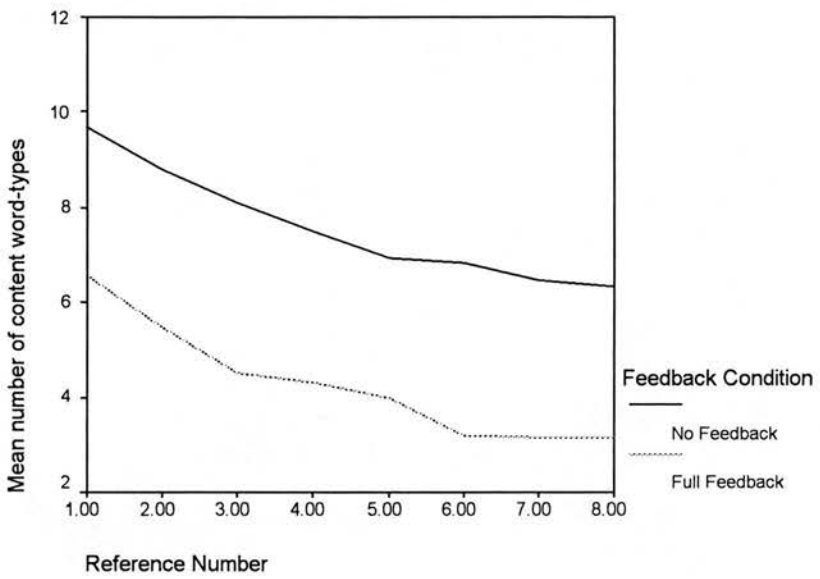
e.g. a) 'Behind the triangle is a square' ('behind' is function word)
b) 'His behind is pointy like an arrow' ('behind' is a content word)

Since function words contribute little to descriptions, the consistency between one description and the next may show up most clearly when examining only the content words – the words which are most important and relevant to the description. The majority of priming studies tend to focus on the priming of content words such as verbs and nouns (syntactic priming: Pickering and Branigan, 1998, Cleland and Pickering, 2003; lexical priming: Wheeldon and Monsell, 1992) rather than function words. Additionally, Bock (1989) found that syntactic priming does not depend on the repetition of function words; her study demonstrated that prepositional phrases using 'for' primed prepositional phrases using 'to', despite the different function words, and Fox Tree and Meijer (1999) found that participants were able to recall the content of target sentences even though the syntactic form of their recollected sentence (and therefore function words) was influenced by the syntax of an intervening prime sentence.

2 X 8 repeated-measures ANOVAs on the number of content word-types per description demonstrated main effects of Reference Number ($F1(7,98) = 26.56$, $p>.001$; $F2(6,66) = 26.49$, $p<.01$), and Condition ($F1(1,14) = 8.43$, $p=0.01$; $F2(1,11) = 12.21$, $p<.01$), where there were fewer content words in later references, and fewer in the FF condition than in the NF condition. There was only an interaction between these two effects in the by-item analysis ($F1(7,98) = .035$, NS; $F2(6,66) = 4.25$, $p<.01$).

The pattern is demonstrated in Figure 17 below, which bears a strong resemblance to Figure 14 (which involved all the word-types, both content and function). In this graph, only the first turn by the speaker was included in the Full Feedback condition (similarly to Figure 14).

Figure 17: Mean number of content word-types per description



Looking again at consistency within speakers' successive references, the mean proportions of *shared* content word-types are shown below in Table 15.

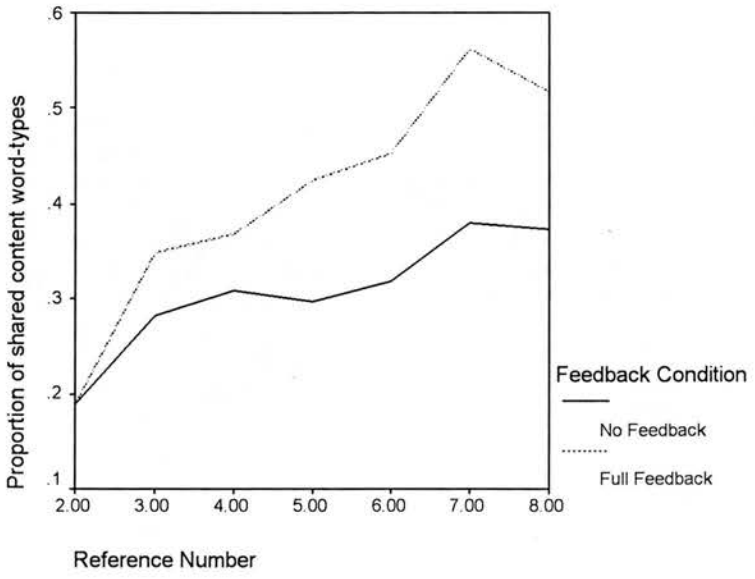
Table 15: Proportions of shared content word-types

Reference Number	No Feedback	Full Feedback
2	.53	.50
3	.69	.65
4	.69	.71
5	.74	.79
6	.76	.85
7	.77	.86
8	.75	.87
Mean	.70	.75

The proportion of content words which were shared was the same as the proportion of all words which were shared (mean shared proportion: .73 in both cases).

Repeated measures 2X7 ANOVAs on the shared content words, similarly to the previous analyses, produced a main effect of Reference Number ($F1(6,84) = 29.8$, $p<.01$; $F2(6,66) = 38.3$, $p<.01$), where there was more consistency in later references. There was an effect by Condition only in the by-items analyses ($F1(1,14) = 1.41$, NS; $F2(1,11) = 7.49$, $p<.05$), suggesting that the FF condition was only marginally more consistent than the NF condition. There was also an interaction ($F1(6,84) = 2.34$, $p<.05$; $F2(6,66) = 3.25$, $p<.01$), demonstrated by Figure 18 below. This again suggests that the FF condition increased in consistency more than the NF condition over the course of the experiment, and it bears a very strong resemblance to Figure 15, which included all words, both content and function.

Figure 18: Proportion of shared content word-types with repeated references



As before, I factored out reference number by analysing each reference number separately. Comparing correlations between the number of content word-types and the number of shared content word-types in monologue and dialogue, only one reference number showed a significant difference – reference 4 (z=4.19). This value remained significant at the 0.05 alpha level even after applying Bonferroni corrections. With the exception of this one significant result, the majority of the results suggest again that once the factors of length and reference number are taken into account, the proportion of content words that were repeated between successive descriptions did not differ significantly between the FF and NF conditions; that is, that feedback had little effect on speakers' consistency.

4.5 Discussion

So, what can these results tell us about shortening and consistency, and the effect of feedback on both these processes? The first finding was that when a tangram was described several times, the descriptions shortened with every repetition, corroborating the findings from Clark and Wilkes-Gibbs (1986) and Krauss and Weinheimer (1966), and supporting Hypothesis 1. Secondly, speakers became more consistent in their descriptions over time, tending to use more and more of the same words in their descriptions, supporting Hypothesis 2. This much seems clear, but the relationship between shortening, consistency, and feedback requires some exploration.

The length of tangram descriptions was affected by the presence or absence of feedback from an interlocutor. It seems that feedback led to more shortening of descriptions when all a speaker's turns were counted: second descriptions of tangrams in dialogue were significantly shorter than those in monologue, despite beginning at the same length. This finding equates with those of Krauss and Weinheimer (1966). However, it could be argued that looking at all of a speaker's turns in dialogue does not give an accurate indication of his intentions; after the first turn, the amount of information subsequently given by the speaker will be largely determined by the addressee, as it is she who asks for more information, and she who ultimately finishes a description, by confirming that she has chosen a tangram. So the number of total words in the speaker's description could be considered to be as much a function of the addressee's behaviour as of the speaker's own intentions. Therefore it seems that an analysis of the number of words in only the speaker's first turn may be more appropriate than an analysis of all the words (which is possibly the type of analysis Krauss and Weinheimer carried out). This new analysis on only the speakers' first turn words demonstrated that although the descriptions were shorter *overall* in the dialogue condition, descriptions shortened at the same rate

during the experiment in dialogue and monologue, which now contrasts with Krauss and Weinheimer's findings[42].

The shortening of references with repetition seen in this study cannot be a result of partner feedback because it occurred to an equal extent in both feedback conditions. Instead it may simply be a result of the speakers' practice at describing tangrams. It is probable that as descriptions were repeated, most of those elements which were not helpful to the addressee (for example, excessive repetition) were eliminated as the speakers gained confidence in their technique and in the individual descriptions. However despite the lack of difference in shortening between the two conditions, the references in dialogue *began* at a shorter length than those in monologue. As suggested in section 4.2.3, it may be the case that dialogue descriptions were shorter from the first reference because in the first trial the addressee told the speaker when she knew enough (and therefore he could stop describing). It is possible that this initial advice from the addressee set the standard length for the rest of the game, after which dialogue behaved just like monologue in terms of shortening. Another possibility is that the shorter first references in dialogue were due to the speaker's awareness that he *could* receive feedback, and expectation that the addressee would ask for more information if she needed it, rather than being a direct effect of the actual feedback received. It is impossible to distinguish between these two theories from the current analyses, and it may well be that a combination of both is responsible.

---

[42] The other possible explanation for the discrepancy between Krauss and Weinheimer's findings and the current results is the speakers' ignorance about the addressees' inability to give feedback, although this would not explain why the monologue and dialogue descriptions in their study *began* at similar lengths, when in the present study the dialogue descriptions were shorter from the start.

Another finding from this study is that consistency between utterances increased during the course of the experiment. This was to be expected, as speakers' practice at describing tangrams will have allowed them to re-use the same elements again and again, and every additional description of a given tangram should have reinforced that description further. Perhaps more interestingly, though, consistency (in terms of lexical overlap) was found to occur more in shorter noun phrases than in longer ones. It is unclear what the exact relationship is here between consistency and shortening; whether utterances become consistent because they're short, or they become short because as a result of consistency. It may be that shortening is the driving force, so that when a speaker produces a shortened reference, he decides to employ the same words as in the last description, in order to maximise his addressee's chances of understanding him. It would also be likely that the speaker would remember the words in a previous description more clearly if it had been short (in comparison with a longer description), facilitating his repetition of them and thus his consistency. Alternatively, a speaker might decide initially to repeat the words which he considers to be the most important from the last description, and this consistency would reduce the need for some of the additional words in that description (particularly those which were produced in earlier descriptions, when the speaker was still working out how to conceptualise the tangram), so the description would be shortened overall. This might lead to descriptions which were initially longer being shortened by a greater proportion than those which were shorter to begin with, because the longer descriptions may have included a larger number of unnecessary words which can be pared down with repetition.

I found that speakers' descriptions were equally consistent in the monologue and dialogue conditions. That is, they re-used a similar proportion of words in successive descriptions, no matter whether they received feedback or not, which supports Hypothesis 3, which stated that there would be no difference between the NF and FF conditions in this respect. This result does not concur with Brennan and Clark's (1996) proposal that consistency (or lexical entrainment) is a result of

collaboration between partners, which implies that it should not occur in the absence of feedback from an interlocutor. Nor does this finding agree with Hadelich et al (2004), who found that consistency in between-partner descriptions occurred more in the absence of verbal feedback than with it. It is possible that this difference in findings here may be due to the tasks employed; while the monologue condition of my study involved only the speaker speaking, Hadelich et al had both partners produce descriptions alternately. This difference in task structure also dictated the style of analysis carried out; their analyses were on between-partner alignment, whereas ours were on within-partner consistency, and this may be the crucial difference. My findings on within-partner consistency are however broadly consistent with the theory put forward by Pickering (unpublished) regarding the role of feedback in alignment between speakers. Pickering suggests that although feedback may not be required for alignment, it should facilitate its occurrence, by giving some kind of meta-commentary on the addressee's progress. However there was no notable facilitation here, and so although the findings here do not clash with this theory, neither do they support it entirely.

We cannot rule out the idea that this slight difference may be a result of a difference in context (between-partner versus within-partner) between this study and Pickering's theory. It is possible that feedback does facilitate alignment *between* speakers in some way, but not within partners. In a between-speaker analysis, the speaker's description of a particular item will be compared with the addressee's description of that item, which will then be compared with the speaker's next description of the same item and so on. That is, in the first few descriptions, the addressee's description of an item is likely to follow her feedback on the speaker's first description of that item. The process of giving feedback in response to a speaker's description (and sometimes changing the speaker's perspective in doing so) may make an addressee more likely to use that changed description in her next reference, since she has played a role in determining its content, and this may make her feel a responsibility towards her partner to re-use those descriptions when

her turn comes. This may be something that does not happen in within-partner repetitions simply because although the speaker has received feedback on his own descriptions, he may not remember this as clearly as if he had produced it himself, and so may not change his descriptions much in response to feedback. This would explain the lack of difference between monologue and dialogue here, as well as the discrepancy between my results on within-partner consistency and Pickering's predictions on between-partner consistency[43].

Our conclusion, then, is that although receiving feedback from an interlocutor allows speakers to produce shorter descriptions overall, for this task it made little or no difference to the shortening of descriptions with repetition (when we consider the first references only), nor to the increasing consistency of descriptions, once description length was taken into account.

---

[43] However this still would not explain why Hadelich et al (2004) found *less* consistency in dialogue than in monologue.

Chapter 5: Experiments 3 and 4: Investigating the benefit of dialogue for overhearers

5.1 Chapter overview

This chapter presents Experiments 3 and 4, which both involved replaying manipulated versions of the tangram descriptions from Experiment 2 to a new set of participants, and asking them to complete the same task as the original addressees. Fox Tree (1999) found that dialogue descriptions resulted in greater accuracy of tangram identification by overhearers than monologue descriptions, and, along with replicating that result, these two experiments tested two hypotheses she proposed to explain this. Experiment 3 analysed the effect of splicing out the addressee's feedback (and therefore to some extent, the addressee's perspective) on overhearers' performance. Experiment 4 kept the addressee's feedback intact, but spliced out the discourse markers in the speaker's speech. A second feature of the design was that in both experiments I varied the reference number of the descriptions; that is, the number of times the same tangram had been previously described by the speaker, in order to determine what effect this had on overhearers' performance.

5.2 Introduction

The communicative effectiveness of speakers appears to be largely influenced by the presence or absence of feedback, and may also be affected by the quality of that feedback (Bavelas et al, 2000; Kraut et al, 1982). Additionally, the practical benefit to addressees of being able to *give* feedback was verified by the findings of Schober

and Clark (1989) and Kraut et al (1982), who demonstrated that interacting in dialogues leads to better performance than simply overhearing them. Schober and Clark attributed this advantage to three characteristics implicit to dialogue: the ability to collaborate with a partner, the ability to become 'grounded' (to know that your partner understands you) and the finding of a perspective that both partners agree on.

All these processes seem to aid communication between interlocutors, but confer less of an advantage upon overhearers, presumably because they are specific to the partners involved. Nevertheless, it appears that dialogical interactivity still helps overhearers in some sense; Kraut et al (1982) demonstrated not only that overhearers can understand and retell film plots more competently when they overhear a dialogue rather than a monologue, but also that the extent of this advantage relates directly to the amount of feedback they hear. This finding led Kraut et al to propose that speakers can better design their speech for *all* addressees when they have information about *any* addressee's comprehension. However, a substantial concern with their experimental design means that this issue needs to be tested again[44]. In Kraut et al's study, participants were in separate rooms, communicating via an audio link. In the monologue condition, the connection from the addressee to the speaker was muted, so that the speaker could not hear the addressee's feedback. Crucially, though, neither of the participants knew this, and so the speaker will have expected to hear feedback from his partner. When he didn't, he may have made incorrect assumptions about her understanding. Similarly, the addressee was unaware that the speaker was not receiving her feedback, and his lack of response to this may have confused her. Additionally, although the addressees in the monologue condition of this study were instructed to

---

[44] See Chapter 2, section 2.7.2 for a description; this concern also applies to Bavelas et al (2000)

only provide single-word feedback, approximately a third of them flouted this rule and their data was still retained, which raises concerns about the validity of the monologue condition for the addressees. A similar paradigm was employed in Bavelas et al (2000), who reported that speakers told stories more competently with feedback than without. It could be argued, then, that these experiments did not use conditions equivalent to real monologues, since a lack of expectancy of feedback is surely part of the working definition of a monologue; the most common types of monologue, such as answer-machine messages and radio shows, all involve an presumption on the part of the speakers that they will not receive any feedback. Thus the results obtained using this paradigm are not necessarily applicable to those types of monologue mentioned above.

Fox Tree (1999) used a different experimental paradigm to test whether it was more useful for a person completing a tangram-selection task to overhear descriptions in monologue or dialogue. She asked participants to listen to descriptions of tangrams (hearing either a dialogue between two people, one describing tangrams to the other, or monological instructions given to an addressee; in the monologue condition both speakers and addressees were aware of the feedback restrictions) and then to select the tangram that best matched the description from a visual display. The proportion of tangrams correctly selected by the overhearers was taken to be indicative of how informative the descriptions were in the monologue and dialogue conditions.

Fox Tree pointed out that either outcome of this experiment (monologue being better than dialogue, or the reverse) could be easily explained. Monologue might be more useful because the overhearer is in a similar situation to the speaker's intended audience, in the sense that the speaker isn't expecting to receive any feedback, and neither the addressee nor the overhearer are able to give any. As a result, speakers in monologue might over-iterate points and be more descriptive

than in dialogue, knowing that their addressee cannot ask them for clarification[45]. Additionally, monologue speakers might speak in a more orderly fashion than dialogue speakers, since they have no interruptions to side-track them or alter their train of thought.

However Fox Tree also suggested that despite these possible advantages of monologue, dialogue descriptions might still be more useful than monologues for overhearers. She gave two grounds for this proposal. The first reason was that dialogues sometimes offer two perspectives on a given object, rather than the usual one perspective in monologues. The idea of 'perspective' is difficult to define, but in this context, I mean that having different perspectives on an item (for example a tangram) involves categorising the item in different ways, by using different names. This involves giving the item different semantics, rather than using rough synonyms for the same thing (for example "car" or "automobile"). Fox Tree's point was that if an overhearer does not understand the speaker's perspective (for example if the speaker describes a tangram as 'the chicken' and the overhearer cannot identify the tangram on the basis of this), then in a dialogue condition he will often have a second chance at recognising the object with the addressee's perspective. Alternatively, even if the overhearer doesn't understand either the speaker or addressee's initial perspectives, he may understand the compromise that the two interlocutors settle upon, which in some cases may differ from either of the interlocutors' initial suggestions, but incorporate elements of both.

It is likely that hearing an addressee's feedback might be beneficial for overhearers in another way: an addressee in dialogue is likely to prompt the speaker to expand on vague or unclear points, which would then aid the overhearer's comprehension.

---

[45] Unless the speaker is unaware of the restriction, as in Bavelas et al (2000) and Kraut et al (1982).

This is particularly relevant when considering the description of tangrams, as these can be interpreted figuratively (for example as a "chicken") as well as with regard to their geometric features (for example "it has a triangle with a rectangle coming out of the base") or more generally holistically ("it's a large spiky triangular shape which is wider at the top"). Figurative descriptions may be more difficult to understand than feature-based or non-figurative holistic descriptions, and so hearing an addressee's feedback might be particularly useful to the overhearer in these situations.

Despite the apparent advantages of overhearing dialogues mentioned above, there are many potential disadvantages too. One key problem is that the information given by the speaker is tailored to his specific addressee, not to the overhearer, and so once the addressee has selected a tangram, the speaker will move on to the next one, regardless of whether the overhearer has also finished the task (before they reach their 'conjecture point', in Schober and Clark's (1989) terms). In many cases this will require the overhearer to finish processing the previous utterance whilst also trying to listen to the following one. Additionally, those occasions where one partner is interrupted by the other, or where the speech of both partners overlaps, are likely to create a distraction for the overhearer and make it less clear what is being said.

Finally, Brennan and Clark (1986) propose that in dialogues, conceptual pacts are created between partners which comprise implicit agreements of how to refer to particular objects. These pacts would facilitate the partners' communication within a given discourse. Metzing and Brennan (2003) further propose that pacts are specific to the partners involved (see also Horton and Gerrig, 2005), and, if this is the case, it seems likely that this would make it even more difficult for a naïve overhearer to understand dialogues. To investigate this idea further, the present experiment will study how overhearing a speaker's description that was

produced early in the game (for example, the first time the speaker had described it) compares with overhearing a description from later in the game (after they had described the tangram seven times previously). This should allow us to determine if the number of times a tangram has been previously described (and the formation of conceptual pacts as a result of this) affects overhearers, and also if this interacts with the amount of feedback allowed in the experiment.

5.2.1 Theories accounting for the benefit of dialogue for overhearers

Fox Tree's results demonstrated that participants chose tangrams more accurately in the dialogue condition than the monologue condition, apparently corroborating the benefit of overhearing the addressee's feedback. The possible reasons for this are as follows: firstly, it might be a result of hearing two people in dialogue producing two perspectives[46] (as outlined above). This theory will be referred to as the *'Dual-Perspective Theory'*. Secondly, the benefit for overhearers might directly relate to overhearing the addressees' questions and the speakers' answers, again as described above. This theory will be referred to as the *'Feedback Clarification Theory'*. It may be difficult to separate this effect experimentally from that of the Dual-Perspective Theory, given that both effects reside in the addressee's feedback, and so if overhearing feedback is beneficial for overhearers, it may remain unclear which of these theories is responsible.

Thirdly, in this type of referential communication task, the speaker is typically the main person providing information. In the monologue condition, he will often

---

[46] Although it is certainly possible for a speaker in monologue to produce more than one perspective, this seems more likely to happen when there are two people in the discussion.

produce a lot of information within a short period of time, which may prove difficult for the addressee to comprehend. In the dialogue condition, the addressee's interjections will often enforce breaks in these descriptions (for example, when they say 'Wait, I'm just looking for it...'), which may be useful to the overhearer in terms of giving him time to catch up (thus reducing the problem of the speaker moving onto the next tangram before the overhearer has selected the current one). However, Schober and Clark (1989) compared overhearers of a dialogue in real time with overhearers who were allowed to start and stop the cassette tape of the dialogue as many times as they liked, and this self-pacing of the task did not seem to be of any benefit to the overhearers. Thus it seems unlikely that the advantage of dialogue here related to the extra time provided for the overhearer by the addressee's contributions, unless the act of starting and stopping the cassette player had any negative impact on the overhearers' performance in Schober and Clark's study, perhaps cancelling out any benefit of self-pacing. Nevertheless, I shall test this theory again in the present experiments, referring to it as the *'Additional Time Theory'*.

Fourthly, Fox Tree noted that the distribution of certain discourse markers (specifically: 'well', 'I mean', 'you know', 'like' and 'oh') was unequal between the dialogues and monologues that were produced as part of her experiment; the dialogues contained many more of these discourse markers than the monologues. She proposed that this might provide an alternative explanation for the apparent benefit of dialogue, producing what I will refer to as the *'Discourse Marker Theory'*[47].

---

[47] See the introduction to Experiment 4, Section 5.4, for a summary of the definition and functions of discourse markers in dialogue.

A fifth and final potential reason for the advantage of dialogue over monologue, and one which Fox Tree does not focus upon, is the apparently improved performance of speakers when they receive feedback. Bavelas et al (2000) showed that receiving 'specific' feedback (that is, feedback which is related to the content of what is being said, for example demonstrating surprise or shock, or laughing) from an addressee makes a speaker's language production more well-structured and expressive, more so than either 'generic' feedback (for example just saying 'uh huh') or no feedback at all. This improvement in the speaker's performance may in turn affect the overhearer's comprehension and her subsequent ability to choose the correct tangram. Further evidence for this comes from Kraut et al (1982), who found that speakers described film plots more accurately and expressively when they received feedback than when they didn't, and this then helped those who overheard the film plots to describe them to a third party more competently. Since this theory refers to the improvement in the speaker's speech as a result of receiving feedback, irrespective of the possible benefit to the overhearer of hearing the feedback, I will refer to this as the *'Speaker-Improvement Theory'*. This is distinct from the Feedback Clarification Theory in that whilst the former refers to an improvement in the actual style and structure of the speaker's speech, the latter refers to the benefit specifically accrued from hearing the addressee's questions and the speaker's replies.

The potential reasons I have discussed for the apparent benefit of overhearing dialogue over monologue are summarised below.

Dual-Perspective Theory: The overhearer benefits from hearing an addressee's feedback because this provides them with an additional perspective (Fox Tree, 1999).

Feedback Clarification Theory: The overhearer benefits from hearing an addressee's feedback because the addressee prompts the speaker to clarify ambiguous or unclear descriptions.

Additional Time Theory: Feedback is beneficial to overhearers because it allows them extra time to comprehend the speaker's speech and carry out the task.

Discourse Marker Theory: Discourse markers produced by the speaker aid the overhearer's comprehension (Fox Tree, 1999).

Speaker Improvement Theory: Feedback from an addressee makes the speaker's speech easier to understand, indirectly aiding the overhearer (cf. Bavelas et al, 2000).

I carried out two studies based on Fox Tree's paradigm, to try to discern which, if any, of these theories is correct. Firstly, I investigated whether the advantage Fox Tree found for overhearers of dialogue over monologue was replicable. Secondly, if this was successful, I wanted to narrow down where the benefit of dialogue for the overhearer lies, based on the five theories above.

One way of testing theories 1-3; that is, Dual-Perspective Theory, Feedback Clarification Theory and Additional Time Theory, would be to excise all of the addressees' speech, leaving no pauses, and assess if this puts the overhearers at any disadvantage in comparison with hearing the whole dialogue (Experiment 3). If it does, then further experiments will be necessary to discern which of these three theories is responsible for this. If this does not disadvantage overhearers, we can conclude that there is no support for these theories.

The Discourse Marker theory can be tested by simply excising the discourse markers from the speakers' speech in dialogue, and again comparing the accuracy of the overhearers' identification in comparison with the whole dialogue (Experiment 4). It seems that the Speaker Benefit theory can only be tested by default here; if overhearers do not benefit from hearing addressees' speech, then we must conclude that they are benefiting from hearing some aspect of the *speakers'* speech which differs between monologue and dialogue.

5.3 Experiment 3

5.3.1 Rationale for Experiment 3

The first experiment tested if Fox Tree's (1999) results on overhearing monologues and dialogues replicated, and also assessed the benefit of hearing addressees' speech in dialogue, by comparing the performance of participants who overheard dialogues with those who overheard the same dialogues with the addressees' speech excised (referred to here as 'half-dialogues'). I used the descriptions of tangrams recorded in Experiment 2 (Chapter 4). Although the procedure of this experiment was similar to that in Fox Tree's study, it was not identical. The main

difference was that participants in Fox Tree's experiment heard tangram descriptions in the same order as they were produced; that is, they heard one speaker describing 12 tangrams, and then a second speaker describing another 12 tangrams, with one of the sets being in monologue and the other in dialogue. In contrast, the materials for each overhearer in the present experiment consisted of 12 randomly chosen descriptions, each of a different tangram. The descriptions were presented in three different feedback conditions (run within-participants; see below for details) and the order of feedback condition was also randomised for each overhearer, so that they might have heard one tangram being described in dialogue, then one in monologue, then one in half-dialogue and so on; the descriptions were not grouped according to feedback condition. This may produce a result that differs from Fox Tree; it is possible that her paradigm might have made it easier for the overhearers to understand the descriptions because, firstly, they had a chance to get used to the speaker's (and addressee's in the dialogue condition) style of description, and secondly, they could potentially benefit from accumulated knowledge, where a speaker would refer to a previous tangram in describing a present one. There is a possibility that these advantages could affect the monologue and dialogue conditions differently[48], and that some kind of accumulated information might be responsible for Fox Tree's finding of the benefit of dialogue for overhearers. If this is the case then we would expect that the present experiment would not show the significant difference she found between monologue and dialogue. Additionally, the present task was computerised, whereas Fox Tree's was not, but I have no reason to believe that this will affect the different feedback conditions in different ways.

---

[48] For example, a speaker might be more likely to refer back to a previous description if he is sure that his addressee understood that description. This is more likely to be the case in the dialogue condition, where he received feedback from that addressee.

The descriptions from Experiment 2 were initially produced in two feedback conditions: no feedback (in which the addressee could not give any feedback, equivalent to monologue) and full feedback (in which there was full interaction, equivalent to dialogue). Additionally, for the purposes of the present experiment a third group of 'half-dialogue' sound-files was created. These consisted of the dialogue recordings with the addressee's feedback spliced out, so that the overhearer only heard the speaker's voice. This experiment thus constituted a direct test of the potential advantage of hearing the addressee's speech in a dialogue.

Although in Fox Tree's study only the first descriptions of each tangram were used, the descriptions heard by overhearers in the present experiment did not always represent the first time the speaker had described that tangram. There were three reference numbers used in the present experiment. One third of overhearers heard the first descriptions (reference 1), one third heard the second descriptions (reference 2), and one third heard the eighth (and last) descriptions of each tangram (reference 8). The first descriptions were used to replicate Fox Tree's experiment. The second references were used because these demonstrated a greater difference in number of words between the monologue and dialogue conditions than the first references did (and therefore might demonstrate a greater difference in other ways too), and additionally because these would have given the speakers in the initial experiment more time to adjust to their task than the first references. Finally, the eighth descriptions (the last in my game) were chosen because they displayed the greatest amount of shortening in comparison with the first references (according to Chapter 4); this should give us some indication of the final type of description that was settled upon by the Speakers, and how useful this is to naïve overhearers, who haven't heard the previous descriptions. It is likely that when tangrams have previously been described several times (for example by the eighth references), then overhearers should find them more difficult to identify, partly because they will not benefit from hearing the previous descriptions, and partly because later descriptions contain fewer words (Chapter 4 (all conditions): mean 29.5 words in 1st refs, 12.2

words in 8th refs, $p<.01$)). Therefore later references should be identified less accurately than earlier ones.

It could also be inferred from Brennan and Clark (1996) and Metzing and Brennan (2003) that the formation of a conceptual pact which is specific to interlocutors in dialogue should make it more difficult for dialogue overhearers in particular to understand later stages of the game. If this is the case, I would expect to find an interaction between feedback condition and reference number, showing that the accuracy of overhearers decreases more in later descriptions in the dialogue condition than the monologue condition.

Our hypotheses for this experiment are as follows:

1. Overhearers should identify tangrams more accurately following dialogue descriptions than monologue descriptions when they hear the speaker's first reference, as in Fox Tree (1999).

2. Overhearers should identify tangrams less accurately when they hear later references than earlier ones (Ref 1>Ref 2>Ref 8).

3. If conceptual pacts are created between partners in dialogue, then the factors of feedback condition and reference number should interact, such that overhearers should identify earlier references more accurately in the dialogue condition than the monologue condition, and later references more accurately in the monologue condition than the dialogue condition.

4. If the benefit of overhearing dialogue lies primarily in overhearing the speaker's speech rather than the addressee's, then overhearers should identify tangrams just as accurately in the half-dialogue condition as they do in the dialogue condition, and both more successfully than in the monologue condition.

Participants

72 University of Edinburgh undergraduate and postgraduate students participated voluntarily. None of them had participated in Experiment 2. The experiment lasted approximately 10 minutes.

Materials and design

The experiment was run on E-Prime software, which allowed for simultaneous presentation of both visual and auditory stimuli. 20 tangram pictures (Elffers, 1976) provided the visual stimuli, and the recordings from Experiment 2 of this thesis, comprising descriptions of 12 of those tangrams, were the auditory stimuli. The tangram descriptions had been initially produced in two conditions: *dialogue* (the speaker and addressee could both speak freely) and *monologue* (no response was permitted from the addressee). For the purposes of this experiment, half-dialogues were also created. These used the dialogue sound files, but with the addressees' responses spliced out, using Wavepad acoustic software, to produce sound files somewhat akin to hearing one half of a telephone conversation. The addressees' responses were not replaced with silence; rather, that part of the sound file was actually removed to prevent the introduction of lengthy pauses into the recordings. (Although it may intuitively seem that these new sound files would sound fragmented and confusing, during debriefing none of the participants reported

noticing that any of the descriptions were unnatural-sounding). Thus there were three within-subjects conditions in this experiment, corresponding to the feedback levels of monologue, half-dialogue and dialogue.

Although there were 16 participant recordings made in Experiment 2, problems with the recording apparatus meant that 4 of these (2 in monologue and 2 in dialogue) were not clear enough to use as materials for this experiment, so the remaining 12 were used. The volumes of the descriptions were normalised across all files using Wavepad software. Each of the 12 files contained 96 tangram descriptions: 12 tangrams were each described 8 times. For this experiment, the 12 first, 12 second and 12 eighth descriptions were used from each file, making 432 files to sample from. The variable of 'reference number' (1^{st}, 2^{nd} and 8^{th}) was run between-participants, with 24 participants in each reference number group.

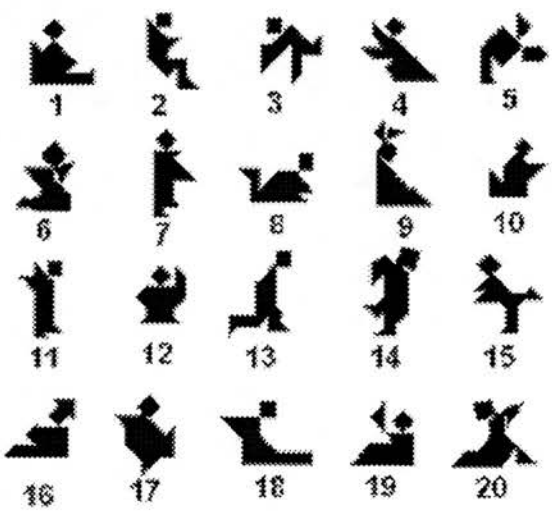
Within each reference number group, each overhearer heard a description of each tangram once, and the speakers were randomised across overhearers. No two overhearers heard the exact same combination of speakers during the experiment. For each participant, four of these tangram descriptions were heard as dialogues, four as half-dialogues and four as monologues, producing a design that was within-subjects by Feedback Condition. The files were counterbalanced so that each tangram item was represented an equal number of times in the monologue, dialogue and half-dialogue conditions. Although it was impossible to have recordings of the same speaker describing the same item in both monologue and dialogue conditions (given that Experiment 2 had a between-subjects design), all the half-dialogue descriptions were produced directly from the corresponding dialogue descriptions, and as such had the same speakers in these two conditions. Finally, the order of tangram descriptions was randomised across participants, such that no two participants heard the same order of tangrams. Overall this represented a 3

(Reference Number, between-participants) X 3 (Feedback Condition, within-participants) mixed design.

Procedure

One participant was tested at a time, and the experiments took place in an experimental cubicle with a PC. The participants were seated with their heads approximately 2 feet from the monitor. They read the instructions displayed on the screen (reproduced in Appendix D), and the Experimenter answered any questions they had before the experiment began. A sample tangram picture was shown to the participant (the same picture for every participant; this tangram did not feature in the experimental trials), in order to familiarise them with the basic structure of the shapes. The Experimenter then left the room to allow the experiment to take place. The experiment comprised 12 trials. In each trial, a layout of 20 numbered tangrams appeared on the computer screen, as in Figure 19 below:

Figure 19: Example Tangram layout (displayed on computer screen)



12 of these tangrams were experimental items (the ones used in Experiment 2) and 8 were filler items. The same selection of tangrams was seen by every participant, but the order of tangrams was changed after every trial; 12 tangram screens were created, each showing numbers 1-20 in the correct order, but the tangram above each number varied from screen to screen. Each participant saw the same order of tangram screens. The participant was instructed to look at the tangram array for a couple of minutes when it first appeared, to familiarise him with the pictures. When the participant then pressed the SPACEBAR, he heard a description of one of the experimental tangrams, played through external computer speakers. The array of pictures remained on the screen while the description was being played. The participant's task was to choose the tangram that they thought was being described, and say the number of that tangram out loud. This response was recorded by a lapel microphone, which also recorded the tangram description (that is, the speaker output), for ease of later analysis. When the participant pressed the SPACEBAR again, the same group of 20 tangrams appeared in a different order on the screen, and upon a further press of the SPACEBAR, a second auditory description, of a different tangram, was heard, inviting their response. In all, 12 descriptions were heard by each participant; one of each experimental tangram. Two dialogue sound files from the first references had to be excluded because they were too long for the E-Prime program to use, so the corresponding half-dialogue files and the two longest monologue files were also excluded to balance the design.

Upon finishing the experiment, each participant was asked if they knew what it had been about. All of them thought it was about "how good people are at following instructions that were intended for someone else" (which was the explanation given in the instructions). Upon questioning, no-one reported being aware of the different feedback conditions within the experiment. For analysis, the DAT tapes were played back to allow the experimenter to mark down the answers given for each description.

5.4 Results

Correctly identified tangrams were identified with a '1', and incorrect ones with a '0', giving a mean value of between 0 and 1 for every participant and every tangram. These are the values that will be used in the following statistical analyses. The mean scores for all conditions and all reference numbers are shown below in Table 16.

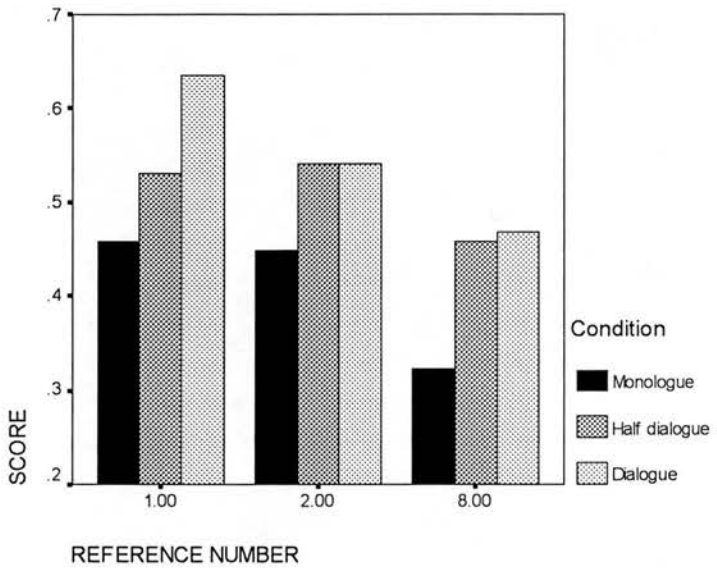
Table 16: Mean identification accuracy scores for 1st, 2nd and 8th references in monologue, half-dialogue and dialogue

Reference	Monologue	Half-Dialogue	Dialogue	Mean
1	.46	.53	.64	.54
2	.45	.54	.54	.51
8	.32	.46	.47	.42
Mean	.41	.51	.55	.49

It is clear that overall accuracy decreased systematically with later references. A 3 (feedback condition, within-participants and items) X 3 (reference number, between-participants and within-items) ANOVA demonstrated a main effect of Reference Number (marginal by items: $F1(2,69) = 4.83$; $p=.01$; $F2(2,22) = 3.22$, $p=.06$), with scores decreasing for later references. There was also a main effect of Condition ($F1(2,138) = 6.94$, $p<.01$; $F2(2,22) = 7.40$, $p<.01$). Independent t-tests showed that the locus of this effect was the difference between monologue and dialogue ($t1(142) = -3.36$, $p<.01$; $t2(70) = -.301$, $p<.01$) and between monologue and half-dialogue ($t1(142) = -2.53$, $p<.01$; $t2(70) = -1.96$, $p=.05$), where dialogue and half-dialogue both produced higher scores than monologue. There was no significant difference between dialogue and half-dialogue ($t1(142) = .965$, NS; $t2(70) = .877$, NS). There was no interaction between Condition and Reference number ($F1(4,138) = .46$, NS; $F2(4,44) = .42$, NS). Considering only the monologue and dialogue conditions, there was still no interaction ($F1(2,69) = .371$, NS; $F2(2,22) = .53$, NS).

Figure 20 below demonstrates the pattern of results.

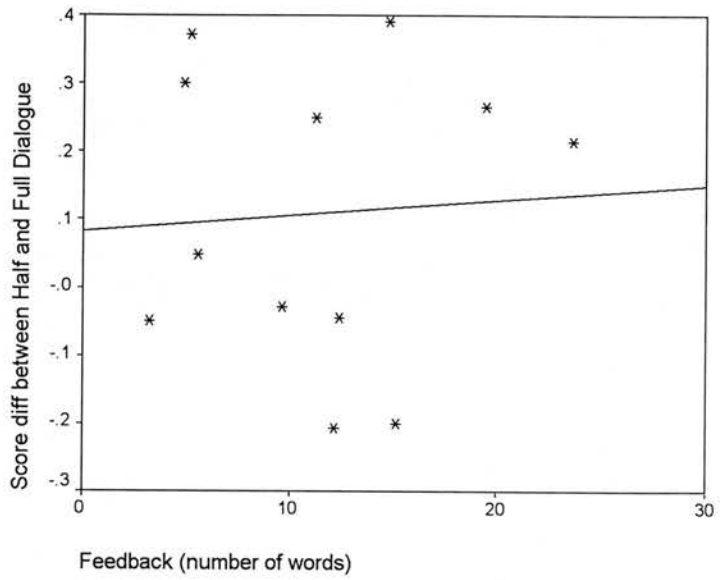
Figure 20: Mean identification accuracy scores for 1st, 2nd and 8th references in monologue, half-dialogue and dialogue



The amount of feedback in the dialogue condition, in terms of the mean total number words spoken by the listener, varied substantially in these recordings; first references tended to elicit a lot of feedback while later ones had very little. The mean number of words of feedback per participant, per tangram was 11.43 for reference 1, 3.93 for reference 2 and 1.13 for reference 8. A one-way ANOVA showed that there was a significant effect of reference number on the amount of feedback, where it reduced in later references ($F1(2,23) = 4.28$, $p<.05$; $F2(7,95) = 22.24$, $p<.01$). As a result of this, the half dialogue (in which the feedback was excised) and full dialogue conditions were more similar in the later references than in the early references, in terms of overall amount of feedback.

Maybe the 1st references, then, will give the clearest indication of the difference between conditions, because it is these references that show the greatest difference between half and whole dialogues in terms of amount of feedback. Looking at reference 1 only, a repeated-measures ANOVA demonstrated an overall effect of Condition ($F1(2,46) = 3.18$, $p=.05$; $F2(2,22) = 3.29$, $p=.05$). Further independent t-tests showed that the locus of this effect was the difference between monologue and dialogue ($t1(46) = 2.53$, $p=.01$; $t2(22) = -2.12$, $p<.05$), with the identifications being more accurate in the dialogue condition. The half-dialogue condition was not significantly different from either the monologue or the dialogue conditions (monologue: ($t1(46) = -1.03$, NS; $t2(22) = -.745$, NS); dialogue: ($t1(46) = 1.54$, NS; $t2(22) = 1.26$, NS). Figure 21 below shows the mean amount of feedback in the dialogue condition (in words) plotted against the mean difference in score between the dialogue and half-dialogue conditions for the first reference to each tangram, thus showing how the overhearer's accuracy at selecting tangrams relates to the amount of feedback. There was no significant correlation ($r(12) = .065$, NS; 1-tailed), meaning that there was no direct relationship between the amount of feedback heard by the overhearer and their accuracy of tangram identification.

Figure 21: The effect of hearing addressees' feedback on overhearers' accuracy at identifying tangrams, reference 1 only



The 8th references also require individual investigation, since any difference between the monologue and dialogue conditions here is unlikely to be due to the overhearer hearing the addressee's feedback, given the very small amount of feedback present in these references. There was a main effect of Condition ($F1(2,46) = 3.63$, $p<.05$; $F2(2,22) = 4.98$, $p<.05$), and t-tests showed a difference between monologue and dialogue ($t1(46) = -2.18$, $p<.05$; $t2(22) = -2.11$, $p<.05$) and between monologue and half dialogue (reliable by participants only: $t1(46) = -2.11$, $p<.05$; $t2(22) = -1.66$, $p=.1$). Additionally, this lack of feedback means that the half-dialogue and dialogue conditions should be very similar by this point, and, as expected, there was no significant difference between dialogue and half-dialogue ($t1(46) = -.168$, NS; $t2(22) = .449$, NS).

5.5 Discussion

The findings from the current study clearly replicate those of Fox Tree (1999), in demonstrating that it is more useful for participants who are trying to complete a task to overhear dialogues than monologues, despite my use of an experimental set-up that differed somewhat from Fox Tree's. This supports Hypothesis 1, which states that overhearers should identify tangrams more accurately following dialogue descriptions than monologue descriptions. This cannot simply be attributed to the length of the descriptions, because Chapter 4, which analysed these descriptions in more depth, showed that first references in monologue and dialogue were approximately the same length when all the words spoken by the speaker were taken into account[49] (The mean number of words in monologue was 29.35, and in dialogue, 28.87; these did not differ significantly). The main effect of reference number in these data indicates that overhearers were less accurate at identifying tangrams from later descriptions than from earlier ones, supporting Hypothesis 2. This is likely to be the case firstly, because overhearers of later descriptions will not understand references to common ground which were based upon the earlier descriptions, and secondly because later references were simply shorter than earlier references.

The combined findings on reference number and feedback condition, however, were less predictable. The fact that dialogue *still* had an advantage over monologue even after tangrams had been referred to seven times previously (indicated by the lack of interaction between feedback condition and reference number), conflicts with the viewpoint that shortening involves partner-specific pacts which may be less

---

[49] That is, not just until the listener's first interruption in the dialogue condition.

comprehensible to overhearers than to addressees (implied by Brennan and Clark, 1996[50]), and suggests instead that utterances which have been agreed upon by two or more people can still be comprehensible to overhearers despite their apparent 'short-hand' quality. It might have been expected that the opposite would be true; that dialogue speakers, making more presumptions about what their partners understand (due to the receipt of their feedback), would produce descriptions that were less transparent to outsiders, but this appears not to be the case. Additionally, dialogue references were significantly shorter than monologue references at reference 8 (see Chapter 4: mean of 6.6 dialogue words, in comparison with 17.8 monologue words), which should again make them less comprehensible simply in terms of the amount of information contained in them, but this was clearly not the case, perhaps because the additional words in monologue were somewhat unnecessary to the description. It appears that overhearing dialogues is *always* more useful than overhearing monologues in this context, despite a number of plausible reasons why the opposite should be the case. This finding contrasts with Hypothesis 3, which predicted an interaction between feedback condition and reference number.

The results from the half-dialogue condition – in which the addressees' speech was excised – are slightly more ambiguous. In my initial analyses it appeared that, since the difference between the half-dialogues and whole dialogues was non-significant overall, we should conclude that the addressees' speech provided no additional benefit to overhearers. However it is important to relate these findings to the fact that there was very little feedback from addressees in the later references, and therefore the difference between the half-dialogues and whole dialogues will have been minimal, particularly by reference eight. The bulk of the addressees' speech

---

[50] Although see Section 2.9.2 for a caveat to this finding.

took place in the first reference to each tangram, and analyses carried out only on this reference showed the half-dialogue condition to be mid-way between the monologue and dialogue conditions, and not significantly different from either. It is possible that the disjointed nature of the half-dialogues contributed to the lower scores in half-dialogues than in full dialogues in reference 1, despite the overhearers not reporting any problems in this area. It seems then that these results, although they initially seemed quite clear-cut, are somewhat inconclusive with regard to the status of the half-dialogue condition, and so Hypothesis 4, which predicted no difference between dialogue and half-dialogue, is not supported by these particular analyses.

However, the results from the other conditions may still shed light on the matter in hand; that is, the usefulness of the addressee's contributions to the dialogue. I will interpret the results in the light of the Dual-Perspective, Feedback Clarification and Additional Time Theories, considering them in order here.

Dual-Perspective Theory

Dual-perspective theory proposes that the benefit to overhearers of hearing dialogues, rather than monologues, lies in the overhearing of more than one perspective on tangrams as a result of hearing two people speaking. As mentioned earlier, in the results from the first reference, the overhearers showed significant differences in accuracy of tangram identification between monologue and dialogue. In this first reference, this advantage could potentially be due to the presence of two perspectives, as suggested by Fox Tree. However by the eighth reference in dialogue, there was typically only a single term used, as collaboration had taken place between the interlocutors and they had settled on a particular description. The mean number of content words used by the speaker in the eighth references of the dialogue condition was 3.23 words (in comparison with 7.31 in monologue),

which leaves little room for multiple perspectives to be mentioned. The example below shows how a description from the present data became more condensed during the course of interaction with an addressee:

1st Reference
S: The next one looks a bit like a duck
A: A duck? I have one that kindof looks like a chicken
S: Yeah I think that's it

2nd Reference
S: Our chicken-like figure

8th Reference
S: Chicken

The small mean number of words by the last reference in dialogue indicate that there was probably typically only one term used by this stage (representing one perspective), but despite this, dialogue descriptions were still found to be more comprehensible than monologue descriptions. This suggests that, even if two perspectives do provide an advantage to overhearers, this factor cannot in itself account for the usefulness of dialogue in my experiment. In this way my results demonstrate that the benefit of dialogue for overhearers does not rely exclusively upon the number of perspectives included, and so the Dual-Perspective Theory is not supported. It must be noted, however, that on occasions the addressee's perspective will be outlined by the speaker, as in the example dialogue below:

S: It looks a bit like an ice-skater
A: Or could you see it as a chicken?
S: Yes, you're right, it could look like a chicken.

In this example, even if the addressee's speech was excised, her perspective would still be evident from the speaker's speech. If this were frequently the case, then it would mean that the half-dialogue descriptions still included the multiple perspectives, which would mean that this experiment did not manipulate what it reported to. However in reality, the above situation happened rarely in the transcripts, so I would not expect it to have influenced the results as a whole.

Feedback Clarification Theory and Additional Time Theory

The Feedback Clarification Theory proposed that hearing feedback is useful to overhearers because addressees can ask speakers to expand on ambiguous descriptions, to the benefit of the overhearer. Crucially, though, I found that although the amount of feedback diminished significantly in later trials in the dialogue condition, there was no interaction between repetition number and feedback condition on overhearers' task success. If this theory was correct, we would have expected to see the advantage of dialogue for overhearers diminishing in later references, as it became more like the monologue condition (additionally, we would have expected to see a significant difference between the half-dialogue and full dialogue conditions overall, which was not evident). Since the benefit of dialogue remained constant until the final references, it appears that it was not a direct result of the feedback. This also applies to the Additional Time Theory; since the amount of feedback diminished in later trials, so did the amount of additional time that the overhearers had as a result of the feedback, and so there should have been an interaction if time was the crucial element here. As there was no interaction, we can conclude that there is little support for either of these theories.

So perhaps we should conclude in favour of the Speaker Improvement Theory, which suggests that receiving feedback makes speakers easier to understand[51]. But why should there still be a benefit of dialogue in the eighth references, which contained almost no feedback from the listener, and very few words even from the speaker? It seems unlikely that style of speech or fluency would play an important role in the comprehensibility of such short descriptions, or that a very small amount of feedback would affect them greatly. It is more plausible that in the dialogue condition, the feedback on earlier trials still had an effect on later descriptions; that the advantage of early feedback (for example, in terms of selecting a perspective on the tangram that was useful for the addressee) continued until the end of the game. Maybe we should consider perspectives to be the main factor here: any description that a dialogue speaker settles upon will have to be comprehensible not only to the speaker himself but also to his partner, which might make it more likely to be comprehensible to others too. This concurs with Kraut et al's idea that speakers can better design their speech for all addressees when they have information about any addressee's comprehension; speakers may be more likely to end up with a widely-recognisable perspective on the given tangrams when their descriptions have been evaluated (and often modified) by an addressee. This added comprehensibility of descriptions in dialogue may be the key to these findings, and at this point, it certainly seems like the most likely explanation for them.

---

[51] Bavelas et al (2000)) reported that their speakers were less disjointed and less repetitive with feedback.

5.6 Experiment 4

Experiment 3 tested Fox Tree's hypothesis that the benefit of dialogue for overhearers lies in the additional perspectives contributed by the addressee. The present experiment will test her second hypothesis; that the benefit of dialogue was a result of the additional discourse markers she found in the dialogue condition. Experiment 4 will involve excising certain discourse markers from speakers' dialogue tangram descriptions and then playing the remains of the dialogues to overhearers, in an attempt to assess if the presence of discourse markers in speakers' speech confers an advantage to overhearers.

5.6.1 Uses of discourse markers

Discourse markers perform many functions in everyday spontaneous speech. Fox Tree and Schrock (1999) provide a comprehensive review of discourse marker uses reported in the literature; among them, to signal upcoming repairs or topic shifts, to add emphasis, to form idioms, to draw attention, to elicit information from an interlocutor, and to retain the floor in discussions. Because discourse markers are used in so many different ways, it is difficult to define them precisely, however Fraser (1999), who devoted a paper to this very topic, reported, "I have defined discourse markers as a pragmatic class, lexical expressions drawn from the syntactic classes of conjunctions, adverbials, and prepositional phrases. With certain exceptions, they signal a relationship between the segment they introduce, S2, and the prior segment, S1. They have a core meaning which is procedural, not conceptual, and their more specific interpretation is 'negotiated' by the context, both linguistic and conceptual" (p950). The most important element of this definition for the purposes of the current discussion is that discourse markers are seen as having a *procedural* meaning, not a *conceptual* one; that is, they are used to determine the form in which the information is conveyed; in Fraser's words, "...an expression with a

procedural meaning specifies how the segment it introduces is to be interpreted relative to the prior [segment]" (p944). Fuller (2003) provided three requirements for discourse marker status: they are "used to signal relationships between discourse units", they are "optional", and they "do not change the truth conditions of the propositions in the utterances they frame" (all p11). Given their filler nature, discourse markers may also be considered to be grammatical disfluencies, in contrast with non-grammatical disfluencies such as 'um', 'er' etc[52].

In common with Fraser, Fox Tree (1999) suggested that discourse markers allow the addressee (and overhearer) to structure the content of the dialogue and separate the ideas contained within it more effectively, and thus are beneficial to listeners of both types. It seems equally likely though, that these extra markers could potentially serve to confuse the addressee (for example, if the word 'like' is used repeatedly during a statement), lessening whatever the *real* advantage of dialogue might be, or else that they might have no effect at all. The apparent relationship between score and discourse markers reported by Fox Tree does not necessarily mean that either is dependent upon the other; it may be that both are caused by an unknown third factor. The simple correlational relationship reported in her paper makes it difficult to discern what exact role, if any, discourse markers play in the apparent benefit of dialogue.

One piece of evidence that supports the practical benefit of discourse markers for listeners comes from Fox Tree and Schrock (1999), who focused on the use of the discourse marker 'Oh'. They found that listeners recognised words in speech more quickly (in particular, the words they had been instructed to listen out for)

---

[52] By 'grammatical' I mean here items which have a specific lexical meaning, although Clark and Fox Tree (2002) would also attribute specific meanings to some non-grammatical items, such as 'um' and 'uh'; this however is a matter of some debate.

following an 'oh' than if the 'oh' was replaced with a pause or simply excised from the heard speech. Fox Tree and Schrock attributed this advantage of 'oh' to its role in helping addressees to integrate discourse; that is, to add the upcoming information to what is already known (similarly to Fraser's reported role of discourse markers on integrating S2 and S1, above). Two ways were proposed in which this might take place: firstly, 'oh' may inform addressees to stop their current integration in preparation for processing upcoming information, or secondly, it might prime speakers to expect a change of topic in the following speech, which in turn will aid their comprehension of this new topic.

Bangerter and Clark (2003) propose that one of the roles of discourse markers (which they call 'project markers') is to enter and exit joint projects, which is a similar idea to that of changing topics mentioned above. Bangerter and Clark point out that a joint project, like a solo activity by an individual, can be viewed as a series of projects and sub-projects put together to attain a given goal, for example cooking a meal together or planning a holiday. But the crucial element of *joint* projects is that they involve communication and coordination between partners at each stage of this process. Partners use dialogue to navigate joint projects, and they may use discourse markers to signal 'vertical transitions' in the dialogue, for example to mark a complete change of topic, a request for more information, or a brief digression and then return to the main topic. In contrast with vertical transitions, 'horizontal transitions' involve continuing on with the same speech topic, and are often accompanied by continuers rather than discourse markers (for example 'uh-huh' or 'yeah'). Bangerter and Clark point out that 'On one plane, people create dialogue *in service of* the basic joint activities they are engaged in, making dinner, dealing with the emergency, operating the ship [horizontal transitions]. On a second plane, they manage *the dialogue itself* – deciding who speaks when, establishing that an utterance has been understood etc [vertical transitions]' (p196, emphases added). If discourse markers are helpful to addressees, then, it may be

because they mark these vertical transitions and allow the addressee to follow the speaker's train of thought more easily.

Despite my having considered all discourse markers in a summary fashion above, they do not all necessarily have the same meanings, and neither are they used consistently in the same contexts. The markers that Fox Tree chose to analyse, and that I subsequently chose to excise for this experiment ('well', 'I mean', 'you know', 'like' and 'oh') have been studied for their literal meanings and uses. Fuller (2003) notes that Jucker and Smith (1998) considered the markers 'well', 'I mean' and 'you know' to be examples of *presentation markers*, with which a speaker advises his addressee how the following words should be processed. In contrast, 'like' and 'oh' were described as *reception markers*, which an addressee uses to indicate to the speaker how well she is integrating the given information into her current knowledge state. However Fuller proposed that, contrary to Jucker and Smith's proposal, the function and distribution of discourse markers is more heavily determined by the role of the speaker and the relationship between the interlocutors than by characteristics of the individual discourse markers. She commented that, "[Discourse markers] can be viewed as particles which are employed in a variety of speech genres, but show distinctions in frequency and function that correlate with the speaker's role as either an interviewee or an interlocutor in an informal, symmetrical interaction".

5.6.2 Rationale for Experiment 4

In the present experiment, I excised the same five discourse markers as Fox Tree (1999) ('well', 'I mean', 'you know', 'like' and 'oh') from dialogues in preparation for playing them to overhearers alongside the full dialogue and monologue sound files. Fox Tree's reasoning for choosing these particular markers was that in most of their occurrences it was easy to separate the discourse marker uses of these from the

semantic uses; with many other lexical items which double up as discourse markers it can often be difficult to tell in what sense they're being used. For example in the sentence below, the first 'like' is an obvious discourse marker (possibly signalling the inclusion of more information), the second 'like' is an adverb, and actually forms part of the description.

e.g. "Their feet go off to the left, *like* along the floor, and one of the feet are pointing down. Kinda looks *like* they've got three feet maybe".

Although the current transcripts contained many more discourse markers than these five, I excised *only* the same discourse markers as Fox Tree did in order to see if her findings on these particular markers replicated. The decisions about which occurrences of these words to excise (i.e. how to distinguish between discourse marker use and non-discourse marker use) were based on Fox Tree's criterion for exclusion, which was, 'Nonpragmatic [non-discourse marker] uses were defined as those which could carry normal nominal, verbal or adverbial functions, as in "do you know which one I mean?" and "it looks like a seal". Pragmatic [discourse marker] uses did not fit these criteria, as in "the other foot is is supporting him and like he's leaning over against the wall" and "it's just just you know pushed towards the left" ' (p48).

This experiment therefore involved a direct test of the effect of discourse markers upon overhearers' comprehension, in addition to investigating if the effects of feedback condition and reference number found in Experiment 3 were replicable.

The hypotheses for this experiment are as follows:

1. Overhearers should identify tangrams more accurately following dialogue descriptions than monologue descriptions when they hear the speaker's first reference, replicating Experiment 3.

2. Overhearers should identify tangrams less accurately when they hear later references (Ref 1>Ref 2>Ref 8), again replicating Experiment 3.

3. The factors of feedback condition and reference number should not interact, replicating Experiment 3.

4. Overhearers should identify tangrams just as accurately in the dialogue without discourse markers condition as they do in the dialogue condition, and both more successfully than in the monologue condition.

Participants

72 University of Edinburgh undergraduate and postgraduate students participated voluntarily. None of them had participated in Experiment 2 or 3. The experiment lasted approximately 10 minutes.

Materials and Design

The materials and design were exactly the same as in Experiment 3, except that instead of using the half-dialogues as the third feedback condition, I used the

whole dialogues with the 5 discourse markers chosen by Fox Tree ('well', 'I mean', 'you know', 'like' and 'oh') excised from only the speakers' speech (not the addressees'). They were not replaced with silence; rather, the parts of the files containing these words were actually excised, on the assumption that this would keep them as natural-sounding as possible. Debriefing demonstrated that none of the participants were aware of any artificiality in the descriptions resulting from this. The 1st, 2nd and 8th references were used again, in a between-participants design, with 24 participants in each reference number group. The 8th references did not contain any discourse markers at all, but they were still included for the sake of completeness. Again, it was a 3 (Reference Number, between participants) x 3 (Feedback Condition, within participants) mixed design. The mean total frequencies of the discourse markers which were excised are shown in Table 17 below for all reference numbers.

Table 17: Mean frequencies of five selected discourse markers per reference, in monologue and dialogue

Reference	Monologue	Dialogue	Mean
1	.033	.771	.804
2	.021	.354	.188
8	.000	.000	.000
Mean	.018	.375	.331

The means above demonstrate that there were more markers in the dialogue condition than in the monologue condition. However, given that the tangram references were significantly shorter in the monologue condition than in the dialogue condition (as described in Chapter 4) and shortened further during the course of the experiment, it makes more sense to calculate the number of discourse markers per 100 words of description (including the discourse markers).

These new values are shown below in Table 18, after outliers more than 2.5 Standard Deviations greater than the mean were truncated to the 2.5SD level (calculated separately for each reference number and condition).

Table 18: Mean number of discourse markers per 100 words in monologue and dialogue

Reference	Monologue	Dialogue	Mean
1	.068	3.51	1.79
2	.062	3.26	1.11
8	.000	.000	.000
Mean	.043	2.26	.967

A repeated-measures ANOVA (analysed by speaker, with feedback Condition between-participants and within-items, and Reference Number within-participants and within-items) demonstrated that there was a significant effect of Condition ($F1(1,10) = 17.03$, $p<.01$; $F2(1,5) = 8.39$, $p<.05$), where dialogue contained more discourse markers than monologue. There was also an effect of Reference Number ($F1(2,20) = 5.37$, $p=.01$; $F2(2,10) = 6.60$, $p<.05$), where earlier references contained more discourse markers. There was also an interaction ($F1(2,20) = 4.97$, $p<.05$; $F2(2,10) = 6.24$, $p<.05$), where although the number of discourse markers reduced to zero with later references in both conditions, the dialogue condition began with more, and so the reduction there was greater. Independent t-tests demonstrated that the difference in number of discourse markers between monologue and dialogue was significant individually for references 1 ($t1(10) = -2.68$, $p<.05$; $t2(10) = -2.64$, $p<.05$) and 2 ($t1(10) = -4.06$, $p<.01$; $t2(10) = -2.86$, $p<.05$). There were no discourse markers in the 8th references.

The lack of discourse markers in the 8th references to tangrams could be the result of a number of factors. It has been proposed that discourse markers indicate hesitancy or uncertainty (cf. Schiffrin, 1987), and by the 8th reference, the speakers

are so sure of their description, having described the item seven times before, that the markers are no longer required. In the same vein, since Chapter 4 demonstrated how tangram descriptions shortened with repetition, it is plausible that the reduction in the number of discourse markers is simply a by-product of this reduction in length. Alternatively, as Fox Tree suggests, they might help interlocutors to manage conversations, and although the speakers are still in some sense taking part in conversations by the eighth reference (in that they are participating in a dialogue, where their partners are free to respond), the response from addressees by this point was minimal; Chapter 4 details how the mean number of turns taken per pair per tangram diminished over the course of the experiment, to 2.07 turns by the eighth reference. This number represents little more than a speaker's description of a tangram and his addressee's acceptance. It is possible, then, that the number of discourse markers used in speech reflects the amount of interaction with an interlocutor, and that this amount of interaction might be a better predictor of overhearers' task success than the actual number of discourse markers present. That is, interaction with an interlocutor may cause, firstly, an increase in overhearers' task success (as is indicated by the benefit of dialogue over monologue for overhearers in Experiment 3), and secondly, an increase in the use of discourse markers by the speaker. These two factors are not necessarily directly related, and the present study will test if this is the case.

Procedure

The experimental procedure was identical to that of Experiment 3, except that the half-dialogue sound files were replaced by the dialogues without discourse markers.

5.7 Results

The scores were calculated in the same manner as in Experiment 3, and these values will be used in the following statistical analyses. The mean tangram identification scores for all three conditions and all three reference numbers are shown below in Table 19.

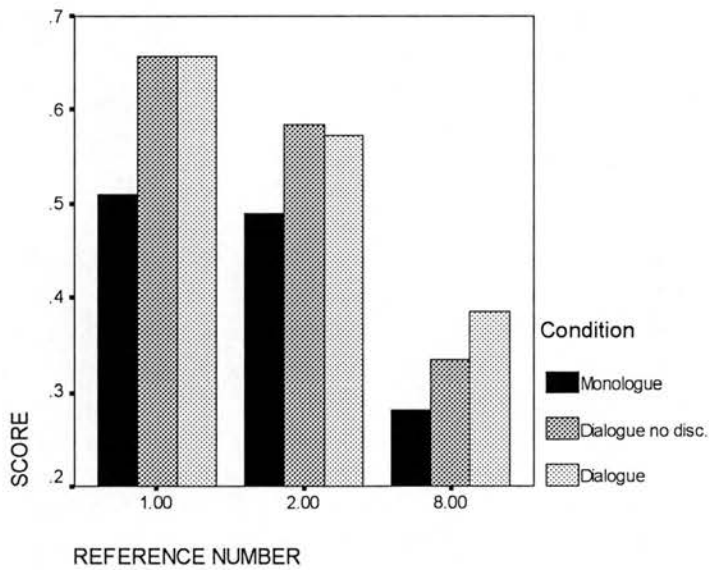
Table 19: Mean identification accuracy scores for 1st, 2nd and 8th references in all conditions

Reference	Monologue	Dialogue	Dialogue no DMs	Mean
1	.51	.66	.66	.61
2	.49	.57	.58	.55
8	.28	.39	.33	.33
Mean	.43	.54	.52	.50

3X3 repeated measures ANOVAs (condition within-participants and items, reference number between-participants and within-items) demonstrated an effect of feedback Condition by participants ($F1(2,138) = 4.09$, $p<.05$; $F2(2,22) = 2.12$, $p=.1$). Independent t-tests showed that the monologue condition was overall significantly less accurate than both the dialogue (marginal by items: $t1(142) = -2.29$, $p<.01$; $t2(70) = -1.86$, $p=.07$) and the dialogue without discourse markers (by participants only: $t1(142) = -2.06$, $p<.05$; $t2(70) = -1.53$, $p=.1$), but that there was no significant difference between the dialogue and dialogue without discourse markers conditions ($t1(142) = .282$, NS; $t2(70) = .224$, NS). There was also a main effect of Reference Number, with identifications being more accurate in the earlier references ($F1(2,69) = 19.51$, $p<.01$; $F2(2,22) = 25.60$, $p<.01$). There was no interaction ($F1(4,138) = .269$, NS; $F2(4,44) = .341$, NS).

3X2 ANOVAs on just references 1 and 2 (because the 8th reference contained no discourse markers) showed an effect of condition by participants only ($F1(2,92) = 3.16$, $p<.05$; $F2(2,22) = 1.90$, $p= .17$) and no effect of reference number ($F1(1,46) = 2.05$, NS; $F2(1,11) = 1.78$, NS). Again, there was no interaction ($F1(2,92) = .193$, NS; $F2(2,22) = .382$, NS). The pattern of results is demonstrated in Figure 22 below.

Figure 22: Mean identification accuracy scores for 1st, 2nd and 8th references in monologue, dialogue without discourse markers and full dialogue

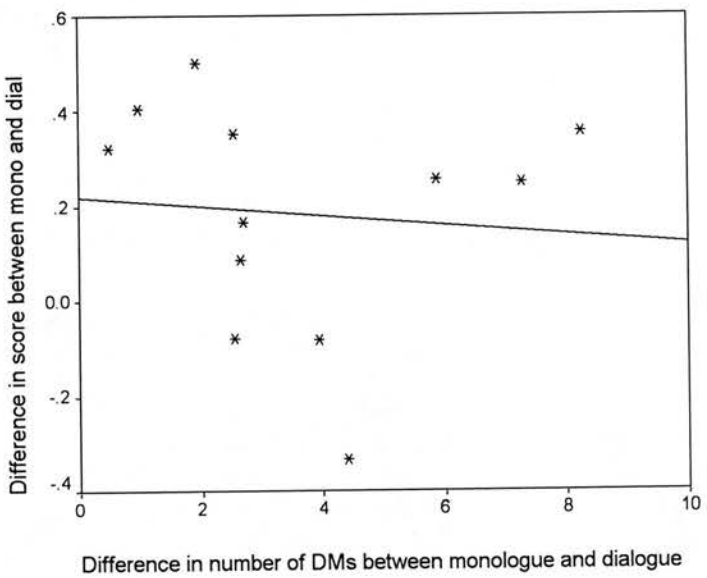


Fox Tree found that the difference in overhearer scores between monologue and dialogue was highly correlated with the difference in number of discourse markers between monologue and dialogue. I carried out the same analysis with mean values for each tangram, and using the number of discourse markers per 100 words rather than the raw scores.

Once outliers exceeding 2.5 Standard Deviations from the mean had been truncated to the 2.5SD level, there was no significant correlation between these variables in

the 1st references of my data (which are the ones Fox Tree used; (r(12) = -.1, NS)). The data set is plotted in Figure 23 below.

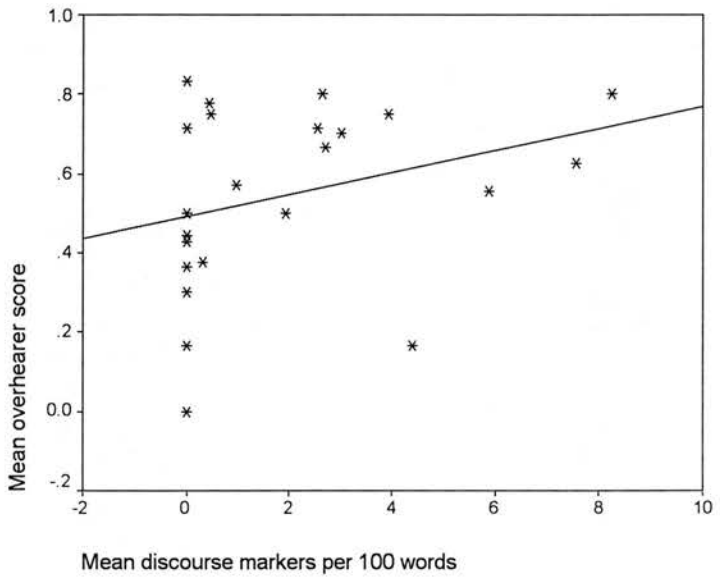
Figure 23: The difference in mean score between monologue and dialogue plotted against the difference in mean number of discourse markers between monologue and dialogue, first references only



I also carried out a correlation between overhearer score and the mean number of discourse markers per 100 words in both monologue and dialogue conditions, and found no correlation between these factors (r(22) = .135, NS).

The scatter plot is shown below in Figure 24.

Figure 24: The relationship between the mean number of discourse markers per 100 description words and overhearer score



5.8 Discussion

The results again replicate the findings from Experiment 3 that dialogue is more useful for overhearers than monologue, and that earlier descriptions are more useful than later descriptions, but that these two factors do not interact. These findings support Hypotheses 1-3. The results also demonstrate that the discourse markers selected by Fox Tree (1999) and excised from the speakers' speech here had no significant effect on overhearers' identification accuracy, as was predicted by Hypothesis 4, which stated that there would be no difference between the overhearers' performance in the dialogue and dialogue without discourse markers conditions, but that both would give better results than the monologue condition. The dialogue without discourse markers condition produced very similar results to the whole dialogue condition.

Thus the Discourse Marker Theory – that the benefit of overhearing dialogues over monologues is a result of the additional discourse markers in dialogue - is not supported.

It should be noted, though, that Fox Tree's choice of just a few markers to excise may have diluted any apparent reduction of their benefit. Presumably deleting *all* the discourse markers in the transcripts would show the greatest effect, although with many other terms that double as discourse markers, it would have been more difficult to distinguish their discourse-marker uses from their non-discourse-marker uses (which is why Fox Tree chose the particular markers she did to excise). Nevertheless, if the five markers I did excise were used to structure the content of dialogues at all, then deleting them should have thrown the overhearers into confusion, or at least have had an effect on their task completion; without them, performance should have been poorer than in the comparable full dialogue descriptions.

Perhaps most crucially, my data did not show the same pattern of discourse markers as Fox Tree's; although there were more discourse markers in dialogue than in monologue, and scores in dialogue were overall higher than those in monologue, there was no significant correlation between the difference in number of discourse markers between monologue and dialogue and the difference in score for monologue and dialogue. Nor did I find a significant correlation between overall number of discourse markers and overall score for the overhearers.

5.9 General discussion

Experiments 3 and 4 both support Fox Tree's (1999) finding that overhearing a dialogue is more useful than overhearing a monologue, at least in terms of task completion. However they did not support either of her hypotheses: that the benefit of dialogue might be due to the additional perspectives or the extra discourse markers in dialogue. Although the findings from Experiment 3 on dialogues with the feedback excised were somewhat ambiguous, we can still draw conclusions from the findings on the dialogue and monologue conditions. There was no interaction between feedback condition and reference number, demonstrating that the performance of overhearers, which was initially better in dialogue than in monologue, did not diminish more over repetitions in the dialogue condition than in the monologue condition, despite a significant decrease in feedback during the course of the experiment. This suggests that the benefit of dialogue for any description was not a direct result of the feedback on that particular description, and so along with not supporting the Dual-Perspective Theory, neither can these findings be accounted for by any of the other three theories I proposed, Feedback Clarification, Additional Time Theory or Speaker Improvement Theory, which all rely upon the presence of feedback for their explanations.

The finding that the amount of feedback produced by the addressees did not correlate with the overhearers' scores again suggests that there was little or no benefit from overhearing feedback on individual descriptions. Additionally, it seems that the benefit of dialogue for an overhearer was not dependent upon how much information was provided by the speaker, as in the first references, the number of words spoken by the speaker was shown to be the same in monologue and dialogue.

Following on from this, if these findings are not a result of feedback provided by the addressee, or of the *amount* of information produced by the speaker, then by default, they must be reliant upon some *quality* of the speaker's speech.

Experiment 4 demonstrated that the presence of discourse markers does not provide an advantage to overhearers of dialogues, and so this cannot explain the useful quality of dialogical speech. This did not support the Discourse Marker Theory. However, this does not mean that discourse markers do not benefit comprehension at all; it is of course possible that they might affect addressees in a different manner from overhearers, or people in a completely natural dialogue differently from those participating in a referential communication task. Additionally, these results do not show that discourse markers have no effect on comprehension per se, and results from Fox Tree and Schrock (1999) on the discourse marker 'oh' seem to indicate to the contrary. These findings simply suggest that, whether they benefit comprehension or not, discourse markers do not appear to have an impact on overhearers' performance at this type of task.

A second key finding that arises from both the experiments reported in this chapter is that, as predicted, repeated references were more difficult for overhearers to understand than initial ones. This can presumably again be attributed to the fact that the most detailed information about tangrams will be provided on the first reference to them, with later descriptions relying on shortened terms that refer back to the initial descriptions. Brennan and Clark (1996) suggest that in a dialogue situation, conceptual pacts are formed, where the interlocutors reach an agreement regarding how each tangram should be referred to. This should prove a disadvantage to people who only overhear later dialogue descriptions, as they have not heard the earlier, more explicit descriptions that were involved as the partners created their pact. Those who overhear later *monologues* should be at less of a disadvantage, as the speaker will not have had any indication of how much his

addressee understands of his descriptions, and so should continue to use fairly detailed descriptions throughout. Surprisingly, though, my results demonstrated that the disadvantage of overhearing later tangram descriptions did not differ between the monologue and dialogue conditions. This suggests that either conceptual pacts were not created in the dialogues (contrary to Brennan and Clark), or, alternatively, that they were created but did not put the overhearers at a disadvantage. It is possible that the disadvantage of not hearing the formation of a conceptual pact could be compensated for by the added comprehensibility of a description that arises from it being a result of both the speaker and the addressee's agreement, as suggested above.

So if the benefit of feedback does not relate to what the addressee says, or how much the speaker says, where does it lie? Maybe the most plausible interpretation for the current results relates to Kraut et al's (1982) idea that speakers can better design their speech for all addressees when they have information about any addressee's comprehension; that it is the specific word choice of the speaker which determines the success of a description. Maybe it is only through receiving feedback that a speaker learns what a listener (whether addressee or overhearer) needs to know. That is, it may not be important to the overhearer what the addressee says, but the addressee's effect on the speaker's speech may be crucial, as together they weed out inadequate descriptions and replace them with more accurate, comprehensible ones. Since the bulk of the feedback occurred in the first description, it is probable that the speaker's adjustment in response to feedback will have happened most at this point. The beneficial effect of this feedback will then have filtered down to future descriptions. This would explain the benefit of dialogue for overhearers even in later descriptions, despite these descriptions containing very little feedback in themselves.

Chapter 6: Experiments 5 and 6: Feedback and tangram discriminability

6.1 Chapter overview

Experiment 1 (Chapter 3) demonstrated that it is more useful for people completing referential communication tasks to interact in dialogues than to hear monologues, at least in terms of success at the task in hand. Additionally, Experiments 3 and 4 (Chapter 5) showed that even if we are not interacting ourselves, it is still beneficial to overhear others' interaction; overhearers identified tangrams more accurately when they overheard two people in dialogue rather than one person giving a monologue. This chapter presents two experiments that studied in more detail the benefit of feedback for both addresses and overhearers. In particular, I looked at how the *difficulty* of a task influences the benefit of feedback for both these groups. Experiment 5 was a referential-communication tangram identification task, somewhat similar to that described in Chapter 4. It was run in three feedback conditions and at two levels of task difficulty. The addressees' abilities to correctly identify the tangrams described by their partner were measured. Experiment 6 used the recordings of tangram descriptions from Experiment 5 to again analyse the effects of feedback condition and task difficulty, this time on accuracy of identification by overhearers.

6.2 Introduction

The ability to interact with a partner has a highly beneficial effect on performance: taking part in a dialogue, rather than a monologue, improves addressees' performance in narrative, object selection and building tasks (Kraut et al, 1982;

Clark and Krych, 2004[53]; Experiment 1, this thesis). Additionally, people participating in dialogues perform better than overhearers of the same dialogue (Schober and Clark, 1996). Looking more closely at overhearers, Fox Tree (1999) demonstrated that it is more useful for people to overhear dialogues than monologues, and this finding was corroborated by results from Experiments 3 and 4 of this thesis. These results all suggest that whether people are giving feedback, or just overhearing it, their task performance benefits as a result. But does the benefit of feedback for addressees or overhearers vary with the requirements of the task concerned? It seems plausible that in more difficult tasks, an addressee or overhearer might benefit more from being able to give or overhear feedback than in easier tasks. Fussell and Krauss (1992) propose that the receipt of feedback is one of two factors that allow speakers to implement audience design (the other being their prior knowledge of the addressee), and it seems from studies like Krauss and Weinheimer (1966) that it allows speakers to shorten their descriptions to a greater extent than would occur otherwise (although Experiment 2 from this thesis showed contrasting results). Nevertheless, if feedback allows speakers to tailor their utterances to fit the need of their addressees, then in a situation where those needs are particularly demanding (if the task is more difficult than usual, for example), then it may be of even greater value.

There are many potential ways of increasing task difficulty in language production experiments. Some researchers have had participants carry out two tasks concurrently (Simon and Sussman, 1987), concurrently remember a digit load (Waters, Caplan, & Rochon,1995), ignore distractors (Kingma, La Heij, Fasotti & Eling, 1996) or take another person's perspective when giving directions (Chiu, Hong, & Krauss, unpublished). The task difficulty in the present experiments was

---

[53] Although we have concerns with the experimental designs of this study: Clark and Krych had speakers produce monologues into tape recorders: see Chapter 3, Section 3.2.2 for details.

manipulated by varying the 'discriminability' of the tangrams; that is, the ease by which they can be distinguished from other tangrams in the selection. In a situation where there are two tangrams which roughly match the speaker's description (and therefore are less discriminable), the addressee will have to think harder about which one to choose, and hence the task difficulty will be increased.

Hupet, Seron and Chantraine (1991) found that when speakers are describing tangrams which are less discriminable, more collaborative effort is required to refer to them, in terms of number of turns, and number of words spoken by both partners, than for more discriminable tangrams. In this kind of scenario, the presence of feedback should be of particular benefit, because it allows the addressee to ask specific questions which will allow her to distinguish between the two similar tangrams, and also allows common ground to build up between the partners as the descriptions proceed. This may be less important for the more easily discriminable tangrams because even in the no feedback condition, the tangrams may be so easy to tell apart that the speaker may initially give enough information about them for the addressee to make an accurate selection, without the speaker needing to be aware of details of the addressee's perspective.

It is likely, then, that imposing a restriction on the amount of interaction allowed, by not allowing free dialogue, will have more of a negative effect for less discriminable tangrams, causing an interaction between discriminability (or similarity, in my terms) and the amount of feedback permitted. Of course if an interaction between difficulty and feedback does occur, it might simply be because a listener in a more challenging situation is likely to ask more questions and produce more feedback than one in an easier situation, which may benefit both himself and any overhearer to a greater extent.

However Experiment 3 demonstrated that the amount of feedback heard by overhearers (in terms of the numbers of words) did not correlate with their accuracy of tangram identification[54].

The experiments presented in this chapter are as follows: Experiment 5 looked at how the difficulty of tangram selection affects the benefit for addressees of being able to give feedback. This also allowed us to test Kraut et al's finding that it is more useful for an addressee to participate in a dialogue than a monologue. Experiment 6 investigated the issue of difficulty, but for people *overhearing* the same tangram descriptions. This experiment also allowed us to see if my results from Experiments 3 and 4 replicated, by testing if monologue or dialogue descriptions are more useful for overhearers. Additionally, by analysing the two experiments together, I assessed if Schober and Clark's (1989) finding that it is better to participate in a dialogue than to overhear one was supported. I also examined if this finding extended to monologue, despite the fact that addressees in the monologue condition are not able to interact with the speakers any more than overhearers.

6.3 Experiment 5

The first experiment employed a referential-communication tangram identification task, in which one partner, the 'describer', had to describe an array of tangrams to his partner, the 'matcher', who then had to select the correct tangrams from her (larger) collection and put them in a similar arrangement. Half the describer's

---

[54] It may be the case, though, that the amount of feedback will correlate with *addressee* success, particularly since they are the ones contributing this feedback.

tangrams had very similar counterparts in the matcher's set (alongside the correct tangrams), in which case the correct tangrams were considered to be particularly difficult to identify correctly. The other half of the tangrams had no similar counterparts, so these should have been easier to identify. It was predicted that a difference in difficulty for particular tangrams should influence the likelihood of accurate identification, and that this might interact with the amount of feedback permitted on the trial.

There were three feedback conditions in the present experiment: No feedback (in which the addressee was not allowed to speak to the speaker), Restricted feedback (in which the addressee was only allowed to speak to the speaker to say 'finished' when she had selected a tangram) and Full feedback (in which full verbal interaction was permitted between the partners). No visual feedback was present in any of the conditions. The Full feedback and No feedback conditions were used because the aim of the experiment is to analyse the potential interaction between difficulty and partner feedback, and these two conditions will produce the greatest possible contrast in amount of feedback. The Restricted feedback condition (Note: this differs from the Minimal feedback condition in Experiment 1) was included to analyse what effect it had on the speaker to only be told when his partner had heard enough information, without receiving concurrent feedback during the task. I also intended to determine which feedback condition Restricted feedback is most like; whether it is comparable to the Full feedback condition because notification of the addressee's completion of the task is the most important component of feedback, or whether it is more like No feedback, because the *content* of the addressee's feedback about the speaker's descriptions is crucial. Alternatively, it may produce results that are between the two; Krauss and Weinheimer (1966) found that when speakers in a picture-matching task were given confirmation that their addressees had selected the correct picture after each trial, this allowed the speakers to shorten their subsequent descriptions, but concurrent feedback from the addressees allowed an even greater amount of shortening. In a sense my Restricted feedback condition

is similar to this confirmation condition, in that it allows the addressee to confirm that she is satisfied with the speaker's description. It may be the case, then, that the participants in this condition produce results which are mid-way between the No feedback and Full feedback results.

The hypotheses for the first experiment are as follows.

1. Tangrams with no similar counterparts ('dissimilar' tangrams) will be more accurately identified by addressees than those with similar counterparts ('similar' tangrams).

2. Tangrams will be identified more accurately by addressees in the dialogue condition than in the monologue condition, corroborating the findings from Experiment 1 (this volume) and Clark and Krych (2004).

3. Identification of tangrams in the Restricted feedback condition will be less accurate than in the Full feedback condition, but more accurate than in the No feedback condition.

4. This benefit of feedback will be greater for similar tangrams, producing an interaction between tangram difficulty and feedback condition (cf. Hupet et al, 1991).

6.4 Pilot test

A pilot test was carried out on a set of materials selected by the experimenter, in order to assess if the pairs of tangrams categorised as 'similar' and 'different' were considered to be so by participants who were drawn from the same population as in the main experiment.

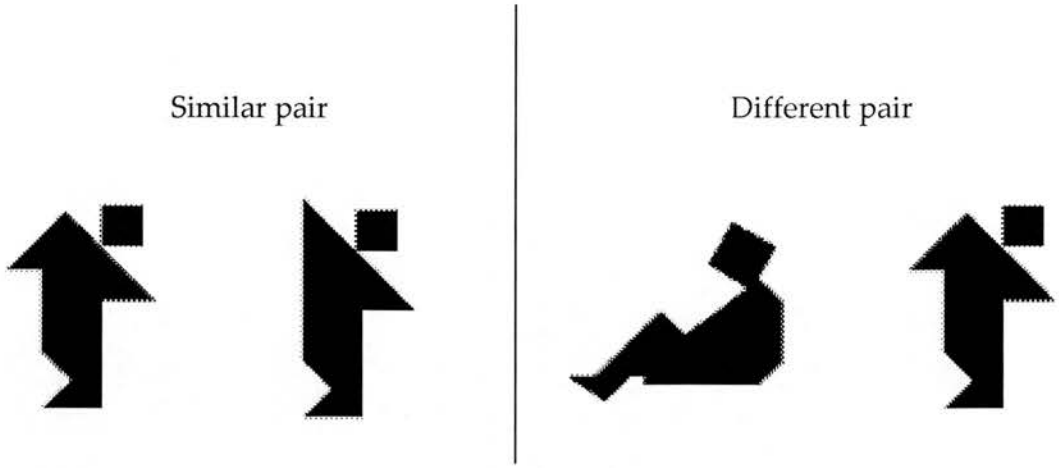
Participants

There were ten participants, all students at the University of Edinburgh. They participated voluntarily.

Materials and Procedure

Each participant was given a sheet of instructions, an answer sheet and 12 sheets of A6 paper; on each sheet there were two tangrams (taken from Elffer, 1976), which the experimenter judged to be either very similar to each other or very different. The materials are reproduced in Appendix E. Example tangram pairs are shown below in Figure 25.

Figure 25: Similar and dissimilar tangram pairs used in Experiment 6.



There were 24 pairs of items overall: 12 pairs of similar tangrams, and 12 dissimilar. The participants were split into 2 groups to allow the item pairs to be balanced. Each group saw 12 pairs of items: Group 1 saw experimental items 1-6 with similar counterparts, and items 7-12 with dissimilar ones. Group 2 saw experimental items 1-6 with dissimilar counterparts and items 7-12 with similar ones. That is, each participant saw the same 12 experimental items, but the tangrams they saw alongside them differed. The order of presentation of the pairs was randomised individually for each participant. Participants were asked to look at the tangram sheets in order and rate how similar the pairs were, on a scale from 1-5, with 1 being 'very different' and 5 being 'very similar'.

6.5 Results

The mean rating (between 1 and 5) given by pilot participants for the item pairs that were designated 'similar' was 4.3 (range: 4.0-4.6), and for 'different' item pairs was 1.7 (range: 1.4-2.2). Paired t-tests showed that this difference was highly significant by participants and items ($t1(9) = -24.53$, $p<.001$; $t2(11) = -27.3$, $p<.001$).

Participants

24 pairs of people participated in the experiment proper, in exchange for payment. All were students at the University of Edinburgh, none of the pairs knew each other prior to the experiment and none of the participants had taken part in any of experiments 2-4, or in the pilot test. One member of each pair was randomly designated the role of 'Describer', and the other, 'Matcher'. The participants were tested one pair at a time, and the experiment took between 10 and 35 minutes, depending on the pair.

Materials and design

The describer was given a set of 12 different tangram cards; describers in all conditions had the same cards. The matcher was given a set of 24 cards. 12 of these were exact replicas of the describer's cards, and 6 more were similar counterparts of 6 of these. A further 6 cards were dissimilar counterparts (the similarity and dissimilarity of counterparts was assessed by the pilot test). Two sets of matcher materials were used: 12 matchers saw Set A and 12 matchers saw Set B. Set A contained the describer tangrams 1-12, 6 counterparts similar to tangrams 1-6, and 6 dissimilar tangrams. Set B also contained describer tangrams 1-12, 6 counterparts that were similar to tangrams 7-12, and 6 dissimilar counterparts. In this way the similarity comparisons were balanced, so that each tangram was seen by half the matchers along with a similar counterpart (and so was 'similar'), and by half the matchers along with no similar counterpart ('dissimilar'). The describer knew that the matcher had 24 tangrams to choose from, but was not informed about which tangrams had similar partners, or about any aspect of the manipulation.

Great care was taken to ensure that in both sets, there were no two tangrams that were alike except for those which were intended to be similar for the purpose of the experiment.

The experiment was run in three feedback conditions: No Feedback, in which no feedback was permitted from the matcher; Full Feedback, in which there was free interaction between the partners; and Restricted Feedback, in which the matcher was not allowed to communicate with the describer except to say 'finished' when she had chosen a tangram. In terms of equipment, in the monologue condition, the describer was given a lapel microphone which was attached to a DAT recorder, and wore a pair of sound attenuating headphones to block out the sound of the matcher placing her cards. In the dialogue and restricted feedback conditions, no headphones were worn and both participants wore lapel microphones; these were coupled before inputting into a DAT recorder, producing sound files in which each participant was recorded on a separate channel. A third of the pairs (crossed across matcher item sets) participated in each feedback condition.

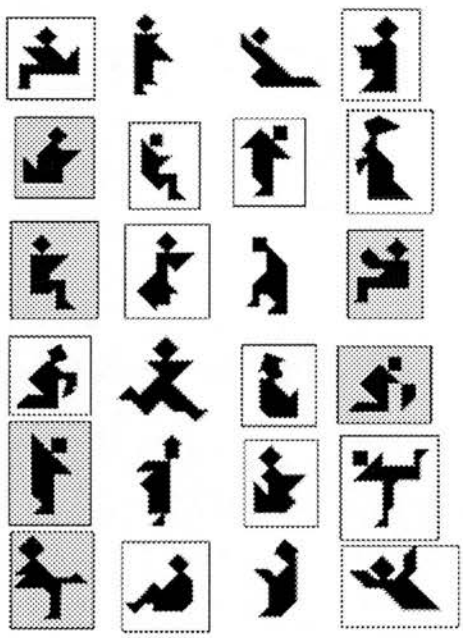
The experiment had a 3 (Feedback Condition, between-participants, within-items; no feedback, restricted feedback, full feedback) X 2 (Difficulty, within-participants and items; similar tangrams, dissimilar tangrams) design.

Procedure

The participants were seated at separate tables in the experimental room, facing opposite directions, so that they could hear each other's speech, but could not see each other or each other's materials. On the table in front of the describer were 12 tangram cards, face down, in two rows of six cards. In front of the matcher were 24 tangram cards, again face down, in four rows of six cards. A sample matcher

tangram card set is shown below in Figure 26. In the actual experiment all the tangrams were on identical white cards, but in this figure, the white boxes represent the 12 experimental tangrams (the tangrams which the describer described) and the grey boxes represent tangrams which are very similar to 6 of the experimental tangrams (and thus made those experimental tangrams 'similar').

Figure 26: Example matcher tangram cards



The order in which the describer and addressees' cards were arranged was randomised separately for every participant. Both participants were given instructions detailing their roles (reproduced in Appendix F) and were allowed to ask the experimenter any questions regarding the task before it began. Both participants were made aware of the feedback conditions imposed on the matcher.

The aim of the experiment was for the describer to describe his cards, and their position in the array, to the matcher, who would them pick out the matching

cards from her set and put them in a similar arrangement. Once the participants were sure of their task, they were asked to turn over their cards, keeping them in the same vertical orientation (to ensure that any lack of matching accuracy was not due to differing orientations of the cards between the partners) to reveal the tangrams. They were then requested to study the cards in front of them for approximately two minutes, in order to familiarise themselves with the images. When the experiment began, the describer described his cards to the matcher in any order he chose, subject to the restriction that each card could only be described once; the describer was not allowed to return to cards later on to describe them further. Once every card had been described once, the experiment was finished.

The main element for analysis was the proportion of tangrams identified correctly. A correct identification and placement was given a score of 1, and an incorrect one, 0, giving a mean value between 0 and 1 for every participant in every condition. The mean accuracy by feedback condition and difficulty is shown below in Table 20.

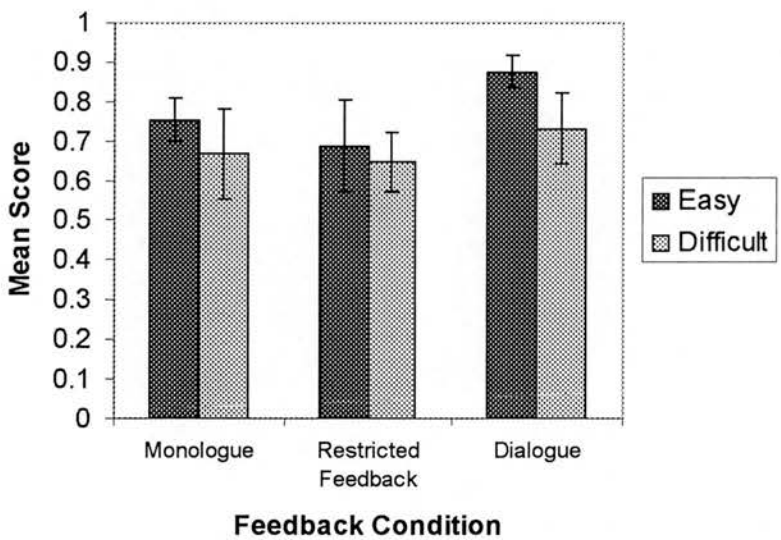
Table 20: Proportion of tangrams correctly identified by addressees, in three feedback and two difficulty conditions

Feedback	Similar	Dissimilar	Mean
Monologue	.67	.75	.71
Res Feedback	.65	.69	.67
Dialogue	.73	.88	.80
Mean	.68	.77	.73

3 (feedback condition) X 2 (difficulty) repeated-measures ANOVAs showed no main effects of either Feedback ($F1(2,21) = 1.58$, NS; $F2(2,33) = .951$, NS) or Difficulty ($F1(1,21) = 1.54$, NS; $F2(1,33) = 2.60$, NS) and no interaction ($F1(2,21) = .220$, NS;

F2(2,33) = 2.73, NS). The full feedback and no feedback conditions were then compared directly, in case the restricted feedback condition (which I expected to show results half-way between the other two conditions) had clouded the data. 2 X 2 ANOVAs comparing only the monologue and dialogue conditions, then, also showed no significant effects (Feedback: (F1(1,14) = 1.33, NS; F2(1,22) = 1.06, NS); Difficulty: (F1(1,14) = 2.25, NS; F2(1,22) = 2.79, NS); Interaction: (F1(1,14) =.25, NS; F2(1,22) = .31, NS)). The pattern of data is demonstrated in Figure 27 below, with the error bars representing standard errors.

Figure 27: Proportion of tangrams correctly identified by addressees, in three feedback and two difficulty conditions



The mean times taken for each whole experiment (all 12 tangrams described and matched by each participant pair) are shown below in Table 21, after being

truncated to 2.5SDs above the mean for each condition.

Table 21: Mean timings per experiment in all feedback conditions.

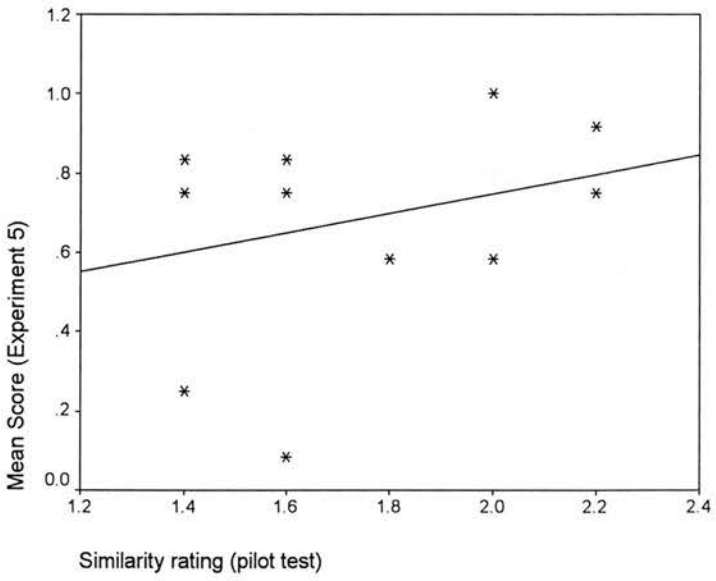
Feedback	Time (sec)
Monologue	510.2
Res Feedback	505.2
Dialogue	909.3

One-way ANOVAs showed a main effect of Time ($F1(2,23) = 8.28$, $p<.01$), and independent t-tests demonstrated that there were significant differences between monologue and dialogue ($t1(14) = -2.92$, $p=.01$) and between restricted feedback and dialogue ($t1(14) = -3.28$, $p<.01$), where the dialogue condition was longer in both cases. There was no difference between the monologue and restricted feedback conditions ($t1(14) = -.07$, NS). There was no significant correlation between length of experiment and mean accuracy ($r(24) = -.032$, NS).

The scores for each tangram in the similar context were compared with their ratings in the pilot experiment; there was no correlation ($r(11) = .284$, NS). The scatter plot is shown below in Figure 28.

Figure 28: Correlation between tangram identification score and ratings in pilot test

for 'similar' tangrams.



6.6 Discussion

The results obtained are inconclusive: there was no support for any of the four hypotheses. There was no significant difference between the tangrams in similar and dissimilar contexts, which contrasts with Hypothesis 1, which predicted that tangrams in the dissimilar contexts would be identified more accurately. The lack of correlation between tangram identification score and rating in the pilot test concurs with this null finding.

Possibly the more surprising aspect of these results, though, is that I failed to find a significant difference in score between the no feedback and full feedback conditions, and so Hypothesis 2, which predicted that tangrams would be identified more accurately in the full feedback condition, was not supported. It might have been expected that participating in a dialogue would produce more accurate

identifications than hearing a monologue, if only because in a dialogue, the describer's speech will be tailored to the matcher's requirements. Being able to give feedback has shown to be beneficial in almost every other context tested (including other referential communication tasks – see Experiment 1 and Clark and Krych, 2004, and a narrative context, Kraut et al, 1982), and so it would be surprising if this were the exception. However the present result suggests that participating in a dialogue guarantees no more addressee success than participating in a monologue. Neither was Hypothesis 3, that the restricted feedback condition would produce scores mid-way between no feedback and full feedback, supported. There was also no interaction between feedback condition and difficulty. This gives no support to Hypothesis 4, which predicted that the benefit of feedback would be greater for tangrams in difficult contexts.

One reason why there were no significant effects, despite a trend in means in the expected directions (.77 for dissimilar tangrams, .68 for similar; .71 for no feedback; .80 for full feedback) may be that there were too few participants in this study. Since feedback condition, in particular, was run as a between-participants variable, this means that there were only 8 participant pairs in each feedback group, which may not have been enough to show a significant effect. A comparison with Experiment 2 from this thesis might be helpful. Although it is unclear whether the addressees in Experiment 2 performed better in the dialogue condition than the monologue condition (because the scores were not recorded at the time), Experiments 3 and 4 demonstrated that overhearers of those descriptions performed significantly better on overhearing the dialogue descriptions than the monologues. Although the performance of addressees and overhearers on this type of task is not typically identical (for example Schober and Clark (1989) demonstrated that addressees perform better than overhearers), it seems that they are both likely to be correlated with the quality of the speaker's speech; Kraut et al (1982) found that variations in the quality of the speaker's speech (which in this case was improved by feedback) affected both addressees' and overhearers' performance, but

crucially, it affected addressees' performance to a greater extent. Drawing from this, I would expect that the addressees' performance in Experiment 2 is likely to have shown an even greater difference between the monologue and dialogue conditions than was shown by those who subsequently overheard the descriptions (in Experiments 3 and 4), and so that study may have shown the effect that is lacking in the present experiment[55].

It may be the case that a comparison between other aspects of Experiment 2 and the present experiment will explain the difference in findings I found here and those I expect were present in Experiment 2. I will make two comparisons here, on the time-course of both experiments and on the number of turns in dialogue.

Comparison of times and turns taken in Experiments 2 and 5

The mean times taken to describe 12 tangrams for the first time in the monologue and dialogue conditions in Experiments 2 and 5 are shown below in Table 22. These 12 descriptions represented one eighth of Experiment 2 (since the tangrams were then described 7 more times each), and the whole of Experiment 5 (since each tangram was only described once).

---

[55] Experiment 1 from this thesis also showed that addressees hearing dialogues attained higher scores than those hearing monologues, however the experimental paradigm was not directly comparable to this one, as it was a building task rather than a picture-matching task.

Table 22: Mean description times for first 12 tangram descriptions in Experiments 2 and 5.

Condition	Expt 2 (sec)	Expt 5 (sec)
Monologue	316.6	510.2
Dialogue	347.3	909.3

It is clear from this comparison that the times taken to describe tangrams differ markedly between the two experiments. In Experiment 2, the difference between the times taken for the monologue and dialogue conditions was not significant ($t1(14) = -.85$, NS); in Experiment 5, dialogue trials took significantly longer than monologue trials ($t1(14) = -2.92$, $p=.01$). One thing that is notable here, however, is that the mean times for descriptions in Experiment 5 are much longer than those for Experiment 2.. This gave rise to a main effect of Experiment ($F(1,14) = 21.07$, $p<.001$), although the interaction between Experiment and feedback condition did not reach significance ($F(1,14) = 2.74$, $p=.1$). There was also a difference between the number of turns taken in dialogue in Experiments 2 and 5; the mean turns taken for the first 12 tangram descriptions are shown below in Table 23. These results are also significantly different, with the number of turns in Experiment 5 being greater than that in Experiment 2 ($t1(14) = 4.04$, $p<.001$).

Table 23: Mean turns taken in dialogue for first 12 tangram descriptions in Experiments 2 and 5

	Expt 2	Expt 5
Turns	53.2	134.8

It seems that the task in Experiment 5 was more difficult than that in Experiment 2. In Experiment 5, the addressees had to choose 12 tangrams from a selection of 24

and put them in the right order, whereas in Experiment 2, they had only the correct tangrams in front of them, and simply had to put them in the right order. It is possible that this extra difficulty in Experiment 5 could account for the dramatic difference in time and number of turns taken in dialogue; the speakers in dialogue may have been alerted to the difficulty of the task (and the added difficulty that was caused by the presence of similar tangrams) by their addressees, and so they may have interacted more with their partners in an effort to complete the task successfully.

Because the participants apparently found Experiment 5 so difficult, it might seem surprising that dialogue did not confer an advantage over monologue in this situation. The explanation for this may lie in the addressees' technique. There is a slight possibility (albeit without any experimental evidence to support it) that the actual process of matching tangrams will have differed between the monologue and dialogue conditions in the present experiment. Perhaps when an addressee in the monologue condition heard the speaker describe a tangram, she scanned the set of tangrams while looking for one that fit the description, and continued to scan the set while the speaker was speaking, to make sure that she had selected the most appropriate tangram. In doing this, she would be highly likely to have noticed any other tangrams that were similar to the one she had provisionally chosen. In dialogue, things may have happened slightly differently; when an addressee noticed a tangram that might have fit the speaker's description, she will have begun to ask questions about that tangram, and answered questions about it from the speaker (this is certainly suggested by the large number of turns taken). In a sense, then, she had committed to this tangram from an early point and needed to focus all her attention on it. This, along with her interaction with the speaker, will have given her less opportunity to keep scanning the set, and less chance of noticing any similar tangrams. (This would be less of an issue in Experiment 2, because there are fewer tangrams to scan, and so the addressee was likely to have been able to memorize these at an early stage of the game). This may have resulted in two

effects cancelling each other out; while in dialogue the addressees had the benefit of being able to ask questions, they may have focussed less on possible alternatives, and as such been lulled into a false sense of security about the tangrams they were focussed on. In monologue, the addressees may have focussed more on the whole tangram selection, but they did not have the advantage of being able to ask questions of the speaker. Hence one possibility is that the interaction with a partner in this particular context actually distracted the addressee from completing the task successfully.

6.7 Experiment 6

Experiment 6 was an overhearer experiment which employed the same experimental paradigm as Experiments 3 and 4, using the sound files of tangram descriptions generated in Experiment 5. The sound files from the whole experiments were divided up into descriptions of individual tangrams, which were randomised and played to participants who had exactly the same tangrams to choose from as the addressees in Experiment 5 had. The overhearers' accuracy of identification was again noted.

It is hypothesised that:

1. Dissimilar tangrams will be more accurately identified by overhearers than similar ones.

2. Tangrams described in a dialogue condition will be more accurately identified by overhearers than those described in a monologue condition (replicating Fox Tree, 1999, and Experiments 3 and 4 of this thesis).

3. Identification of tangrams in the Restricted feedback condition will be less accurate than in the Full feedback condition, but more accurate than in the No feedback condition.

4. The benefit of feedback will be greater for similar tangrams, producing an interaction between tangram difficulty and feedback condition.

Participants

24 participants completed the experiment, all students at the University of Edinburgh. None of them had taken part in any of Experiments 2-5. The experiment took approximately 10 minutes.

Materials and design

This experiment, similarly to Experiments 3 and 4, was run on E-Prime software. Each participant heard a description of each of the 12 experimental tangrams from Experiment 5, and also simultaneously saw an array of 24 tangrams on the computer screen (12 of which were the experimental tangrams), corresponding to the tangrams seen by the matcher in Experiment 5. As in Experiment 5, the participants were split into two groups: Group A saw tangrams 1-12 in addition to 6 other tangrams that were very similar to tangrams 1-6 (and a further 6 dissimilar tangrams), and Group B saw tangrams 1-12 in addition to 6 tangrams that were

similar to tangrams 7-12 (and a further 6 dissimilar tangrams). The descriptions heard by overhearers in these two groups were those produced by addressees in the same two groups (hence the overhearers saw the same tangrams as the addressees saw in Experiment 5). Although for any one participant the same tangrams appeared on the computer screen throughout the whole experiment (so that they had the same opportunity to become familiar with the tangrams over the trials as the original addressees had), their order was randomised after each description. The order of tangram descriptions heard was also randomised, such that no two participants heard descriptions of the 12 tangrams in the same order, nor did they hear the same describers describing them. The three levels of feedback condition and two levels of task difficulty condition were crossed so that every participant heard two descriptions in every possible combination of conditions.

Procedure

The procedure was exactly the same as in Experiments 3 and 4: each participant saw an instruction screen, then when they pressed the SPACEBAR, the first array of tangrams appeared. They were instructed to familiarise themselves with the array for a couple of minutes before pressing the SPACEBAR again to hear the first description. After picking out the tangram that they thought matched that description, they pressed the SPACEBAR again to see the next tangram screen and hear the next description. This continued for 10 more trials, making a total of 12 trials overall.

6.8 Results

Again, correct responses were allocated a score of 1, and incorrect responses, 0. The mean scores per feedback and difficulty condition are shown below in Table 24.

Table 24: Proportion of tangrams correctly identified by overhearers, in three feedback and two difficulty conditions

Feedback	Similar	Dissimilar	Mean
Monologue	.58	.60	.59
Res Feedback	.67	.65	.66
Dialogue	.60	.64	.62
Mean	.62	.63	.64

3 (feedback condition) X 2 (difficulty condition) repeated-measures ANOVAs demonstrated no effect of Feedback condition, ($F1(2,46) = .72$, NS; $F2(2,22) = .76$, NS), or Difficulty ($F1(1,23) = .03$, NS; $F2(1,11) = .006$, NS) and no interaction ($F1(2,46) = .52$, NS; $F2(22) = .564$, NS). As before, 2X2 ANOVAs on just the monologue and dialogue conditions showed a similar lack of effect in all areas (all ps>.1).

6.9 Discussion

Again, these results are somewhat unexpected; there were no significant effects at all. Similarly to the previous experiment on addressees, there was no difference between dissimilar and similar tangrams, so Hypothesis 1 (which predicted such a difference) was not supported. Neither did this experiment demonstrate a

significant difference in accuracy between overhearers of monologue and dialogue descriptions, contrary to findings by Fox Tree (1999), and the results from Experiments 3 and 4 earlier in this thesis. This does not support Hypothesis 2, which predicted that dialogue should produce more accurate identifications than monologue. Neither was Hypothesis 3 – that Restricted dialogue should give results midway between monologue and dialogue – supported. Finally, Hypothesis 4, which predicted that there should be an interaction between the expected effects of difficulty and feedback, was not supported.

As in the previous experiment, the most surprising result here is the lack of effect of feedback condition. The similarities between this experiment and Experiments 3 and 4 lend themselves to a direct comparison here, since all three experiments involved overhearers hearing recordings of tangram descriptions which were produced within a referential communication context. Maybe such a comparison can give us an idea why there was no difference in identification accuracy between the monologue and dialogue conditions in this experiment, when a significant difference was found in both Experiment 3 and 4. The mean accuracy scores for overhearers from all three experiments (reference 1 only for Experiments 3 and 4) are shown below in Table 25.

Table 25: Overhearers' accuracy of tangram identification in monologue and dialogue conditions: comparison between experiments 3, 4 and 6

Experiment	Monologue	Dialogue
3	.46	.64
4	.51	.66
6	.59	.62

These means strongly suggest that the monologue condition in Experiment 6 (the current experiment) showed higher scores than expected, rather than the dialogue condition showing lower scores. It is possible that this can be explained by the context in which the tangrams were seen. In Experiment 2 of this thesis (from which the sound files for Experiments 3 and 4 were taken), the matcher saw only the 12 experimental tangrams, with no fillers. In the present experiment, each matcher saw 24 tangrams: 12 experimental and 12 fillers, and so there was a smaller ratio of experimental items to fillers. Since the describers in Experiment 5 *knew from the outset* that their matchers had a large number of tangrams to choose from, it is possible that this affected their descriptions; that they were more detailed in describing tangrams than they would have been otherwise[56]. The average duration of 12 tangram descriptions in Experiment 5, as mentioned earlier, was 510 seconds, in comparison with 317 seconds for 12 descriptions of tangrams in Experiment 2 (where there were only 12 tangrams to choose from). This obviously shows a higher level of detail in the present descriptions.

But why should this affect the monologue and dialogue conditions differently? One of the key benefits of dialogue appears to be that speakers can tailor their utterances to meet their addressees' needs, which will involve giving them whatever extra information they ask for. If a describer in monologue is aware of the need to be particularly precise and if he therefore deliberately gives as much specific information as he can in order to make his matcher's apparently difficult task easier, then this might reduce the disadvantage of being in monologue. In a sense, then,

---

[56] In Fox Tree's experiment, whose results concurred with Experiments 3 and 4 but not Experiment 6, the matchers saw 16 tangrams, 12 of which were experimental, and 4 of which were fillers, and so again, as in Experiments 3 and 4, they did not have a large selection of tangrams to choose from. Additionally, it is unclear whether the describers in Fox Tree's study were aware of the number of tangrams held by the matchers.

the describer's awareness of the difficulty of the matcher's task may provide knowledge which could affect his behaviour. Certainly Krauss and Fussell (1992) mention two sources of information on which speakers can base audience design: firstly, concurrent feedback, and secondly, prior knowledge of the situation, which seems applicable to the speaker's awareness of the addressee's task. However the findings on timing (Table 22 above) do not support this explanation; dialogue trials took significantly longer than monologue trials, where we would have expected the opposite.

Experiments 5 and 6 together: A comparison of listener and overhearer effects

How did the accuracy of overhearers compare with the accuracy of the original addressees? Looking at all the data together, the mean accuracies of addressees and overhearers, in only the monologue and dialogue conditions, are shown below in Table 26.

Table 26: Accuracy of tangram identification for addressees and overhearers in three feedback and two difficulty conditions

	Condition	Similar	Dissimilar	Mean
Experiment 5	Interact Mon	.67	.75	.71
	Interact Dial	.73	.88	.81
Experiment 6	Overhear Mon	.58	.60	.59
	Overhear Dial	.60	.65	.63
	Mean	.65	.72	.69

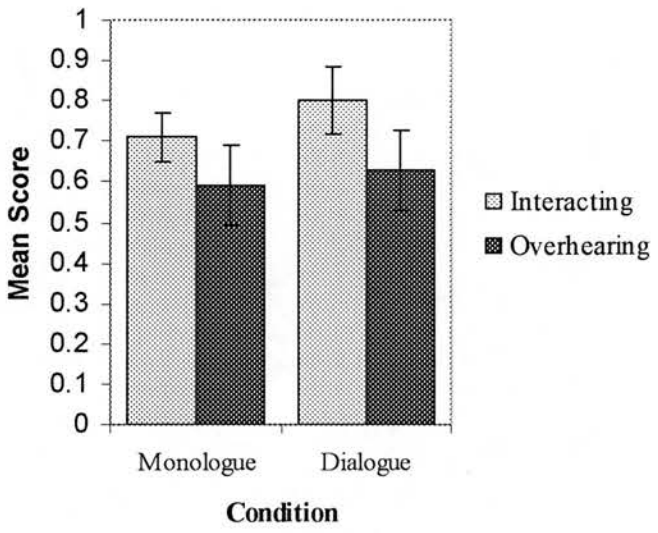
2 (Difficulty, within participants and items) X 2 (Feedback, between participants[57], within items) X 2 (Experiment number, between participants, within items) ANOVAs on only the monologue and dialogue conditions demonstrated that there was no effect of Feedback condition ($F1(1,60) = .64$, NS; $F2(1,11) = .97$, NS) or Difficulty ($F1(1,60) = 1.36$, NS; $F2(1,11) = 2.15$, NS). There was however an effect of Experiment number, where Experiment 5 (addressees) showed higher scores than Experiment 6 (overhearers) (marginal by participants and reliable by items: $F1(1,60) = 3.46$, $p=.068$; $F2(1,11) = 20.7$, $p<.01$). None of the interactions were significant (all $ps>.1$). Splitting up the monologue and dialogue conditions, in independent t-tests only the dialogue condition showed an individual effect of Experiment number, although it was marginal by participants (Dialogue: ($t1(30) = 1.74$, $p=.09$; $t2(22) = 2.38$, $p<.05$); Monologue: ($t1(30) = .96$,NS; $t2(22) = 1.27$,NS)).

The data are shown in graphical form below (Figure 29), demonstrating the accuracy of identification for monologue and dialogue, collapsed over task difficulty, where the participant was either interacting (Experiment 5) or overhearing (Experiment 6). The error bars represent Standard Error margins.

---

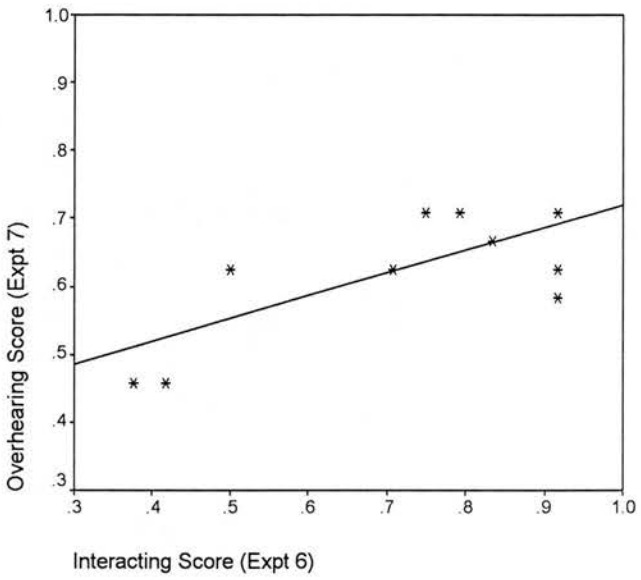
[57] Although Feedback was a within-participants variable in Experiment 6, it was between-participants in Experiment 5, and therefore we treated it as between-participants overall; the effects obtained may be overly conservative as a result.

Figure 29: Accuracy of tangram identification for addressees and overhearers in two difficulty conditions



Plotting the overhearer accuracy against the addressee accuracy for individual tangrams, these two factors correlated significantly ($r(10) = .719$, $p<.01$), suggesting that those tangrams that were easier for addressees to identify (possibly because the speaker described them well) were also easier for overhearers to identify. The data pattern is shown below in Figure 30.

Figure 30: Comparison of Overhearer and Addressee accuracy for individual tangrams



6.10 General discussion

The results from Experiments 5 and 6 did not support any of my hypotheses, as there were no main effects of tangram difficulty or feedback condition, and no interactions. The lack of a difference between monologue and dialogue in both experiments has brought to my attention a crucial aspect of studies such as this; that is, that the results obtained in such studies may rely greatly upon even small details of the experimental paradigm. Speakers' knowledge of even something as apparently insignificant as the number of tangrams which their partners have to choose from may have been used to aid their use of audience design in this context. This concurs with reports from Isaacs and Clark (1987), Kingsbury (1968, in Krauss and Fussell, 1996) and Fussell and Krauss (1992) that speakers' speech is affected by

perceptions of their addressees' knowledge or current situation, as distinct from the effect of feedback. Additionally, the lack of significant effects in Experiment 5 may reflect power problems, as a result of having too few participants, as discussed in Section 6.3.6.

It appears from the comparison between the two experiments that participating in the referential communication task (in Experiment 5) was more beneficial than overhearing it (in Experiment 6), but only for the dialogue condition. Although there was no interaction between feedback condition and experiment, there was an individual effect of experiment on the dialogue condition alone. This concurs with Schober and Clark's (1989) finding that being a participant in a dialogue is more helpful than overhearing that dialogue. Additionally, I have extended this finding to demonstrate that, within this context, there was no significant difference between being the addressee of a monologue and being an overhearer. This is understandable, given that even an addressee (as opposed to an overhearer) in a monologue condition is not permitted any interaction whatsoever with the speaker, and so the roles of addressee and overhearer are more similar in monologue conditions than they are in dialogue.

However we must be cautious with drawing these conclusions, as there were many differences between the addressee and overhearer experiments that could account for the difference between them. The addressees in the dialogue condition of Experiment 5 may have had more motivation to select tangrams correctly because they felt responsibility towards their partners, which would have inclined them to put a lot of effort into the task because they did not want to let the speakers down. This may not have been the case for the overhearers in Experiment 6, who carried out the task alone on computer, and therefore were not involved in such a 'team effort'. The addressees also heard a whole monologue or dialogue, in the order it was presented, by only one speaker. In comparison with this, the overhearers

heard a mixture of speakers in all three feedback conditions, and in a mixed-up order, which may have impaired their performance[58]. Even the simple fact that Experiment 6 took place on a computer, in comparison to Experiment 5 taking place on a table with plastic tangram cards, may have had an effect. Any or all of these differences could be responsible for the apparent difference between experiments.

---

[58] Although Experiments 3 and 4 also used the same procedure, and they produced comparable results to Fox Tree (1999), in which the monologues and dialogues were played whole to overhearers.

Chapter 7: Conclusion

The previous four chapters have described a total of six experiments that studied the effect of feedback on speakers, addressees and overhearers: three referential-communication tasks (between speakers and addressees) and three overhearer experiments (where the descriptions from the referential-communication experiments were replayed to overhearers). Since this thesis set out to examine the effect of feedback on three distinct roles, those of speakers, addressees and overhearers, the results will be summarised according to these themes.

7.1 Effect of feedback on speakers

The purpose of the first two experiments was to investigate the effect of feedback on two main aspects of speakers' productions: the length of their object descriptions, and their consistency of word-choice.

Length of object descriptions

The results from Experiments 1 and 2 demonstrated that receiving feedback from an addressee affected the number of words speakers used to describe objects. Experiment 1 employed a referential communication task in which one person from the participant pair (the 'describer') had to describe to his partner (the 'builder') how to construct a Lego model identical to the one the describer had. This was run in three feedback conditions: no feedback/monologue (in which the builder was not allowed to speak), full feedback/dialogue (in which both partners could speak freely) and minimal feedback (in which the builder was only allowed to use

single word, non-contentful terms such as 'expand', 'wait', 'yes' and 'no'). Each pair of participants completed nine trials, building a different model in each trial.

Over the course of the experiment, speakers reduced the length of the first noun phrases they used to describe the Lego models. This finding was supported by a reduction in later trials in the mean number of features in the first description of each Lego block. Looking at the effect of feedback on this shortening, first descriptions of Lego blocks were longer at the start of the experiment in the monologue condition than in the dialogue condition, but they then proceeded to reduce more in length than the dialogue descriptions. This resulted in descriptions of new blocks being approximately the same length in monologue and dialogue by the end of the experiment, and it was surmised that the greater decrease in length of monologue descriptions occurred only because those descriptions began at a longer length.

Experiment 2 was similar to Experiment 1, but used tangrams instead of Lego models, and involved a matching task rather than a building task. It looked further at the shortening of descriptions by speakers, but this time, instead of looking at how references to *new* objects shortened during the course of the experiment, it examined how references to the *same* objects shortened with repetition. The findings demonstrated overall shortening of repeated descriptions in both monologue and dialogue conditions. Looking at only the first utterance from the speaker in each description (until the addressee spoke), dialogue descriptions were initially shorter than monologue descriptions (as in Experiment 1), and then descriptions in both conditions reduced at a similar rate over repetitions, resulting in the dialogue descriptions being shorter overall than the monologue descriptions.

The fact that descriptions in dialogue were initially shorter than those in monologue in both Experiments 1 and 2 could reflect speakers taking account of their

addressees' ability to give feedback even at this early point, possibly describing objects briefly because they knew that their addressees could ask for more information if they needed it. This pattern could also have arisen because the addressees in the dialogue condition may have interrupted at a point before the speakers would have naturally stopped speaking, thus shortening the first descriptions.

The difference in shortening of descriptions between Experiment 1 (where references to new objects ended up similar lengths in monologue and dialogue) and Experiment 2 (where repeated references to the same objects ended up shorter in dialogue than in monologue) may be attributable to the difference in the amount of common ground in these two cases. It is likely that the shortening of descriptions with repetition in Experiment 2 was a result of those objects entering the common ground between the speaker and addressee, as a result of previous references to the same objects. This may not have occurred with the new references in Experiment 1 because the speakers had not described those objects previously, and so there was no relevant common ground to appeal to. This apparent role of common ground in shortening certainly concurs with the Collaboration framework of dialogue (e.g. Clark and Brennan, 1991), but, interestingly, these results suggest that feedback from an addressee does not play a major role in inducing shortening with repetition, contrary to reports from Krauss and Weinheimer (1966).

Consistency of descriptions

Another aspect of speech that did not seem to be directly affected by the amount of feedback was the choice of words used by the speakers. In particular, the consistency of the nouns chosen to refer to Lego blocks the first time they were described (e.g. 'block', 'piece', 'bit') in Experiment 1 was the same in monologue and dialogue conditions. However the overall repetitiveness of speakers increased

more in monologue than in dialogue, suggesting that other elements of the speaker's productions may have been repeated more consistently than the nouns. In Experiment 2, tangram descriptions became more consistent with repetition, in terms of lexical overlap between subsequent descriptions, but as in Experiment 1, this occurred to a similar extent in monologue and dialogue once the length of descriptions was taken into account. This suggests that there is a direct relationship between the length of descriptions and their consistency; which factor is the driving force here is not clear from the results of this experiment. These findings do however seem to suggest that conceptual pacts are not responsible for lexical consistency in dialogue, since the consistency noted in dialogue occurs to an equal extent in monologue. Of course it is possible that the increase in consistency in monologue may occur for a different reason than that in dialogue.

7.2 Effect of feedback on addressees

Experiment 1 showed a notable effect of feedback on the behaviour of addressees. Addressees in the dialogue condition performed better than those in the monologue condition overall (in terms of task score), despite the speakers taking the same mean amount of time per trial and using approximately the same number of words in both conditions. The minimal feedback condition showed results that were roughly halfway between the monologue and dialogue conditions, suggesting that although even a little feedback was beneficial to interaction in this context, it was not equivalent to being in a fully communicative situation, and so *quantity* of feedback mattered here, rather than its mere presence or absence.

7.3 Effects of feedback on overhearers

Fox Tree (1999) found that dialogue descriptions resulted in greater accuracy of tangram identification by overhearers than monologue descriptions. Experiments 3 and 4 replicated that result, but more importantly tested two hypotheses she proposed to explain this. Both experiments involved replaying manipulated versions of the tangram descriptions from Experiment 2 to a new set of participants, and asking them to complete the same task as the original addressees. Experiment 3 analysed the effect of splicing out the addressees' feedback (and therefore to some extent, the addressees' perspective) on overhearers' performance. Experiment 4 kept the addressees' feedback intact, but spliced out the discourse markers in the speakers' productions. However neither manipulation caused a disadvantage for overhearers in comparison with the full dialogue descriptions, suggesting that neither the dual perspectives nor any aspect of the actual feedback itself were responsible for the benefit of dialogue.

A second feature of the design was that in both experiments I varied the reference number of the descriptions; that is, the number of times the same tangram had been previously described by the speaker. Overhearers heard either the first, second or eighth description of each tangram. In both experiments, dialogue descriptions still produced more accurate identification than monologue descriptions even in the eighth descriptions, in which there was a minimal amount of feedback. This further supports the idea that the benefit of feedback for overhearers does not stem from hearing the actual feedback (as there was very little by this stage). I concluded that the important factor for overhearers must be the effect that the addressees' feedback has on the speakers' speech. It is likely that in the dialogue conditions, the interaction between speakers and addressees led to the formulation of descriptions which were more comprehensible to others. Kraut et al (1982) suggested that speakers can better design their utterances for *all* addressees when they have

information about *any* addressee's comprehension, and this may have been the case here. This manipulation of reference number also allowed me to analyse the effect of another aspect of dialogue on overhearers: the presence of conceptual pacts between partners in dialogue. If these are built up gradually during the course of a dialogue (as proposed in Brennan and Clark, 1996), then they should be drawn upon more in later references than in earlier ones. The fact that there was still a benefit of dialogue over monologue for final references suggests that conceptual pacts, if they were formed, did not present any disadvantage to overhearers in this context.

7.4 Effects of task difficulty and feedback on addressees and overhearers

Two final studies assessed the effect of task difficulty on addressees' and overhearers' performances. Experiments 5 and 6 again used the referential-communication and overhearer paradigms. Experiment 5 was similar to Experiment 2, in that it involved two participants working together to match a set of tangrams, although in this case, the addressees had to select 12 tangrams from a set of 24, rather than simply re-ordering the 12 tangrams they had. The 'difficulty' of certain tangrams was manipulated by introducing very similar counterparts into the set, and the experiment was run in monologue, dialogue and restricted feedback conditions.

The results showed no effect of either feedback condition or tangram difficulty on the accuracy of tangram identification by addressees. Experiment 6, which involved playing the descriptions from Experiment 5 to overhearers, also showed no effect of either variable. It was suggested that the lack of significant results was attributable to the experimental design in Experiment 5, where the difficulty of the addressees'

task was increased by having a particularly large number of tangrams to choose from, and where the power was reduced by not testing enough participants.

A comparison between Experiments 5 and 6 showed that addressees performed better than overhearers in the dialogue condition, in keeping with Schober and Clark (1996). Additionally, it showed that there was no such benefit for addressees in the monologue condition, suggesting that it is the actual ability to give feedback that benefits addressees, rather than any other aspect of the social context.

The results of the experiments reported here demonstrate the important role of feedback in communication, such that it enables speakers to design their utterances for the benefit of their addressees, and concurrently allows addressees to partially determine the content of the speakers' productions. Together these processes ensure both the speakers' efficiency, and the addressees' comprehension of the speakers' utterances.

7.5 Future directions

Although the experiments reported above answer some of the outstanding questions regarding the role of feedback in dialogue, they also open up a number of areas for future research. Experiment 1 demonstrated that some general shortening of descriptions occurred even in the apparent absence of what we would consider to be common ground (previous reference to the same object) [59]. It would be interesting to see if this shortening of new descriptions only relates to descriptions

---

[59] Of course this may have been an artefact of the design; given that all the Lego blocks will have had elements in common with some of the previous blocks (for example, being 'blue' or 'long'), these descriptive terms may have constituted common ground for the interlocutors.

of objects of similar types (as in this study), or if the previous production of Lego descriptions would cause initial shortening of, say, tangram descriptions (where we can be fairly sure there is little common ground). A referential-communication study on this would enable us to assess how much of this shortening effect is due to a building up of common ground between the speaker and addressee, and how much is simply a result of the speaker understanding that the addressee will ask for more information if it is required. The increasing overall repetitiveness found in the monologue condition of Experiment 1 is another effect of interest, primarily because it is not echoed in the consistency of nouns used to describe Lego blocks; this suggests that it may have resulted from a repetition of instructions to the addressee (for example, 'Fasten that block onto the final four squares of the long red one'), or from less object-related utterances, such as 'I hope you've got that one, now I'll move on'. Alternatively, it may have been due to a repetition of other aspects of the noun phrases, for example the adjectives. A discourse analysis of the speech used in monologue and dialogue might show us the roots of this effect. Additionally, it may be interesting to go further into the structure of the Lego descriptions: whether speakers referred to 'the six by two' or 'the two by six', for example, and if there was any method to the order of dimension descriptions in this sense.

The findings on shortening and consistency in Experiment 2 have interesting implications for the idea of collaboration in dialogue; it appears that shorter descriptions tend to be more consistent than longer ones. It may be the case, then, that the main focus of collaboration is to shorten descriptions, and this leads to lexical consistency, or else lexical consistency may be the driving force that leads to shortening. Since this study showed that lexical consistency occurred simultaneously with shortening, it would be interesting to find a way of dissociating these effects. This could be tested by running a pair of referential communication experiments. In the first experiment, speakers would be given particular words that they must include in their descriptions (thus ensuring lexical consistency to some extent), but would otherwise be free to describe the objects in any way they liked.

We would measure the length of descriptions. In a second study we could restrict the *number* of words speakers were allowed to use to describe objects, and measure the resulting consistency.

A comparison of the relationship between shortening and consistency in these two experiments would show us which experiment produced results most like Experiment 2 of this thesis, and so this might allow us to see whether consistency or shortening is the driving force.

Experiments 3 and 4 led to the conclusion that descriptions in dialogue become more comprehensible to overhearers as a result of the collaboration between speakers and addressees. This could be explored in a very controlled way, where pairs or larger groups of participants decided on appropriate descriptions for tangrams and then a second set of individual participants had to select the correct tangram from a selection. It would be expected that the more people were involved in creating the original tangram description, the easier the tangram would be for other people to identify. Additionally, further studies on discourse markers could be carried out to investigate the effect of different discourse markers (besides those chosen by Fox Tree) on overhearers.

Experiments 5 and 6 did not produce the expected results, and I suggested that this was a result of methodological factors. Since one issue with these experiments may have related to the number of tangrams the addressees had to choose from being too large, it would be beneficial to replicate them using a smaller proportion of non-experimental tangrams to experimental ones, and also test more participants. We would expect that this would produce a main effect of feedback condition and of tangram difficulty for both addressees and overhearers, and that there might be an interaction between these two factors.

One concern brought up in this thesis regards the validity of previous studies where in the monologue conditions, the speakers did not know that their addressees could not give feedback (e.g. Bavelas et al, 2000; Kraut et al, 1982). I have hypothesised that this would have had a substantial effect upon the speakers' performances, but a direct comparison between monologue conditions where the speakers were either aware or unaware of the restriction on the addressees would confirm this.

The results from these studies demonstrate that although feedback benefits the performance of people producing and overhearing it, its effects on speakers are more complex. It appears that the main benefit of feedback for speakers is in terms of allowing them to produce more succinct speech from the outset, accomplishing more with fewer words, rather than affecting their actual choice of words. In this sense, then, feedback is highly useful in terms of allowing speakers to convey information with the minimum of effort; put simply, *more* interaction results in *better* communication.

References

Bangerter, A., & Clark, H.H. (2003). Navigating joint projects with dialogue. *Cognitive Science, 27,* 195-225.

Bard, E.A., Anderson, A.H., Sotillo, C., Aylett, M., Doherty-Sneddon, G., & Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language, 42,* 1-22.

Bard, E.G., & Anderson, A.H., (1994). The unintelligibility of speech to children: the effect of referent availability. *Journal of Child Language, 21,* 623-648.

Bard, E.G., & Aylett, M.P. (2004). Referential form, word duration, and modelling the listener in spoken dialogue. In J.C. Trueswell and M.K. Tanenhaus (Eds), *Approaches to Studying World-Situated Language Use: Bridging the Language-as-Product and Language-as-Action Traditions.* Cambridge, MA: MIT Press.

Bavelas, J.B., Coates, L., & Johnson, T. (2000). Addressees as co-narrators. *Journal of Personality and Social Psychology, 79,* 941-952.

Bell, C. (2003). L2 speech rate in monologic and dialogic activities. *Linguagem & Ensino, 6,* 55-79.

Bilous, F.R., & Krauss, R.M. (1988). Dominance and accommodation in the conversational behaviors of same and mixed-gender dyads. *Language and Communication, 8,* 183-194.

Bock, J.K. (1986). Meaning, sound and syntax: Lexical priming in sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 575-586.

Bock, J.K. (1989). Closed-class immanence in sentence production. *Cognition, 31,* 163-186.

Bock, J.K., & Griffin, Z.M. (2000). The persistence of structural priming: Transient activation or implicit learning? *Journal of Experimental Psychology: General, 129,* 177-192.

Bock, J.K., & Loebell, H. (1990). Framing sentences. *Cognition, 35,* 1-39.

Bosch, L. ten, Oostdijk, N., & Boves, L. (2005). On temporal aspects of turn-taking in conversational dialogues. *Speech Communication, 47,* 80-86.

Branigan, H.P., Pickering, M.J., & Cleland, A.A. (2000). Syntactic coordination in dialogue. *Cognition, 75,* B13-B25.

Branigan, H.P., Pickering, M.J., & McLean, J.F. (2005). Priming prepositional-phrase attachment during comprehension. *Journal of Experimental Psychology: Learning, Memory & Cognition, 31,* 468-481.

Branigan, H.P., Pickering, M.J., McLean, J.F,. & Cleland, A.A. (in press). Syntactic alignment and participant role in dialogue. *Cognition.*

Branigan, H.P., Pickering, M.J., Pearson, J., McLean, J.F., Nass, C.I., & Hu, J. (2004). Beliefs about mental states in lexical and syntactic alignment: Evidence from human computer dialogs. *Proceedings of the CUNY Conference on Human Sentence Processing.*

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

Brown, P.M., & Dell, G.S. (1987). Adapting production to comprehension: The explicit mention of instruments. *Cognitive Psychology, 19,* 441-472.

Buhl, H.M. (2001). Partner orientation and speaker's knowledge as conflicting parameters in language production. *Journal of Psycholinguistic research, 30,* 549-567.

Chartrand, T.L., & Lakin, J.L. (2003). Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychological Science, 14,* 334-339.

Chen, K., & Krauss, R.M. (1991). The functionality of backchannel responses in conversation. *Paper presented at the Eastern Psychological Association annual meeting, New York.*

Chiu, C.-y., Hong, Y.-y., & Krauss, R.M. (unpublished manuscript). Gaze direction and speech dysfluency in conversation.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* Cambridge: MIT Press.

Clark, H.H. (2002). Speaking in time. *Speech Communication, 36,* 5-13.

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

Clark, H.H. (1992). *Arenas of language use.* Chicago: University of Chicago Press.

Clark, H.H., & Brennan, S.E. (1991). Grounding in communication. In L. B. Resnick, J. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition,* 127-149. Washington, DC: APA.

Clark, H.H., & Fox Tree, J.E. (2002). Using *uh* and *um* in spontaneous speech. *Cognition, 84,* 73-111.

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

Clark, H.H., & Marshall, C.R. (1978). Reference diaries. In D. L. Waltz (Ed.), *Theoretical issues in natural language processing,* Vol. 2, 57-63. New York: Association for Computing Machinery.

Clark, H.H., & Marshall, C.R. (1981). Definite reference and mutual knowledge. In A. K. Joshi, B. Webber, & I. Sag (Eds.), *Elements of discourse understanding,* 10-63. Cambridge: Cambridge University Press.

Clark, H.H., & Murphy, G.L. (1982). Audience design in meaning and reference. In J.F. L. Ny & W. Kintsch (Eds.), *Language and Comprehension.* New York: North Holland.

Clark, H.H., & Schaefer, E.F. (1989). Contributing to discourse. *Cognitive Science, 13,* 259-294.

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

Clark, H.H., Schreuder, R., & Buttrick, S. (1983). Common ground and the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior, 22,* 245-258.

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

Cleland, A.A., & Pickering, M.J. (2006). Do writing and speaking employ the same syntactic representations? *Journal of Memory and Language, 54,* 185-198.

Cleland, A.A., & Pickering, M.J. (2003). The use of lexical and syntactic information in language production: Evidence from the priming of nounphrase structure. *Journal of Memory and Language, 49,* 214-230.

Danet, B. (1980). Language in the legal process, *Law and Society Review, 14,* 445-564.

Elffers, J. (1976). *Tangram. The ancient Chinese shapes game.* New York: McGraw-Hill.

Fay, N., Garrod, S., & Carletta, J. (2000). Group discussion as interactive dialogue or serial monologue: The influence of group size. *Psychological Science, 11,* 487-492.

Feke, M.S. (2003). Effects of native-language and sex on back-channel behavior. In: L. Sayahi (Ed.), *Selected Proceedings of the First Workshop on Spanish Sociolinguistics,* 96-106. Somerville, MA: Cascadilla Proceedings Project.

Fernández, R., & Ginzburg, J. (2002). Non-sentential utterances: Grammar and dialogue dynamics in corpus annotation. In *Proceedings of the 19th International Conference on Computational Linguistics (CoLing)*, 253-259, Morgan Kaufman Publishers, San Francisco.

Ferreira, V.S., & Dell, G.S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology, 40,* 296–340.

Fodor, J.A., Bever, T.G., & Garrett, M.F. (1974). *Psychology of Language.* New York: McGraw Hill.

Fowler, C., & Housum, J. (1987). Talkers signalling 'new' and 'old' words in speech, and listeners' perception and use of the distinction. *Journal of Memory and Language, 26,* 489-504.

Fowler, C. (1988). Differential attenuation of repeated content words produced in various communicative contexts. *Language and Speech, 28,* 47–56.

Fox Tree, J.E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language, 34,* 709-738.

Fox Tree, J.E. (1999). Listening in on monologues and dialogues. *Discourse Processes, 27,* 35-53.

Fox Tree, J.E., & Schrock, J.C. (2002). Basic meanings of *you know* and *I mean. Journal of Pragmatics, 34,* 727-747.

Fraser, B. (1999). What are Discourse Markers? *Journal of Pragmatics, 31,* 931-952.

Fuller, J. M. (2003). The influence of speaker roles on discourse marker use. *Journal of Pragmatics, 35,* 23-45.

Fussell, S.R., & Krauss, R.M. (1992). Coordination of knowledge in communication: Effects of speakers' assumptions about others' knowledge. *Journal of Personality and Social Psychology, 62,* 378-391.

Garrett, M.F. (1975). The analysis of sentence production. In G.H. Bower (Ed.), *The Psychology of Learning and Motivation,* 133-177. New York, NY: Academic Press.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic coordination. *Cognition, 27,* 181-218.

Garrod, S., & Clark, A. (1993). The development of dialogue coordination skills in schoolchildren. *Language and Cognitive Processes, 8,* 101-126.

Garrod, S., & Doherty, G. (1994). Conversation, coordination and convention: an empirical investigation of how groups establish linguistic conventions. *Cognition, 53,* 181-215.

Garrod, S., & Pickering, M.J. (2004). Why is conversation so easy? *Trends in Cognitive Science, 8,* 8-11.

Giles, H. (1973). Accent mobility: A model and some data. *Anthropological Linguistics, 15,* 87-105.

Ginzburg, J. (1996). Dynamics and the semantics of dialogue. In J. Seligman (Ed.), *Language, Logic and Computation,* Vol. 1, CSLI Lecture Notes, CSLI: Stanford.

Gregory, S.W., & Webster, S. (1996). A non-verbal signal in voices of interview partners effectively predicts communication accommodation and social status. *Journal of Personality and Social Psychology, 70,* 1231-1240.

Grice, H.P. (1975). Logic and conversation. In Peter Cole and Jerry Morgan (Eds.), *Syntax and Semantics: Volume 3, Speech Acts.* New York: Academic Press

Hadelich, K., Branigan, H.P., Pickering, M.J., & Crocker, M. (2004). Alignment in dialogue: Effects of visual versus verbal feedback. *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue, Catalog'04, Barcelona, Spain,* 35-40.

Hartsuiker, R.J., & Kolk, H.H.J. (1998). Syntactic facilitation in agrammatic sentence production. *Brain and Language, 62,* 221-254.

Haywood, S.L., Pickering, M.J., & Branigan, H.P. (2005). Do speakers avoid ambiguities during dialogue? *Psychological Science, 16,* 362-366.

Heinz, B. (2003). Backchannel responses as strategic responses in bilingual speakers' conversations. *Journal of Pragmatics, 35,* 1113-1142.

Hermer-Vazquez, L, Spelke, E.S, & Katsnelson, A.S. (1999). Sources of flexibility in human cognition: Dual-task studies of space and language. *Cognitive Psychology, 39,* 3-36.

Hohlfeld, A., Sangals, J., Sommer, W. (2004). Effects of additional tasks on language perception: An event-related brain potential investigation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 1012-1025.

Horton, W.S., & Gerrig, R.J. (2005). Conversational common ground and memory processes in language production. *Discourse Processes, 40,* 1-35.

Horton, W.S., & Gerrig, R.J. (2002). Speakers' experiences and audience design: Knowing *when* and knowing *how* to adjust utterances to addressees. *Journal of Memory and Language, 47,* 589-606.

Horton, W., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition, 59,* 91-117.

Hupet, M., & Chantraine, Y. (1992). Changes in repeated references: Collaboration or repetition effects? *Journal of Psycholinguistic Research, 21,* 485-496.

Hupet, M., Seron, X., & Chantraine, Y. (1991). The effects of the codability and discriminability of the referents on the collaborative referring procedure. *British Journal of Psychology, 82,* 449-462.

Isaacs, E.A., & Clark, H.H. (1987). References in conversations between experts and novices. *Journal of Experimental Psychology: General, 116,* 26-37.

Johnson-Laird, P.N. (1983). *Mental Models: Toward a cognitive science of language, inference and consciousness.* Cambridge, MA: Harvard University Press.

Jucker, A.H., Smith, S.W. (1998). And people just you know like 'wow': Discourse markers as negotiating strategies. In Jucker, A.H., Ziv, Y. (Eds.), *Discourse Markers: descriptions and theory.* Pragmatics and Beyond Series: 57. John Benjamins Publishing Company, Amsterdam, pp. 171–1201.

Jucker, A.H., Smith, S.W., & Ludge, T. (2003). Interactive aspects of vagueness in conversation. *Journal of Pragmatics, 35,* 1737-1769.

Kent, G.G., Davis, J.D., & Shapiro, D.A. (1978). Resources required in the construction and reconstruction of conversation. *Journal of Personality and Social Psychology, 36,* 13-22.

Keysar, B., Barr, D.J., Balin, J.A., & Brauner, J.S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science, 11,* 32-38.

Keysar, B. & Henly, A.S. (2002). Speakers' overestimation of their effectiveness. *Psychological Science, 13,* 207-212.

Kingma, A., La Heij, W., Fasotti, L., & Eling, P. (1996). Stroop interference and disorders of selective attention. *Neuropsychologica, 34,* 273-281.

Kraljic, T., & Brennan, S.E. (2005). Prosodic disambiguation of syntactic structure: For the speaker or for the addressee? *Cognitive Psychology, 50,* 194–231.

Krauss, R.M., & Fussell, S.R. (1996). Social psychological models of interpersonal communication. In E. T. Higgins & A. Kruglanski (Eds.), *Social psychology: A handbook of basic principles,* 655-701. New York: Guilford.

Krauss, R.M., Garlock, C.M., Bricker, P.D., & McMahon, L.E. (1977). The role of audible and visible backchannel responses in interpersonal communication. *Journal of Personality and Social Psychology, 35,* 523-529.

Krauss, R.M., & Glucksberg, S. (1977). Social and nonsocial speech. *Scientific American, 236,* 100-105.

Krauss, R.M., & Glucksberg, S. (1969). The development of communication: Competence as a function of age. *Child Development, 40,* 256-266.

Krauss, R.M., & Weinheimer, S. (1966). Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology, 4,* 343-346.

Kraut, R.E., Fussell, S.R., & Siegel, J. (2003). Visual information as a conversational resource in collaborative physical tasks. *Human-Computer Interaction, 18,* 13-49.

Kraut, R.E., Lewis, S., & Swezey, L. (1982). Addressee responsiveness and the coordination of conversation. *Journal of Personality and Social Psychology, 43,* 718-731.

Levelt, W.J.M. (1983). Monitoring and self-repair in speech. *Cognition, 14,* 41-104.

Levelt, W.J.M. (1989). *Speaking. From intention to articulation.* Cambridge, MA: MIT Press.

Levelt, W.J.M., & Kelter, S. (1982). Surface-form and memory in question answering. *Cognitive Psychology, 14,* 78-106.

Lickley, R.J. (1998). HCRC Disfluency coding manual. *Technical report HCRC/TR-100.* HCRC, University of Edinburgh, UK.

Lockridge, C.B., & Brennan, S.E. (2002). Addressees' needs influence speakers' early syntactic choices. *Psychonomic Bulletin and Review, 9,* 550-557.

Malt, B.C., & Sloman, S.A. (2004). Beyond conceptual pacts: Enduring influences on lexical choice in conversation. *Memory and Cognition, 32,* 1346-1354.

Markman, A.B., & Makin, V.S. (1998). Referential communication and category acquisition. *Journal of Experimental Psychology: General, 127,* 331-354.

Metzing, C., & Brennan, S.E. (2003). When conceptual pacts are broken: Partner-specific effects in the comprehension of referring expressions. *Journal of Memory and Language, 49,* 201-213.

Oviatt, S.L. (1995). Predicting spoken disfluencies during human-computer interaction. *Computer Speech Language, 9,* 19-36.
Oviatt, S.L., & Cohen, P.R. (1989). The effects of interaction on spoken discourse. *Proceedings of the 27th Annual Meeting of the Association of Computational Linguistics,* 126-134.

Pardo, J.S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America, 119,* 2382-2393.

Pearson, J., Hu, J., Branigan, H.P., Pickering, M.J., & Nass, C.I. (2006). Adaptive language behavior in HCI: How expectations and beliefs about a system affect users' word choice. *Proceedings of CHI 2006,* 1177-1180.

Pickering, M.J. (2005). Feedback and alignment in dialogue. Unpublished manuscript.

Pickering, M.J., & Branigan, H.P. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language, 39,* 633–651.

Pickering, M.J. & Branigan, H.P. (1999). Syntactic priming in language production. *Trends in Cognitive Sciences, 3,* 136-141.

Pickering, M.J., & Garrod, S. (2004). Towards a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27,* 169-190.

Potter, M.C., & Lombardi, L. (1998). Syntactic priming in immediate recall of sentences. *Journal of Memory and Language, 38,* 265-282.

Roβnagel, C. (2000). Cognitive load and perspective-taking: Applying the automatic-controlled distinction to verbal communication. *European Journal of Social Psychology, 30,* 429-445.

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

Schegloff, E.A. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in Society, 29,* 1-63.

Schiffrin, D. (1987). Discourse markers. *Studies in Interactional Sociolinguistics, 5,* Cambridge: Cambridge University Press.

Schober, M.F. (1993). Spatial perspective-taking in conversation. *Cognition, 47,* 1-24.

Schober, M.F., & Brennan, S.E. (2003). Processes of interactive spoken discourse: The role of the partner. In A. C. Graesser, M. A. Gernsbacher, & S. R. Goldman (Eds.), *Handbook of discourse processes,* 123-164. Hillsdale, NJ: Lawrence Erlbaum.

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

Simon, T.J., & Sussman, H.M. (1987). The dual-task paradigm: Speech dominance or manual dominance? *Neuropsychologica, 25,* 559-569.

Smith, S.W., Noda, H.P., Andrews, S., & Jucker, A.H. (2005). Setting the stage: How speakers prepare listeners for the introduction of referents in dialogues and monologues. *Journal of Pragmatics, 37,* 1865-1895.

Smith, M.C., & Wheeldon, L.R. (2001). Syntactic priming in spoken sentence production: An online study. *Cognition, 78,* 123-164.

Stalnaker, R. (1978). Assertion. *Syntax and Semantics, 9,* 315-332.

Waters, G.S., Rochon, E., & Caplan, D. (1998). Task demands and sentence comprehension in patients with dementia of the alzheimer's type. *Brain and Language, 62,* 361-397.

Watson, M.E., Pickering, M.J., & Branigan, H.P. (in press). An empirical investigation into spatial reference-frame taxonomy using dialogue. *Proceedings of the 26th Annual Conference of the Cognitive Science Society,* Vancouver, July 2006.

Wilkes-Gibbs, D., & Clark, H.H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language, 31,* 183-194.

Wheeldon, L.R., & Monsell, S. (1992). The locus of repetition priming of spoken word production. *The Quarterly Journal of Experimental Psychology, 44,* 723-761.

Wood, S., Hiscock, M., & Widrig, M. (2000). Selective attention fails to alter the dichotic listening lag effect: Evidence that the lag effect is preattentional. *Brain and Language, 71,* 373-390.

Yngve, V.H. (1970). On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society.*

Zwaan, R.A., & Radvansky, G.A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123,* 162-185.

Appendix A

Experiment 1: Participant instructions

Instructions – Describer

This experiment will investigate how good people are at building Lego models under different conditions.
Your role is that of the DESCRIBER.
You will be given a number of abstract Lego models, which you will have to describe to your partner (the BUILDER). You must describe each model in enough detail to allow your partner to build an identical model. That means you need to mention the colour, shape and size of each block, and describe how it is to be attached to the other blocks. Try and do this fairly quickly – there is a time limit for each model.
Your partner's model will be marked for correctness, and the best-performing pair of participants will receive a £5 bonus each.

Instructions – Builder

This experiment will investigate how good people are at building Lego models under different conditions.
Your role is that of the BUILDER. Your partner (the DESCRIBER) has a number of Lego models, which he/she will try to describe to you in such a way that you will be able to build the same models with your own Lego pieces. You will only be given the pieces you need to construct each model. Listen to your partner carefully and try to follow his/her instructions exactly. Try and do this fairly quickly – there is a time limit.
For some of the models you will be restricted in what you are allowed to say. The experimenter will tell you what restrictions you are under for which models. If you, by mistake, break the rules of any of the conditions, the experimenter will warn you.
Your models will be marked for completeness and accuracy – the highest-scoring pair of participants will receive an extra £5 each.

Appendix B

Experiment 1: Example transcripts

In the Full feedback and Minimal feedback conditions, the symbol '<>' represents the Builder's feedback; the content of this was not transcribed. The Lego model being described in all three transcripts is pictured below from two perspectives:

Example Transcript: No Feedback

"Right okay start off with a green square block with four circles on and then put a long green one and the long green one has to have six dots on and it has to connect to the bottom green one with the third dot and then get a big yellow block which is two across and six up and connect that to the longer side of the long thin green one but so that it's on the same level as the square green one so you should have two floors so far if you see what I mean, and then get a blue thin one that's one across in width and three along and attach that on top of the yellow block, at right angles to the long thin green one and say one two three four five, to the fifth like circle of the long thin green one, next to it and then get a long thin yellow one that's two across and one two three four five six seven eight up and attach that to the blue thing so that it's on the fourth circle up but so that only one of the circles of the blue one is on it. It should kindof have an H at the moment with the two yellow ones and the square on the first level of the H and then the blue one like being the kindof across bit of the H and the long thin green one just being an extra bit and now you have to get a thin blue one that's one circle across and four circles on it and attach that to the top of the H so that it covers the top two of the circles of the right hand yellow one, but just one of the circles of the left hand yellow one and then when you've done that get a blue block that's three circles across and two circles up and put that beneath the top blue one but so that..." (time limit).

Example Transcript: Minimal Feedback

"Okay put the two yellow ones in front of you on the table, so that they're lined up at the top, the long one on the right, the shorter one on the left, with a little gap in between <> yeah. Then you put the green square underneath the yellow one on the left, so that they're both the same length now, then you get the blue one with four holes, starting at the top right of the long yellow one, you put it horizontally so that it's going over the top two of that one, the gap, and just the first one of the top of the second one <> Just the top right of the left yellow one, yeah? <> So it's going across, and there's a gap of one hole in the middle. Then the blue one with six on you put underneath that, so that it's taking up the next four holes, so it's overhanging by two on the right yellow one, you put it like horizontally again, with the two overhanging on the right <>. And then you get the little blue one and you miss out a line on the right yellow, so it's on not the next one but the next one <>. And then you put it so that one's on each yellow one and one's in the middle, get it <>. Then you get the green one, and you put it far left of the left yellow and green, so that there's three yellow holes above it and then you put it horizontally down the left hand side so it's overhanging by one <>. So you've got three yellow holes going down the far left side, then the green, and the green's overhanging by one, got it? <> That's it".

Example Transcript: Full Feedback

"Okay, take the longest yellow piece and run that from side to side, and then take the blue oblong piece with just four single dots, that goes on the right-hand side, and it goes vertical to the yellow piece. It goes on the very right end, so that you have two dots sticking up, two blue dots free <>. Yep, on top of the yellow. Then beside that, take the blue oblong piece, the fat one, and that goes right beside the blue, but runs down, so there should be two dots free at the bottom. That's on top of the yellow. Then if you're going from right to left <> okay we'll just take that top row of yellow dots, okay, you leave a space of one yellow dot and then the other blue piece goes on top there, running parallel with the first blue piece we did, so it's two dots sticking up, okay? <> Then you take the other yellow piece, and that's gonna run from side to side as well <> Yeah, yes. So you've got the other yellow piece running from side to side, and then that's going to join onto the two blue pieces but it's only one dot, it's not two, so there's a space between the two yellow pieces, and the right end goes with the right end of the blue <> Yeah the one we've just put on, yep, that's right. Then take the long green piece and that makes if I say a T on top of the second blue piece, but it's not a T, there's only one green dot to the right of it, so that there should be four dots to the left, and then this green square goes right underneath, it goes right beside that top yellow piece, so it's underneath the long green piece and right beside the yellow piece. So there should be two green dots sticking out below and then on top one green dot right at the left end free".

Appendix C

Experiment 2: Participant instructions

Instructions – Speaker

This experiment is designed to study how people communicate when they can't see each other. It will involve placing pictures in the correct positions on a grid. You and your partner must work together to try and complete your task as accurately as possible.
You have a folder that shows where each picture should be placed on the grid. Look at the first page in the folder, and tell your partner firstly *which* picture to move (describe it as well as you can, so they can identify it correctly), and then *where* to put it on their grid. Then turn that page over and look at the next picture, describe it and its position and so on. Once all the pictures are on your partner's grid, they will have to move them to different positions. Each picture will be moved more than once. Keep going until you come to a sheet saying 'End'.

Instructions – Listener

This experiment is designed to study how people communicate when they can't see each other. It will involve placing pictures in the correct positions on a grid. You and your partner must work together to try and complete your task as accurately as possible.
You have a grid and a number of picture cards. When your partner describes a picture, try and pick out the matching one from your set, and put it where they tell you to on your grid. Later, when all your cards are on the grid, your partner will tell you to move them to different positions. Your task is to have the correct display of pictures at the end of the experiment.

Appendix D

Experiments 3, 4 and 6: Participant instructions

Welcome!

This experiment is designed to test how good people are at following instructions that were intended for someone else.

When you press the spacebar, you will see 20 black 'tangram' pictures on the screen, with numbers below them. Tangrams are pictures made up of shapes, like squares and triangles, put together to look like characters or animals. When you see the tangram selection, you will hear a description of one of the tangrams. Choose the tangram you think the description refers to, and then SAY the NUMBER of that tangram into the microphone. Even if you're not sure of it, make your best guess.

These descriptions were recorded in an earlier experiment, where people had to place pictures in a particular orientation, but just ignore any references given (like 'the top left one is') - that's not relevant to you, and has no bearing on the position of tangrams on your screen.

Press SPACE to move onto the next screen and hear the next description. Choose as accurately as possible, but try to do it quite quickly too! The screen will time out after a set time.

The experimenter will now answer any questions you have, and make sure you're clear about what you're doing before you start. The next screen you see will be example tangrams - you will have a while to familiarise yourself with them before the first description begins.

Press SPACE to start.

(Note: In Experiment 6, there were 24 tangrams on each screen, not 20; the instructions reflected this change).

Appendix E

Experiment 5: Pilot test

Similarity Questionnaire

Below you will see a number of pairs of black 'tangram' pictures. Please look at these, and decide how similar you think each pair is, rating them on a 5 point scale, like this:

1 2 3 4 5

Very similar Very different

For example, the pair below:

...would probably be judged as 'very different'. Ring the number on the form that you think represents their similarity.

1 2 3 4 5

Very similar Very different

Do not worry about spending too long on this – just go with your instincts. You can use characteristics such as the direction they're facing, whether they are standing up, sitting down, lying down etc, to help you decide. Thanks!

Appendix F

Experiment 5: Participant instructions

Instructions – Describer (dialogue)

Turn over the cards on the desk in front of you, being careful to keep them in the same order. Your task is to describe each of these cards in turn to your partner. They will try and select the matching cards from their (much larger) set, and put them in the same order as yours. Your partner can talk to you freely. There is no time restriction, but try and complete the task as efficiently as possible.

Instructions – Matcher (dialogue)

Lay out the cards in the pile in front of you so that you can see them all clearly. Your partner has an array of 12 cards in a particular order, and they will describe each of them in turn to you. You must find the cards they are describing, and put them in the same order as your partner's. Be careful that you look at all the possibilities, because sometimes there will be two or more cards that look similar, and you must be careful to choose the right one. You can ask your partner as many questions as you like. There is no time restriction, but try and complete the task as efficiently as possible.

Instructions – Describer (restricted feedback)

Turn over the cards on the desk in front of you, being careful to keep them in the same order. Your task is to describe each of these cards in turn to your partner. They will try and select the matching cards from their (much larger) set, and put them in the same order as yours. Your partner is not allowed to speak, except to say 'finished' when they believe they have placed the correct card. There is no time restriction, but try and complete the task as efficiently as possible.

Instructions – Matcher (restricted feedback)

Lay out the cards in the pile in front of you so that you can see them all clearly. Your partner has an array of 12 cards in a particular order, and they will describe each of them in turn to you. You must find the cards they are describing, and put them in the same order as your partner's. Be careful that you look at all the possibilities, because sometimes there will be two or more cards that look similar, and you must be careful to choose the right one. You are not allowed to speak to your partner, except to say 'finished' when you believe you have placed a card correctly. There is no time restriction, but try and complete the task as efficiently as possible.

Instructions – Describer (monologue)

Turn over the cards on the desk in front of you, being careful to keep them in the same order. Your task is to describe each of these cards in turn to your partner. They will try and select the matching cards from their (much larger) set, and put them in the same order as yours. Your partner is **not allowed** to speak to you at all. There is no time restriction, but try and complete the task as efficiently as possible.

Instructions – Matcher (monologue)

Lay out the cards in the pile in front of you so that you can see them all clearly. Your partner has an array of 12 cards in a particular order, and they will describe each of them in turn to you. You must find the cards they are describing, and put them in the same order as your partner's. Be careful that you look at **all** the possibilities, because sometimes there will be two or more cards that look similar, and you must be careful to choose the right one. You are **not allowed** to speak to your partner at all. There is no time restriction, but try and complete the task as efficiently as possible.