DISFLUENCY IN DIALOGUE:
ATTENTION, STRUCTURE AND FUNCTION

Hannele Buffy Marie Nicholson

PhD Thesis

The University of Edinburgh

2007

“My work is a game. A very serious game.”
-M.C. Escher

“You can know the name of a bird in all the languages of the world, but when you’re finished, you’ll know absolutely nothing whatever about the bird…So, let’s look at the bird and see what it’s doing – that’s what counts. I learned very early the difference between knowing the name of something and knowing something.”
-Richard Feynman

“Do not fail
To learn from
The pure voice of an
Ever-flowing mountain stream
Splashing over the rocks
-Morihei Ueshiba,
The Art of Peace

CONTENTS

FIGURES

TABLES

DECLARATION

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

Hannele Nicholson

ABSTRACT

Spontaneous speech is replete with disfluencies: pauses, hesitations, restarts, and less than ideal deliveries of information. Disfluency is a topic of interdisciplinary research with insights from psycholinguistics, phonetics and speech technology. Researchers have tried to determine: *When does disfluency occur?*, *Can disfluency be reliably predicted to occur?,* and ultimately, *Why does disfluency occur?* The focus of my thesis will be to address the question of why disfluency occurs by reporting the results of analyses of disfluency frequency and the relationship between disfluency and eye gaze in a collaborative dialogue.

Psycholinguistic studies of disfluency and collaborative dialogue differ on their answers to why disfluency occurs and its role in dialogue. One hypothesis, which I will refer to as Strategic Modelling, suggests that disfluencies are designed by the speaker. According to the alternative view, which I will call the Cognitive Burden View, disfluency is the result of an overburdened language production system. Throughout this thesis, I will contrast these two theories for an ultimate answer to why disfluency occurs. Each hypothesis attaches a functional role to a structural definition of disfluency and therefore in order to determine why disfluency occurs, I will contrast the structural and functional characteristics of disfluency. I will attempt to do this by analysing the dialogue behaviour in terms of speech goals and eye gaze behaviour a speaker is engaged in when they make certain types of disfluencies.

A multi-modal Map Task paradigm was used in this thesis, in which speakers were asked to describe the route on a cartoon map to a distant confederate listener who provided either visual or verbal feedback. Speakers were eye-tracked during the dialogue and a record was kept of when the speaker attended to the listener's visual feedback. Experiment 1 tested the visual feedback paradigm to establish its validity as a baseline condition. Speakers were found to make more disfluencies when they could interact with the visual feedback, suggesting disfluency is more common in interactive circumstances. Experiment 2 added verbal feedback to the experimental paradigm to test whether listeners react differently to the two modalities of feedback. Speakers made more disfluencies when the feedback was more complicated. Structural disfluency types were also observed to fulfil different functions. Finally, Experiment 3 manipulated the motivation of the speaker and found that Motivated speakers gazed more often and were more disfluent per opportunity than Control speakers suggesting that highly motivated subjects are more willing to engage in difficult tasks.

ACKNOWLEDGEMENTS

This thesis was funded by the generous support of EPSRC as part of the MONITOR Project co-organized by Dr. Ellen Gurman Bard and Dr. Anne H. Anderson. I would like to thank my supervisors Dr. Ellen Gurman Bard and Dr. Robin Lickley for continuous advice, support, training, and inspiration so that this document is not merely a '*banalysis'* (I hope). I would also like to thank Dr. Anne H. Anderson who as a collaborator on the MONITOR Project has informally fulfilled the role of a third supervisor. In terms of assistance with the running of experiments and data collection, I would like to thank the Research Associates from the MONITOR Project (in alphabetical order): Yiya Chen, Sara Dalzel-Job, Maria Luisa Flecha-Garcia, Alex Fulton, Catriona Havard, David Kenicer, and Lucy Smallwood. For providing additional disfluency coding with which I could do a reliability study, I would like to thank Lucy MacGregor, Michael Schnadt, and Ryan Gramacy. For conducting an analysis of Krippendorff's alpha with my data, I would like to sincerely thank Dr. Klaus Krippendorff. For assistance with XML, I am indebted to Dr. Jean Carletta and those involved with developing the NITE XML Toolkit. For writing programmes to turn XLabel disfluency coding into XML, I would like to thank Ruli Manurung. For special purpose programming, I would like to thank Cedric MacMartin and Joseph Eddy. I would also like to thank the support staff of Linguistics and English Language, Cedric MacMartin, Morag Brown, Michael Bennett, and Eddie Dubourg, who assisted with making sure that computers ran smoothly, lost files were retrieved, printers were unjammed, Microsoft Word © behaved (for the most part) and for generally making the LEL department a pleasantly memorable place to work. I would like to thank Prof. James Hurford and Prof. April McMahon for advice and support. I also wish to thank Dr. Catherine Mayo for advice throughout the thesis, especially during the darkest hours. Thanks to Dan Wedgwood for most interesting and extremely relevant discussions about intention and meaning. Without the moral support of Yiya Chen, Susana Cortés-Pomocondor, Michael Durrant, Francesca Filiaci, Marisa Flecha-García, Christine Haunz, Ruth Hanson, Annabel Harrison, Lynda Hutchison, Simon Kirby, Timothy and Deena Pandachuck-Mills, Andreas Hilmo Teig, Dan Wedgwood, Lukas Wiget, Ivan Yuen and The Common Roomers in general, I would have finished much worse for the wear and no doubt suffered severe biscuit withdrawal. Likewise, I am also indebted to my Aikido sensei and partners, Neil Blacknell and his students at Azami-kai dojo, for teaching me about coordination, harmony, and discipline and for throwing me forcefully on the mat, relieving any stress instantaneously. Thank You to Babycakes, David and Douglas von Wauchope, Mr. Wheat and Dr. Gaywood for exceedingly splendiferous flatmateship: *Get In!* Thanks to Joseph Eddy for constant moral support, constant laughter, constant willingness to cook and clean when I couldn't, and constant love. Thank you to my sister, Marija, for showing her support, love, humour and for keeping me grounded. Finally, I would like to thank my parents, Sandra Sevre and Darrel Nicholson, for inspiring my love of learning, my interest in disfluent speech and conversation, and for reminding me when it is necessary to have '*SISU*'. *Haakka Päälle!*

CHAPTER 1 – Introduction

Why write a thesis about disfluency? Disfluency is notoriously more common in spontaneous speech than in written text. Take for example an exchange in a court room where the speaker said '*Yes, we were there….I mean we didn't leave the um the place ehm Lake Street Cafe <pause> until about um t- ten-… eh eleven o'clock'*. In this case, the court reporter would have transcribed '*We didn't leave the Lake Street Café until about eleven o'clock'* without all of the *'ums'* and *'uhs'* and restarted phrases. Since spoken speech differs in this manner from written text, the study of disfluency offers a potential insight into human language production and human behaviour and by studying it empirically one can tap into these insights. Since disfluency is one output of language production, it is important to review some of the psycholinguistic models of language production and collaborative dialogue in order to understand disfluency in relation to other features of dialogue. In this chapter, I will outline two theories of collaborative dialogue and introduce the issues considered in this thesis. According to one view, the Cognitive Burden View, disfluency is an output error of an overburdened cognitive system. Alternatively another hypothesis, which I will call the Strategic-Modelling View, argues that disfluency is a signal of delay and commitment to a listener.

It is important to note that disfluency differs from speech errors, or 'slips of the tongue', and from stuttering. A person is considered to have made a speech error when 'the actual utterance differs from the intended utterance' (e.g. *'White Anglo-Saxon prostitute'* instead of '*White Anglo-Saxon Protestant'*) (Wells-Jensen, 1999). Speech errors can be delivered in an entirely fluent manner with no disruptions and not be considered disfluent. Most of the disfluencies considered in this thesis tend to involve correction of some sort or another (e.g. '*yeah, if you just continue down to the left…ehm right, sorry'*) but disfluencies need not always be corrected or detected. All disfluencies are in some way a disruption in otherwise fluent speech. I will not discuss stuttering in this thesis.

As far as disfluencies are concerned, disfluencies are thought by some to express a strategic signal to the listener (Clark, 1996; Clark & Wasow, 1998; Fox Tree & Clark, 1997). It is the purpose of this thesis to investigate this question: '*Why do disfluencies occur?'* In order to do this, I will contrast the strategic signalling proposal with another hypothesis which suggests that disfluencies do not fulfil a signalling function and only occur because it the easiest thing for the speaker to do at that point in time or simply as an error (Bard, Lickley, & Aylett, 2001; Pickering & Garrod, 2004). I will first explain the labelling systems used to describe disfluency and the two models of dialogue which offer potential answers for *why* disfluency might occur.

1.1 Disfluency Description

What is the structure of a disfluency? Here I will address this question by explaining the labelling systems used to describe disfluency.

Spontaneous speech is notoriously unlike written text in part because it frequently includes a variety of extended pauses, filled pauses (e.g. '*um', 'er', 'eh'*), cut off, and repaired utterances. In order to model these disfluency phenomena consistently, several categorisation systems have been developed to label the various components which comprise a speech repair (e.g. Levelt, 1983; Nakatani and Hirschberg, 1994; Shriberg, 1994; Lickley, 1994). For the most part, these labelling systems are interchangeable but the occasional discrepancy does exist. It is the purpose of this section to examine these differences and determine the standard referred to henceforth in this dissertation. The most widely used scheme comes from Levelt (1983).

OU (original utterance)	editing phase	R (repair)
Go from left again to	**uh ….**	**From pink again to blue**

The structure above is useful because it allows reference to particular disfluent regions of speech. The **OU** (original utterance) designates all the speech prior to the Interruption point including the **reparandum**, or portion of the utterance to be repaired. Between the OU and the Interruption point there is a delay period that may range over any number of words. The **editing phase** may contain a filled pause, as it does in the example above or any **editing terms** (*sorry, or, I mean*) or nothing. Following the editing phase is the **Repair** which contains the **alteration**, or speech which is meant to replace the reparandum. Optionally the Repair may also contain **retracing**, or repeated words (eg. *from* in the example sentence above) that occur between the Interruption point and the Repair.

Subsequent study by Nakatani and Hirschberg (1994) also defined the structure of disfluency by decomposition into three intervals. First, the reparandum interval corresponds to Levelt's OU and contains all the 'flawed' speech that is replaced by the Repair. The disfluency interval corresponds to Levelt's 'editing phase' or region of filled pauses, silences and overt markers of correction. The repair interval corresponds to Levelt's Repair and spans from the resumption of speech to the end of the material replacing the reparandum.

Since Levelt (1983) a number of authors have adopted the terms 'reparandum' and 'repair' (Blackmer & Mitton, 1991; Lickley, 1994; Savova & Bachenko, 2002; Shriberg, 1994).

However, there is some disagreement about how to refer to the intervening speech between Reparandum and Repair (i.e. Levelt's 'editing phase'). Shriberg (1994) coins the term 'interregnum' to refer to Levelt's 'editing phase' while Blackmer and Mitton (1991) use the term 'cutoff-to-repair' to refer to this region. Shriberg (1994) states her reason for coining the term is one of maintaining an atheoretical position with respect to the function of disfluency for the speaker. The term interregnum neutrally refers to the period of speech in between repair and reparandum without necessarily ascribing an intentional editing state to the speaker as is implied by Levelt's 'editing phase'.

According to Shriberg (1994), a speech repair can be segmented into a *reparandum* (eg. the portion of speech to be repaired), an *Interruption point* (*IP)*, an *Interregnum* (IR), and a *repair*.

A vertical **IP** uh a horizontal line

Reparandum ← IR------→ Repair

Figure 1. An example of substitution

In Figure 1 above, the reparandum ('a vertical') is interrupted at the Interruption Point (IP), thus beginning the Interregnum stage (IR). The Interregnum in the above example contains a filled pause 'uh' and a silent pause of unspecified length. Immediately following the interregnum, the repair ('a horizontal') begins and the utterance continues with 'line'. It should be noted that the term 'Interregnum' is consistent with the 'disfluency interval' in Nakatani and Hirschberg (1994).

In Figure 1, the speaker began by describing a line as vertical, but then altered his description to 'horizontal'. The term horizontal was substituted for the term vertical and so one could classify such a repair as a substitution. If the speaker had said:

It's.... **IP** it's a bit down from the dead tree

Reparandum < IR---> Repair

Figure 2. A repetition disfluency

where the pronoun + copula combination is repeated in two tokens before the utterance continues, one could label the repair a *repetition* as distinct from a *substitution***.** Clearly, since a speaker can repair an error in a number of ways, some sort of labelling schema is required to distinguish disfluency form by categories which operate consistently across all speakers and all potential speech repairs. Two such approaches will be discussed in Chapter 2, namely Levelt's (1983) cognitive theory of repair as devised on a corpus of Dutch speech and Lickley's (1998) speech repair classification system as employed in the HCRC Map Task corpus (Anderson et al., 1991; Lickley, 1998).

1.2 Models of Dialogue

Disfluencies occur frequently in spontaneous conversations between individuals everyday. Sometimes the speaker will be aware that they have been disfluent and they will rephrase the disfluent utterance, most often they will reprhase it immediately after making it (Nooteboom, 1980). How does the speaker recognize that s/he has made a disfluency in the first place? According to Levelt, (1983, 1989), the language production system of a speaker is equipped with an internal monitor loop which allows the speaker to monitor their own speech and detect disfluencies in the output. The speaker then amends the disfluencies and the dialogue continues normally.

According to the *Principle of Optimal Design***,** speakers monitor their listeners during a dialogue also and formulate utterances for the listener (Clark, Schreuder, & Buttrick, 1983). In other words, speakers must coordinate with the listener in order for a successful dialogue. As any one who has ever taken dancing lessons might know, coordination with another person requires some skill. For dialogue, Clark (2002) suggests that speakers use a variety of signalling devices to indicate their actions. One of these signalling devices is disfluency (Clark, 2002; Clark & Wasow, 1998; Fox Tree & Clark, 1997). If the speaker encounters difficulty during language formulation, Clark and colleagues suggest that speakers use disfluencies as a 'collateral signal' to indicate to the listener when s/he expects to be ready to utter the next portion of speech. For example, if a speaker said *'You want to turn ri- ...eh left at the corner'*, the fragment 'ri-' (presumably *'right'*), the short pause and the '*eh*' would all be signals that the speaker intended to halt speech, delay for a short while and eventually resume speaking (Clark, 2002; Clark & Fox Tree, 2002; Fox Tree & Clark, 1997).

Why do speakers go to all this trouble to signal their intentions to their listener? According to Clark and colleagues, when a speaker engages in conversation, the speaker strives for the ideal delivery (Clark, 2002; Clark & Wasow, 1998). Ideal delivery requires firstly, that the speaker engages the listener's attention at just the right moment and secondly, that the speaker's utterance is well-formed. Participants assume a sort of joint responsibility in designing utterances that are optimal for the current circumstances. Pursuing the ideal delivery requires that the speaker and the listener are synchronised. A speaker will generally begin an utterance once they know that the listener is looking at him/her (Goodwin, 1981). Once they have the listener's attention, the speaker will then try to speak in a fluent manner, with a model of the listener in mind. If an error should occur, the speaker will still attempt to speak in a continuous manner by retracing an utterance from the point at which they left off (e.g. '*If you have a-..., If you have a green car'*) (Clark & Wasow, 1998).

The Principle of Optimal Design is based on the theory that during collaborative dialogue, interlocutors attempt to develop 'common ground' with each other (Brennan & Clark, 1996; Clark, 1996; Clark & Carlson, 1982a, 1982b; Clark & Marshall, 1981; Clark et al., 1983; Pickering & Garrod, 2004). Common ground refers to the knowledge, beliefs and assumptions that two interlocutors might share. Interlocutors determine what constitutes common ground with the aid of three types of information: community membership, linguistic evidence and perceptual evidence from their immediate surroundings (Clark & Marshall, 1981). By referring to the common ground between them, interlocutors can work out what constitutes mutual knowledge.

The theory of Mutual Knowledge is a much discussed topic in a wide range of literature (Austin, 1962; Grice, 1957, 1968, 1989; Johnson-Laird, 1982a, 1983; Smith, 1982; Sperber & Wilson, 1987, 1995). The term 'mutual knowledge' refers to the fact that in order for something to be fully mutually known by another person, that person must also recognize that the speaker intended for the person to know this (Grice, 1957, 1968, 1989). As pointed out by many, attaining full mutual knowledge require an infinite number of recursive steps, which presents problems when considering the rapid nature of dialogue (Clark & Carlson, 1982a, 1982b; Johnson-Laird, 1982b, 1983; Pickering & Garrod, 2004; Smith, 1982; Sperber & Wilson, 1987, 1995). For example, in a conversation between two people discussing a dress, when the speaker refers to 'the dress' she has made some assumptions that the listener knows which particular dress is being discussed, and further more the listener must know that the speaker knows which dress they are discussing and so on (Schober & Brennan, 2003).

Since true mutual knowledge is difficult to obtain, a number of other researchers have suggested that perhaps speakers do not require full mutual knowledge in order to sustain a successful conversation. According to the *Principle of Least Collaborative Effort*, as proposed by Clark and Wilkes-Gibbs (1986), individuals involved in a dialogue have a joint responsibility to make sure that any contribution to the conversation has been mutually understood by the other participant. If Angelina says to Bryce, 'Which dress should I wear to the party, the blue or the green one?' If while deciding Bryce realises that Angelina actually owns *two* green dresses, then it is his responsibility to clarify by asking something like 'Which green dress?' According to this view, it is not solely the speaker's responsibility to ensure that the listener has fully understood his contribution. Rather, speakers and listeners share in the responsibility.

Clark and colleagues have suggested that speakers will attempt to model their listener's perspective when they can. There are others who have argued that modelling the other listener is a cognitively costly and demanding task, given the real-time nature of dialogue and the processing demands on a speaker during dialogue (Barr & Keysar, 2002; Horton & Gerrig, 2005; Horton & Keysar, 1996). I will refer to this view as the Cognitive Burden View in this thesis. According to this view, the speaker does not need to rely on a model of the speaker because s/he can instead rely on his/her own perspective of the conversation to formulate an utterance.

According to the Cognitive Burden View, a disfluency is considered to be an unintentional sign of cognitive difficulty on the part of the speaker (Bard et al., 2001). This differs quite noticeably from the view proposed by Clark and colleagues that disfluency is a strategic signal. Clark and colleagues suggest that disfluency occurs while the speaker is encountering difficulty (Clark & Wasow, 1998), but elsewhere in the literature Clark (e.g. Clark, 1996, 2002) does suggest quite clearly that disfluencies "are genuine signals – collateral signals – that speakers design and produce with skill" (Clark, 2002, p. 13). As it stands then, there seem to be two answers to the question *Why does disfluency happen?* The first suggests that disfluency is not under the volitional control of the speaker, but is merely the error of an overburdened system (Bard et al., 2001). The second theory suggests, according to the Principle of Optimal Design, that disfluencies are strategic signals and speakers design them as careful solutions to problems in dialogue (Clark, 2002).

What is the nature of the evidence to support the Optimal Design and the Cognitive Burden views? There is support from the philosophy of language (Austin, 1962; Grice, 1957, 1968, 1989; Sperber & Wilson, 1995), conversational analysis (Schegloff, 1996; Schegloff, Sacks, & Jefferson, 1977) and some support from psycholinguistics (Clark, 1996; Haywood, 2004; Haywood, Pickering, & Branigan, 2005; Horton & Gerrig, 2005; Horton & Keysar, 1996) for both theories, although traditional psycholinguistic models of language tend to avoid research on dialogue (Haywood et al., 2005; Pickering & Garrod, 2004). The evidence within these fields, with the exception of psycholinguistics, has tended to be descriptive in nature. As far as studies of disfluencies and dialogue are concerned, researchers have conducted corpus studies to discover how disfluencies occur in natural dialogue and then describe their occurrence (e.g. Clark, 2002; Clark & Wasow, 1998; Fox Tree & Clark, 1997). For the most part, corpora are a very valuable and enlightening tool. There is, however, a need for experimental studies since they are an online test of the speaker's ability (Schober & Brennan, 2003). Task-oriented experiments provide the experimenter with easier access to the speaker's intentions since the task helps to constrain the possible range of intentions (Brennan, 2004; Schober & Brennan, 2003). Therefore, I will argue that there is a need for task-oriented experiments that manipulate difficulty and speaker attention in dialogue to determine whether speakers really attend to their listeners fully and signal their intentions through disfluency. Conducting such an investigation is my primary goal in this thesis.

1.3 Investigations in this Thesis

The lack of literature on disfluency and difficulty in dialogue is the main motivation to report the results in this thesis. As explained in Section 1.2, theories of dialogue tend to differ in terms of whether speakers model their listeners and therefore use disfluencies as collateral signals, or whether listener modelling is cognitively taxing and unnecessary, and therefore disfluencies are merely an output error of an overburdened system. In order to determine whether speakers use disfluency as signals, I will test the speaker's cognitive load during a dialogue task and investigate their disfluency patterns in conjunction with their gaze patterns at visual feedback from a listener, their partner in the task. Previous experimental paradigms to test shared knowledge have used tasks that are not tricky enough to simulate a real-world task (i.e. tangrams[1], simple naming of objects in a grid). For this reason, I will use the Map Task (Anderson et al., 1991; Brown, Anderson, Yule, & Shillcock, 1983) as my experimental paradigm: it allows speakers to have a quasi-spontaneous and natural dialogue, yet it is more complex in nature than simple shape description so one might actually expect speakers to encounter cognitive difficulty.

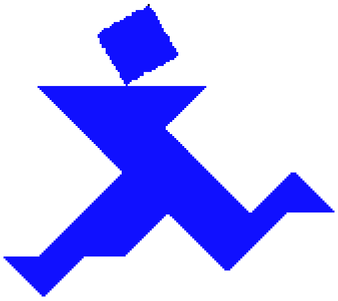


Figure 3. An example of a tangram shape

8/5/078/5/07

[1] An example of a tangram is shown in Figure 3. In experiments using tangram tasks, participants are often asked to describe what the shape looks like to a partner

In the Map Task corpus, originally developed by (Brown et al., 1983), participants were asked to reproduce the route from one participant's map onto the map of the other participants. For the HCRC Map Task Corpus (Anderson et al., 1991), one participant was designated as the 'Instruction Giver' and the other was assigned the role of 'Instruction Follower'. The Instruction Giver was given a map with cartoon landmarks (labelled with names) and a pre-drawn route. It was the job of the Giver to describe this route to the Instruction Follower, who could only see a similar map that did not have a pre-drawn route on it. Givers and Followers saw similar maps which shared some landmarks but differed for others: some landmarks were present on the Follower's map that were not present on the Giver's, some landmarks were labelled with different landmark names but were in the same location, some landmarks occurred twice along the route on the Giver's map but only once on the Follower's map and finally, some landmarks had a contrastive pronunciation feature (i.e. '*Green Bay'* vs. '*Crane Bay'*). In addition to landmark accessibility, Anderson et al. controlled for the familiarity of participants (i.e. participants were either friends or had never met) and the ability to make eye contact (i.e. eye contact versus none). The advantages of analysing a map task experiment are that such a corpus provides spontaneous speech and task-oriented dialogue. As mentioned previously, Schober and Brennan (2003) suggest that a task-oriented dialogue constrains the number of possible intentions that the speaker could have entertained and thus makes it more amenable for determining whether speakers are using disfluencies as intentional signals or out of difficulty, as predicted by the Strategic-Modelling and Cognitive Burden Views respectively. The Cognitive Burden View predicts disfluency will arise when the speaker is under cognitive load and therefore in order to test difficulty, a task that is suitably difficult is required. The map task is perfect for this type of experiment because it requires that speakers guide listeners around a route that they have not seen. Furthermore, since their maps are not perfectly matched, difficult periods of misunderstanding are almost guaranteed. For these reasons, I will report the results of the MONITOR Project, described in further detail in the next section, in this thesis.

Before any analysis can be done to address why disfluency occurs, we need an understanding of what a disfluency is and the classification systems developed for disfluency. Chapter 2, the literature review, begins by differentiating disfluencies from speech errors, explaining disfluency classification systems so that the reader can understand the perspectives in the field. I then discuss the issue of disfluency terminology and the fact that there appears to be some terminological confusion in the field. Next, I introduce fully the hypotheses of collaborative dialogue tested in this thesis and their predictions for why disfluency occurs. Included in these sections is a review of the literature from the fields of Speech Technology and Phonetics, perception of disfluency, and intentionality to understand when disfluency might occur, when listeners can perceive it and what is meant by an intentional signal. Finally, Chapter 2 reviews the literature on the role of eye gaze in dialogue. Chapter 3, the first experimental chapter, is focussed on establishing a baseline experimental paradigm of visual feedback in a Map Task dialogue. Chapter 3 also begins to address the questions of when, where and why disfluency occurs. Chapter 4 tests both the baseline visual feedback paradigm with the addition of verbal feedback in order to discern whether one type of feedback has more of an impact on the speaker. Chapter 4 further investigates the when, where and why of disfluency in addition to how a speaker copes with additional cognitive load. Finally, Chapter 5 asks whether the speaker's behaviour can be changed if a speaker is offered additional motivation. In this way, Chapter 5 is a true test of speaker commitment because one might predict that the speaker who is more committed to producing an ideal delivery and to helping their listener, would signal this fact by signalling more often with disfluencies.

1.4 MONITOR Project

This thesis was not written in a vacuum and the experiments reported in it were by no means of my own creation. As previously mentioned in the Acknowledgements, I received financial support from the EPSRC in order to pursue my PhD. This support was part of the MONITOR Project, a collaborative EPSRC grant held by Dr. Anne H. Anderson at the University of Glasgow and Dr. Ellen Gurman Bard at the University of Edinburgh. During the course of the project, a number of Research Associates and Programmers have run experiments, developed XML tools, transcribed speech, coded eye-gaze data, analysed data and written special-purpose computer programmes. Table 1 below shows which analyses and work were conducted by other individuals and which were conducted by the author. I also benefited from MONITOR Project meetings, mainly with Dr. Anne H. Anderson, Mr. David Kenicer, Dr. Marisa Flechá-Garcia, Dr. Yiya Chen, Ms. Catriona Havard, Ms. Sara Dalzel-Job and Mr. Jim Mullin. Under the auspices of the MONITOR project, I have published previous papers about disfluency and eye-gaze: Nicholson et al. (2003) and Nicholson et al. (2005). Copies of these papers can be found in Appendix A.

Table 1. Distribution of work on the MONITOR Project

	Experiment 1	Experiment 2A & 2B	Experiment 3
Experimenter	Mr. David Kenicer	David Kenicer and Catriona Havard	Alex Fultion and Hannele Nicholson
Eye-tracking paradigm	Mr. Jim Mullin	Mr. Jim Mullin	Mr. Jim Mullin
Gaze Coding	David Kenicer and Catriona Havard	Catriona Havard	Alex Fulton
Gaze Analysis	David Kenicer	Catriona Havard, Alex Fulton and Sara Dalzel-Job	Alex Fulton and Sara Dalzel-Job
Transcription and Speech Coding	Dr. Maria Flechá-Garcia and trained coders	Dr. Yiya Chen and trained coders	Hannele Nicholson, Gabriel Murray and Ken Thomson
Speech Analysis	Dr. Maria Flechá-Garcia and Dr. Yiya Chen	Dr. Yiya Chen	Hannele Nicholson and Sara Dalzel-Job
Disfluency Coding	Hannele Nicholson (with training from Dr. Robin Lickley)	Hannele Nicholson	Hannele Nicholson
Disfluency Analysis	Hannele Nicholson	Hannele Nicholson	Hannele Nicholson
XML Assistance	Dr. Jean Carletta, Dr. Henry Thompson and Dr. Ruli Manurung		
Programming	Joseph Eddy	Joseph Eddy	Joseph Eddy

Note: Dr. Anne H. Anderson and Dr. Ellen Gurman Bard oversaw all elements of the project.

CHAPTER 2 - Literature Review

In this chapter, I review the literature which bears on what constitutes a disfluency, where disfluencies occur and why they occur during dialogue. To address the question of what can be considered a disfluency, I review recent disfluency models within the literature. To answer where and when disfluency occurs, I turn to the field of speech recognition and automatic detection of disfluencies. Answering the question of why disfluency occurs is not straight-forward, and for this reason I review models of collaborative dialogue, speech production, intentionality in speech and models of self repair. I also outline the current experimental research on gaze during dialogue with emphasis on what has been discovered since the advent of eye-tracking in order to motivate the need for further experiments investigating disfluency and eye gaze during collaborative dialogue.

2.1 Disfluency Classification Systems

2.1.1 Cognitive models of Speech Production and Self-Repair

In order to understand why, when and where speech errors or disfluencies occur, we need an understanding of how the speaker is thought to detect and correct mistakes in his or her own speech during dialogue. This question will be the focus of this section. Three major proposals have been put forth within the literature. Laver (1980) defends an account of error detection in speech production that incorporates error detection on a neuromuscular level. Levelt (1983; 1989) proposes that error detection occurs via an auditory-feedback loop and internal monitor. Finally, MacKay (1987) outlines the node structure theory, a connectionist model in which node activation leads to speech production and possibly to disfluencies. In this section, I will briefly summarise and compare the three accounts. For a more extensive review, refer to Postma (2000).

Laver (1980) argues for a distributed editing theory of error detection which employs propositional logic and feed-forward links at various stages throughout the production process to detect errors. Before speech production can occur, the message must proceed from the Ideation phase on to an abstract phase of linguistic programming, on to an abstract phase of motor programming and from there to the conversion of abstract planning into neuromuscular commands. Articulation occurs at this stage, after which there is a period of post-articulatory monitoring. Errors may only be detected after post-articulatory monitoring and only then it is possible to loop back for correction. Laver (1980) shows that it is possible to assign errors of different kind to different phases of this system. For example, if a speaker were to say ‘*Ra**l**e**b**ais*’ when they meant ‘Rabelais’, (error recorded by (Fromkin, 1971)). This error could be attributed to a malfunction in the motor programming section because there has been an error in the serial ordering of the abstract motor program such that segments [l] and [b] were exchanged (Laver, 1980 p. 297). Not all errors involving whole segments arise in the motor programming section, however. A spoonerism, or phrase involving an exchange of two sounds, like *‘a **k**ice **r**eam cone’* (Fromkin, 1971) that involves segmental exchange is more likely to be formed during the linguistic programming phase because the exchanged segments cross morphophonemic boundaries (Laver, 1980). On the other hand, a ‘linguistically unorthodox’ error like *‘he behaved as like a fool’*, a blend of ‘*like a fool’* and ‘*as if/though he were a fool’*, crop up during the linguistic planning phase and are verified later through a postutterance monitoring function (Laver, 1980).

Although the production-based approach of Laver (1980) employs an external mechanism in order to detect errors, connectionist theories traditionally utilise only entities within their own system. This is the case for node structure theory of MacKay (1987) in which language processing and language comprehension are brought about via the same structure of hierarchical layers of interleaving nodes. Nodes may either be shared or specific to production and perception. Mackay's node layers include but are not limited to ‘propositional nodes’, ‘muscle-movement nodes’, ‘phonological nodes’, ‘syllable nodes’ and ‘sensory-analysis nodes’. Any of these node types may be primed in either the output or input direction. The node with the most priming is the one to be activated for inclusion in a particular phase of either perception or production.

According to node structure theory, an error may be detected through backward-priming (Mackay, 1987). For example, suppose that a particular node is activated by mistake. This activated node would submit a signal to the next conceptual node in the network, thus activating it and creating perceptual awareness of the flaw. Corrective action may then commence, though MacKay does not provide the specifics of this process. The node activation system in node structure theory is developed especially to capture MacKay's belief that the perception of one's own speech errors differs from the perception of other's speech errors. MacKay compared Nooteboom's (1980) 75 percent self-correction rate for phonological errors versus Tent and Clark's (1980) much lower rate of phonological correction of other's speech errors.

According to Mackay, should a node be activated wrongly, an error will occur and the speaker will be aware of it. A central tenet of the node structure theory is that once a node has received enough activation to be uttered, the speaker is aware of this activation (Mackay, 1987). This claim is contested by anecdotal evidence provided by Laver (1980) and experimental evidence from Postma & Noordanus (1996) that in fact speakers are not consciously aware of every error they utter. Furthermore, as Postma (2000) argues, Mackay's (1980) node structure theory lacks an external monitoring loop and, thus, predicts that the same number of errors should occur in a silent or noise-masked speech condition as in normal speech since errors are created via node activation. When speakers were asked to report their own errors in either a silent, noise-masked or normal auditory feedback condition by pressing a button, they reported fewer errors if they did not have auditory feedback (Postma & Noordanus, 1996). Similar results have been reported by Dell and Repka, (1992) and Postma and Kolk (1992). In the light of such evidence, the assumption of no external loop in the node structure theory is highly suspect.

Levelt (1989) disputes such results arguing that the self and other-error data sets should not be compared and that monitoring for errors in the speech of others is highly dependent on context. Levelt (1983; 1989) proposes and defends the perceptual loop theory of self-monitoring, which posits that the process of perceiving one's own errors is equivalent to that of perceiving another's. To capture this effect, the editing component of speech is coupled with the comprehension mechanism in a double-loop device. The first phase of speech production is "conceptualization", where the speaker realises an intention to convey information. During the conceptualization phase, the speaker can refer to a *discourse model*, or record of what was said previously in the dialogue, and a *situational* model, or model of the physical world around him and the objects in it, in order to make his or her own contribution relevant to the conversation. The speaker can also *monitor* his or her own speech, whether it is overt or internal. By doing all of these things, the conceptualizer produces a *preverbal message*, which passes to the *formulator* for *grammatical encoding.*

During the formulation phase, the abstract conceptual structure of what the speaker intends to say is mapped onto a linguistic structure (Levelt, 1989). First, the message must be grammatically encoded, or mapped onto an appropriate syntactic structure. The formulator has access to a store of *lemmas*, or units of lexical meaning. For example, the concept behind the verb *to buy* requires that one person spend money in order to obtain ownership of a particular object. A lemma also provides the speaker with syntactic information about the lexical item. For example, *buy* is a verb which requires a subject performing the action, a direct object that is bought and an agency from which the item is bought. Levelt uses the terms *surface structure* for messages that have been grammatically encoded.

Next, the surface structure, or syntactically acceptable string or organized lemmas, is *phonologically encoded*. Each lemma also comes with a phonetic or articulatory plan for how to pronounce the word, which is devised during the phonological encoding phase. Also, during this phase the phrasal stress for the whole message is determined. The end-product of phonological encoding is the *articulatory plan*. Levelt (1989) views this representation as internal speech and with it assumes a certain degree of attention to it by the speaker (McNeill, 1987). However, if the speaker doesn't attend carefully to the articulatory plan, an error might occur because failure to attend to the plan causes an error to go unnoticed.

A speaker can attend to his or her own speech via the monitor. Levelt (1989) posits two loops with his model of speech production. One loop, which travels from the phonetic plan (Figure 4) to the speech comprehension system, is utilised to monitor internal speech, and thus makes it possible to prevent errors from being pronounced (see Section 2.1.1 for a description of a 'covert' repair). The other loop is a route which enables the monitor to detect errors occurring in overt speech via the auditory loop and the language comprehension system. The advantage of such a system is that no other editing devices have to be stipulated; either overt speech or the phonetic plan for the predetermined articulations will suffice.

Since the error detection and correction processes can be extremely rapid, the purpose of any theory must be to capture this effect. Laver (1980) claims that the entirely feed-forward design of his system explains the rapid repair process. Furthermore, Laver suggests that the perceptual system may be preset and thus accelerate the perception process (pg. 301). Laver is not explicit about how the perceptual system is preset or in what way. Both Laver and Levelt (1989) predict that replanning of the utterance will occur after the cut off point in an overt error. Due to its distributed nature, the node structure theory is capable of detecting an error at any point in the production process (Mackay, 1987). Postma (2000) points out that this capability is a definite strength. He cites a study by Oomen and Postma (2000) in which speech rate and error-to-cut-off and cut-off-to-repair rate are examined. When the speech rate accelerates, the rate at which the errors are perceived and repaired also increases (Oomen & Postma, 2001a, 2001b) .

This finding supports either the production-based theories of Laver (1980) or the node structure theory but causes problems for the perceptual loop theory because the perceptual loop theory fails to account for the fact that repair rate increases with speech rate as shown by Oomen & Postma (2001a). The perceptual loop theory of Levelt (1983) relies upon the auditory channel and the comprehension system in order to repair an error.

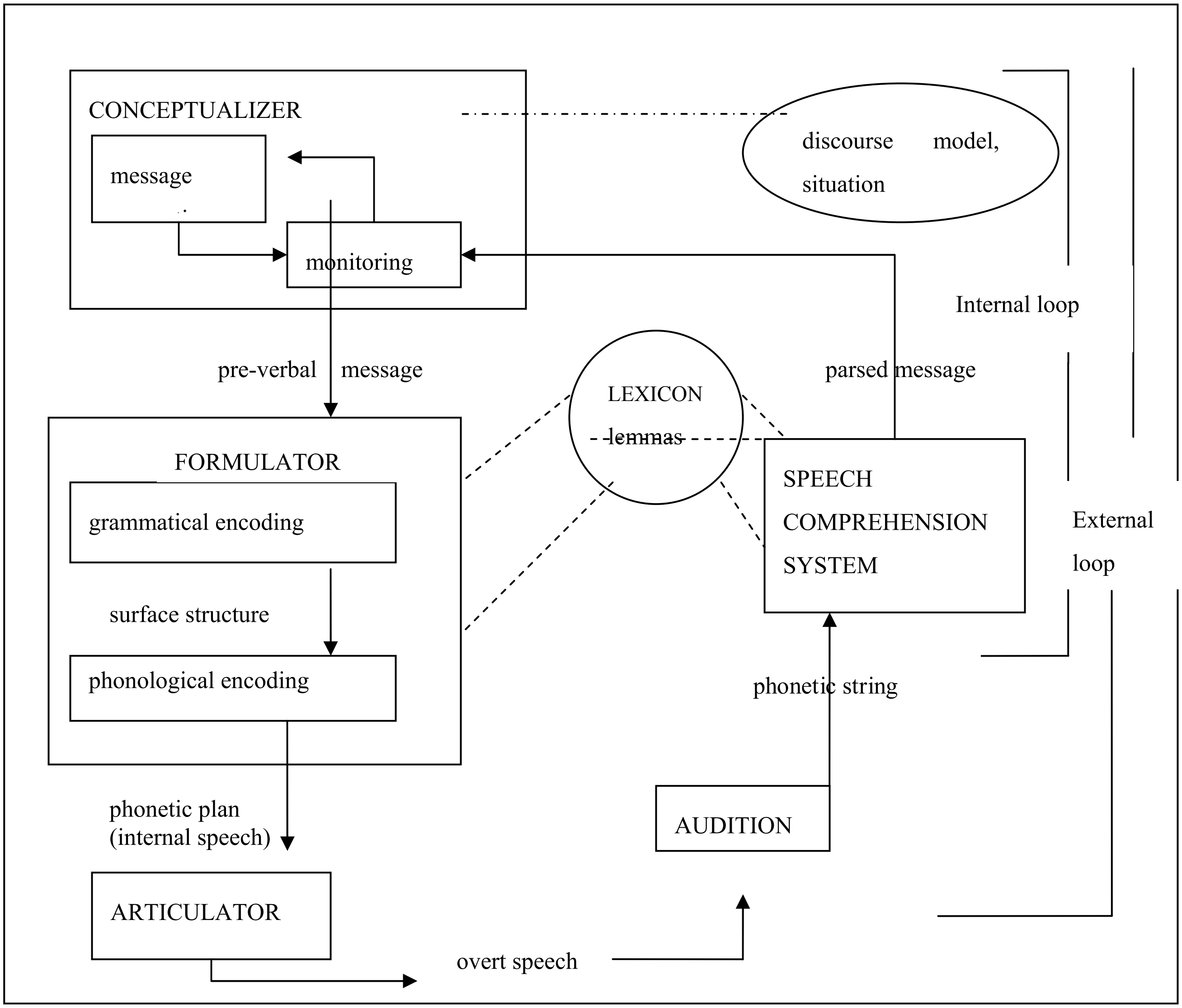


Figure 4. A copy of Figure 1.1 from Levelt (1989) showing the speech production and perception system

There is no particular reason for perception to speed up at faster speech rates. Hartsuiker and Kolk (2001) designed a computational simulation to test the perceptual loop theory in light of Oomen and Postma's (2001a) claims. A simulation of faster speech rates confirmed that the perceptual loop theory could account for error detection in accelerated speech. The perceptual loop theory employs both production and comprehension via its inner and outer loops. Hartsuiker and Kolk incorporated this scenario into their simulation and found that comprehension speed increases in parallel with production rate or that the comprehension constant is small.

The purpose of this section has been to explain three differing theories for how speakers are thought to detect and correct speech errors: the production-based account of Laver (1980), the perceptual loop theory of Levelt, (1983) and node structure theory as proposed by MacKay (1987). Henceforth, I will more or less adopt the view proposed by the perceptual loop theory because this hypothesis stands out as the sole hypothesis to incorporate retrospective processes and error awareness. Levelt proposes that speech error correction is a ‘marginal form of executive control’, meaning that the speaker must expend energy in order to complete it (Levelt, 1989, p. 22). A speaker may be aware that an error or a disfluency has occurred but that does not mean the speaker has used the error or the disfluency strategically. As explained in Chapter 1, speech errors and disfluencies are not the same phenomena

In his perceptual theory of monitoring, Levelt (1983; 1989) devises a categorisation system to classify repairs based on the reasoning behind the repair. According to this view, a potential repair is detected because speech production incorporates an internal monitor. Once an error is detected, the appropriate corrective action is taken. Examples are shown in Figure 5 as reprinted directly from Levelt (1983).

REPAIR TYPE	**TRANSCRIPTION**
D-REPAIR	**We beginnen in het midden met … in het midden van het papier met een blauw rondje** We start in the middle with…in the middle of the paper with a blue disc
A-REPAIR	**We gaan rechtdoor offe….We komen binnen via rood, gaan dan rechtdorr naar groen** We go straight on or…We come in via red, go then straight on to green
E-REPAIR	**Een eenheed, eenheid vanuit de gele stip** A unut…Unit from the yellow dot
C-REPAIR	**En aan de rechterkant een oranje stip, oranje stip** And at the right side an orange dot, orange dot

Figure 5. Levelt's (1983) four major cognitive categories of repair

The first type of repair occurs when the speaker notices that s/he could have formulated the most recent utterance in a more efficient manner. In other words, the speaker makes a *D-Repair* when s/he utters something “different” from the original intention, for example when the speaker decides to say *"of the paper"* before saying "*with the blue disc"*. D-Repairs contrast with A-Repairs because as the example in Figure 5 shows, the speaker attempts to provide more contextual information for a listener "*We come in via red..."* who may require it to understand the best way for proceeding to the green dot. *A-Repairs* are commonly known as 'Appropriateness Repairs' where the speaker is monitoring for the applicability of the information in the given context and *not* for error. Appropriateness repairs differ from D-Repairs in that an Appropriateness repair will never co-occur with an editing term such as '*er, I mean, oops'*, but a D-Repair can (Levelt, 1989).

During an E-Repair, or *Error-Repair*, on the other hand, a speaker has monitored for error and found something to repair. In the example in Figure 5, the error was phonological in nature as the speaker mispronounced the word "*eenheed"*. This sort of repair corresponds directly to the sort of repair considered in the next section (2.1.2.) and also in most of the literature. Levelt (1983) divides E-Repairs into separate categories depending upon their linguistic nature: lexical repairs (EL-Repairs), Syntactic Repairs (ES-Repairs) and Phonetic Repairs (EF-Repairs). For present purposes, these sub-types are not relevant.

The fourth sort of repair type, C-Repairs or *Covert-Repair*, occur when the speaker has made a repair, but has done so in a covert fashion, as shown in example (4) above when the speaker simply repeats "*orange dot, orange dot"*. In this case, the monitoring loop caught the repair in enough time so that the speaker did not overtly pronounce the error that he or she was in the process of monitoring for at the time. Since there is no overt, surface evidence of either error or repair, I leave a more in depth analysis of C-Repairs for future research because I am predominantly interested in investigating whether a speaker used disfluency to signal to the listener or whether disfluency merely displays that the speaker encountered trouble. Until we know more about the intentionality of disfluency, this investigation requires overt errors that the listener could have heard. Finally, Levelt (1983) posits an R-Repair group ('R' for 'Rest') to classify all the anomalous examples which didn't fit into any other categories.

Clark and Wasow (1998) do not develop a classification system of disfluencies per se but they do ascribe a cognitive function closely related to Fox Tree and Clark's (1997) findings for repetitions in collaborative dialogue. Clark and Wasow analyse only repetition disfluencies from the Switchboard and London-Lund corpus. It is their view that disfluencies (or speech repairs as they refer to them) have a strategic function to perform in signalling speaker difficulty to the listener. Repetitions are the focus of analysis because similarly to Fox Tree and Clark's findings they signal a speaker's commitment to a particular utterance. A speaker retraces a portion of the reparandum in order to signal their difficulty in planning and indicate that the previously uttered portion will undergo repair. The term ‘Self Management’ (SM) or in a later article ‘Own Communication Management’ (OCM) is used by some researchers to refer to disfluencies (Allwood, Nivre, & Ahlsén, 1990).

Allwood et al. argue in a similar manner to Clark and Wasow (1998) and Schegloff et al. (1977) that disfluencies can fulfill a pragmatic function in speech. According to their hypothesis, OCMs are a natural part of the linguistic system and its complex rota of turn management and utterance planning. The main thrust of the argument is possibly to provide defense against Chomsky’s (1965) claim that disfluencies (or OCMs) are outside the traditional notions of ‘langue’ and therefore not worthy of linguistic study. Allwood et al. (1990) set out to show that this is incorrect by devising an entire classification system from speech collected in different social situations. OCMs are the speaker’s way of managing their own communication, and as such can either signal ‘choice’ or ‘change’ (Allwood et al., 1990). A speaker signals ‘choice’ through filled pauses, repetition or silent pauses; using these items signals that the speaker needs to stop “to gain time for processes having to do with the continuing choice of content and types of structured expression” (Allwood et al., 1990, p. 10). Allwood et al refer to filled pauses as ‘simple self-management expressions’ and silent pauses as ‘pauses’. Otherwise, an OCM can signal ‘change’ by deleting, reordering, inserting or substituting words; the function of a signal for change is to enable the speaker, on the basis of various feedback processes (internal and external), to change already produced content, structure or expressions. The term ‘disfluent’ is inappropriate to refer to such phenomena since according to Allwood et al. (1990) such hesitations are a fluent part of conversation and are under the speaker’s control. Allwood et al. do not discuss whether they consider ‘tip of the tongue’ states to be a fluent part of conversation that is under the speaker’s control.

To conclude, both Clark and Wasow (1998) and Allwood et al. (1990) suggest that the process of correcting an error or disfluency is under the direct control of the speaker and, furthermore that both speaker and listener are somehow capable of incorporating this information into the intended message during the rapid-fire speech production process. Both studies tend to focus on the role of the speaker and speaker’s strategic intent during speech production. This view is problematic when the assumption of strategic intent becomes the norm rather than a potential deviation. Could strategic intent ever really be the norm and if so, what reasons would a speaker have for mispronouncing words and making their utterances difficult to understand? I will address this issue further in Section 2.1.2 and in subsequent experimental chapters.

2.1.2 Structural speech repair classification schemes

In contrast to Levelt’s cognitive theory of disfluencies, Shriberg (1994) develops a preliminary disfluency classification system based on speech from three corpora, ATIS (Dahl et al., 1994; MADCOW, 1992), SWITCHBOARD (Godfrey, Holliman, & McDaniel, 1992; Wheatley et al., 1992) and AMERICAN EXPRESS/SRI (Kowtko & Price, 1989). The aim of the study is to develop an atheoretical categorization system for disfluencies. She does not investigate disfluencies (or speech repairs) that cross speaker turn boundaries; a disfluency begins and ends within the same turn and is initiated only by the speaker making the correction. Shriberg’s motivations for her classification system are subjective with respect to the discourse history of a referent. For example, if a speaker makes an error while describing a network of nodes (*‘move the block ... the green block’*), this error could be classified as either an error repair or an appropriateness repair in Levelt’s (1983) system. The classification is dependent on the speaker’s model of the listener and whether the speaker believes the listener understood *block* or specifically *green block* the first time around (Shriberg, 1994, p. 13). This classification system was designed to be subjective since in most cases it is not possible to determine the nature of the speaker’s model of the listener, or indeed whether there even was one. Additionally, Shriberg states that disfluency rate depends on cognitive variables such as the complexity of the utterance under preparation and its linguistic structure. Her classification system incorporates eight types of disfluencies (see Table 2, page 39).

Both Shriberg (1994) and Lickley (1994; 1998) categorise disfluencies according to an atheoretical structure. Lickley’s (1994) examples come from a corpus of casual conversations.

REPAIR TYPE	**TRANSCRIPTION**
Repetition	Right **there’s a** ….there’s a line about a quarter of the way down
Substitution	**a vertical** | a horizontal line
Insertion	**two |** about two centimetres above from the bottom of the page
Deletion	on no **what |** the line stops at the flagship

Figure 6. Lickley's (1998) disfluency categorisation system

Their five types are based partially on the word-level adaptations made during the repair and partially on the psycholinguistic aspects of repair. Examples, taken directly from Lickley (1998), are presented below for exposition. Reparanda are denoted in bold and IPs are represented with a horizontal bar.

As shown in Figure 6, the speaker repeats the phrase *there's a* once before describing the line in question. Repetitions are always instances of exact repetition, with no additional words. Unlike Page (1999), who presents a disfluency taxonomy for medical transcription, Lickley allows repetitions to consist of either full word or word fragments (eg. *The ben- the bench)*, on the grounds that the difference is as likely to be a perceptual illusion as a real difference in many cases.

Substitutions are detectable when one word or string of words is replaced by another, as occurs in Figure 6 above. Notice, however, that the indefinite article *'*a*'* occurs in both the reparandum and the repair. As Shriberg (1994) and Lickley (1998) point out, substitutions may contain occurrences of repeated words. The repeated word can often be an anchoring device for detecting the disfluency. Similarly, insertions, as shown in Figure 6, may also contain repeated words. The defining characteristic of an insertion is the fact that a word that did not appear in the reparandum has been added to the repair. In Figure 6, we see this exemplified with the addition of *about* in front of the original *two centimetres*. Substitutions and Insertions would usually be classed among Levelt's (1983) A-Repairs, as they tend to modify the original utterance to contain more accurate or precise information.

Finally, *deletions* occur when the speaker has interrupted herself but has not repeated or substituted any portion of the reparandum for another in the repair. For example, "*on no what ... the line stops at the flagship"* as shown in Figure 6. In a sense, a deletion is a covert repair because the disfluency analyst has little overt knowledge of what error the speaker monitored for during the repair.

Page (1999) developed a disfluency classification system to recognize and remove disfluency patches of speech in medical dictations so that the fluent portions can be automatically transcribed. Savova (2002) employed this system in her thesis. According to this system, disfluencies can be *exact repetitions (***with a…with a***),* *exact substitutions (***five correction seven***), repetition and substitution (***does not…did not***), repetition and insertion (***to clean...to try to clean***)* and *repetition with deletion (***no spotting dysuria or abnormal … correction no spotting or dysuria***)*. Lickley's (1998) definition of deletion differs from that of Page. According to Page's disfluency classification system, deletions must contain repeated words in both the reparandum and the repair. Page's system does not consider deletions without repetition as deletions, and so Page would have no way to classify the example of a deletion given by Lickley in Figure 6. This fact might account for the scarcity of deletions reported in Page (1999) for medical transcription data.

Following Nakatani and Hirschberg's (1994) disfluency description system, Heeman (1997) devises a statistical model for detecting disfluencies in conjunction with discourse markers (DMs) and boundary tones. His approach is similar to the approach used by Hindle (1981) in that both divide their repair taxonomy into three types: fresh starts, abridged repairs and modification repairs. Heeman argues that the solution to finding disfluencies, or speech repairs as he calls them, is intrinsically linked to the solution to finding the other two (cf. Wang and Hirschberg, 1992 for information on detecting tones in conjunction with repairs; Hirschberg and Litman (1993) for information on detecting DMs in conjunction with repairs). Furthermore, Heeman (1997) views part-of-speech tagging (POS tagging) as integral in detection of all three phenomena and so he implemented this into his model. Heeman's (1997) view on disfluency detection is rooted in a desire to design a computational model that can be implemented in speech recognition systems, and thus provides valuable insights to both discourse processing and speech recognition.

A fresh start corresponds to Lickley's (1998) deletion or Hindle's (1981) restart, in that the speaker abandons a turn and starts again anew. For example, a speaker might say ***I need to send*** *... let's see, how many boxcars can one engine take* to use Heeman's own example (pg. 11) where *I need to send* is the reparandum, *let's see* is an editing term, and *how many boxcars can one engine tak*' is the alteration or repair.

In the next type, an abridged repair, the reparandum is viewed as null or empty. Instead, it contains only an interruption point and editing term. To use Heeman's own example once again: *We need to um…manage to get the bananas to Danville more quickly* (pg 13). Here, the interruption point occurs just after *we need to*, the editing term is the filled pause *um*, and *manage to get* begins the continuation. There is no correspondence between this type of repair and anything in Lickley's (1998) classification system. Heeman points out the difficulty in telling whether terms like *let's see* or *well* are editing terms, since they could also be construed as DMs. These terms are only considered to be part of an abridged repair when they occur mid-utterance and seem as if they weren't intended as part of the utterance. Furthermore, he points out that it is sometimes tricky to say whether phrases like *manage to* are not instead intended as substitutions for *need to* as in the following example of a modification repair.

Finally, the third type of repair in Heeman's (1997) system is a modification repair, where there tends to be a strong similarity between reparandum and repair. The reparandum material can be repeated verbatim, as Lickley's (1998) repetition, or partially as in Lickley's (1998) substitution: *You can* **carry them both on**…**tow them both on** *the same engine*. As Heeman points out, Hindle refers to this type as a restart. Heeman admits that it can be difficult to discern whether something should be classified as a modification repair or a fresh start, although filled pauses tend to co-occur with modification repairs while editing terms like *I'm sorry* tend to occur with fresh starts.

Heeman (1997) tested his model on The TRAINS corpus collected at the University of Rochester. With 34 speakers arranged into 25 different pairs, it is most similar to the HCRC Map Task corpus (Anderson et al. 1991). Participants were asked to discuss a circuitous train route with five cities on it. They were provided with information about how many engines and boxcars were available from each city and the location of various factories and warehouses. One person played the role of the system while the other played the role of the user and together they solved fictional problems presented to them. The participants saw similar but not identical maps: the system map contained more information about route time between destinations. Participants sat in the same room but did not have visual contact.

Overall, Heeman's model can detect and correct 65.9% of all speech repairs with a precision of 74.3%, before any syntactic processing has occurred. The full model, which uses POS tags to find DMs, is capable of identifying 97.3% of all DMs with a precision of 96.3%. The model can identify 71.8% of all turn-initial intonational boundaries with a precision rate of 70.8%.

This section has explained six separate disfluency classification systems to explain the differences between cognitive classification systems and structural classification (Allwood et al., 1990; Heeman, 1997; Levelt, 1983; Lickley, 1998; Page, 1999; Shriberg, 1994). As can be seen by studying any one of these disfluency classification systems, disfluencies are not speech errors. Disfluencies occur when a speaker has changed his or her mind and revised a portion of an utterance. Speech errors can go unnoticed or changed. As explained by Levelt (1983), the same mechanism in language production, the monitor, is used to detect both disfluencies and speech errors in speech. Just as there is a terminological difference between disfluencies and speech errors, there is a terminological difference between the terms used to describe disfluency. As shown in this section, different researchers approach disfluency with different methodologies and theoretical purposes in mind. In the next section, I will explain how these approaches can affect disfluency terminology and defend the terminology used in this thesis.

2.2 Disfluency Terminology

As Eklund (2004) points out, the terminology used to refer to the phenomenon under investigation in this thesis is a subject in its own right. This section will investigate the motivations for particular terms in various disciplines and eventually clarify and define the terms used in this thesis.

2.2.1 Disfluencies, Hesitations and Speech Repairs

In Chapter 1, I explained the differences between disfluencies, stuttering and speech errors. This section will review the literature with respect to disfluency terminology while asking the question ‘Does disfluency depend on the ear of the beholder?’ For each of the studies reviewed, I will consider each work according to:

1. corpus type
2. discipline of research (Computational, Pragmatic, Psycholinguistic)
3. the structure of the disfluency classification system
4. disfluency types within the classification system
5. the role of speakers and listeners

Knowledge of the **corpus type** under analysis is of importance because as Shriberg (1994) points out, disfluency frequencies and types tend to vary across corpora. For example, a corpus consisting of dialogues is much more likely to contain what Schegloff et al. (1977) term “other-initiated repairs” where a corpus of a news broadcast is more likely to contain more self-corrections.

Likewise, knowing which **discipline** the researcher came from can tell the reader something about the researcher’s ultimate goals in approaching disfluency. A computational linguist interested in building an effective means of detecting disfluencies is likely to have very different views and methods from a sociologist studying the ways in which people use disfluencies in interaction. Both of these researchers are also likely to differ from the psycholinguist cum phonetician who is likely to be interested in the potential cues listeners employ in perception or the mechanisms causing disfluency in language production.

By knowing the **disfluency structure** and **disfluency types** considered in each classification system, one becomes aware of how inclusive and thorough the system is. This allows the reader an opportunity to classify disfluencies according to a particular system and thereby see how effective and reliable a classification system is. One can also compare different classification systems using the same data for an understanding of the frequencies of disfluency types and the relationships between different types of disfluencies, for example how Levelt's (1989) A-Repair corresponds to a structurally coded disfluency like a substitution or insertion in Lickley's (1998) system.

Finally, as indicated in Chapter 1, there is considerable disagreement in the disfluency field about the 'why and how' of disfluency. Proponents of the Strategic-Modelling view suggest that disfluency is a strategic signal to a listener, while proponents of the Cognitive Burden hypothesis suggest that disfluency is merely an indication of difficulty in language production. For this reason, knowledge of the **role of speakers and listeners** according to a disfluency model will reveal if the approach implies a motivation for disfluency. Are speakers seen as being in control of their errors and if so how does that impact on the implications of the research? Are listeners responsible for attending to (or even capable of perceiving) the speaker's potential cues?

Roughly, the disfluency community can be divided into two groups: those who view disfluency as a means of correcting oneself (Heeman, 1997; Levelt & Cutler, 1983; Lickley, 1994; Shriberg, 1994, 1999) and those who view disfluency as a natural part of conversation, often with a pragmatic or communicative function (Allwood et al., 1990; Clark & Wasow, 1998; Schegloff et al., 1977). Those who attribute a pragmatic or communicative function to disfluencies tend to use terms like *speech repair*, *hesitation*, *other-repair*, *self-repair* or *own communication management*. Such terms imply a communicative function rather than simply just a discontinuity in the speech stream. Those who, like Lickley (1994) and Shriberg (1994), use the term *disfluency* do so primarily only to refer to the speech stream and nothing more; a disfluent patch of speech is one that contains rewordings, filled pauses, hesitations and the like. A neutral term is used precisely because no communicative intent is assumed on the part of either speaker or listener or because communicative intent was irrelevant.

As evident from Table 2, the 'error-correction taxonomists' (eg. Lickley, 1994; Shriberg, 1994; Heeman, 1997; Savova, 2002) tend to come from a computational or phonetic community, and those who view disfluency as a conversational tool are generally members of the psycholinguistic, sociological or pragmatic fields (Allwood et al., 1990; Clark & Wasow, 1998; Schegloff et al., 1977). There of course can be division within a field as is the case within psycholinguistics as will be further explained in Section 2.3. In general, error correctionists seek to develop a method for automating elimination of unwanted text from the preliminary form of a document. Those who study dialogue concentrate on the ways in which interlocutors align (i.e. to show that they have understood what their partner meant) and view disfluency as part of this process. The question to ask is: are these different theoretical ends fundamentally opposed to one another or can they be reconciled? Are there simply two ways of describing the same thing? Are disfluencies by any other name still representative of the same phenomenon?

2.2.2 Roles of Speakers and Listeners in Disfluency Literature

Table 2 describes to some extent how the literature reviewed within this section seems to characterize the roles of speakers and listeners during conversation. When reviewing disfluency terminology, it is important to be mindful of this distinction because usually disfluencies are analysed as a means for understanding how conversation and speech production operate.

Although researchers were grouped into categories according to the discipline their study largely came from, there are exceptions to this classification when considering only the roles of speakers and listeners. Roughly, researchers within the fields of pragmatics or sociology tend to portray disfluencies as a communicative part of conversation (Allwood et al., 1990; Schegloff et al., 1977). In both cases, conversation is viewed as something to manage and control. When an error occurs, both the nature in which it is repaired and the identity of who does the repairing is focused upon. According to Schegloff et al. (1977) speakers may correct themselves or this task may be left up to the listener. Both Allwood et al. (1990) and Schegloff et al. (1977) claim that disfluencies are intentional strategizing on the part of the speaker.

As an alternative to this view, Lickley (1994; 1996), Shriberg (1994) and to some extent Savova (2002) consider speech primarily from the perception or listener's perspective and as such are not primarily interested in how or why the disfluency arose. Lickley's (1994) thesis drives him to examine how a listener recovers the speaker's meaning by examining how soon a problematic area of speech can be detected. In both cases, the speaker makes a mistake for some reason, and if an understanding is to be achieved, it is the listener's goal to process this error. Shriberg (1994) handles disfluencies from the perspective of a computer system by requiring disfluent patches of speech to be removed from the output text.

Clark and Wasow (1998) hypothesize differently by emphasizing the role of the speaker in their model of disfluency. Clark and Wasow view disfluency as a strategic signal to the listener of the speaker's commitment to the utterance (and therefore that the listener should not interrupt). This view assumes the listener's ability to perceive the signal in addition to the speaker's ability to produce such an accurate signal given the rapid pace of conversation.

Table 2. Breakdown of disfluency researchers, their corpora, disciplines, disfluency structure, disfluency types and perceived role of speakers and listeners

Author	Corpus Type	Discipline	Disfluency Structure	Disfluency Types	Roles of Speakers and Listeners
Schegloff et al. (1977)	Taped Conversations; Radio broadcasts	Sociology Pragmatics	Self-initiated vs. Other-initiated Repairs	Self-initiated self repairs, Other-initiated self-repairs, other-initiated other repairs, self-initiated failures, other-initiated failures	Part of dialogue management; self vs. other repairs
Levelt (1983	Dutch description of a network of lines and nodes	Psycholinguistic		D-Repair, C-Repair, A-Repair and E-Repair	Speakers perceive selves
Shriberg (1994)	ATIS/ SWITCHBOARD	Computational	Uses Nakatani & Hirschberg's (1994) structural system	repetition, substitution, insertion, deletion, filled pauses, editing terms, word fragments, extra discourse markers	Neutral with respect to intention
Lickley (1994)	Conversations	Psycholinguistic / Phonetic	Uses Nakatani & Hirschberg's (1994) structural system	repetition, substitution, insertion, deletion, filled pauses, silent pauses	No speaker intention; Listener editing
Clark and Wasow (1998)	London-Lund; Switchboard	Psycholinguistic / Sociology	Commit-and-Restore Model	repetitions, filled pauses, hesitations	Speakers use disfluency as signal to listener
Heeman (1997)	TRAINS corpus	Computational; Statistical Modelling	Levelt (1983) but follows Nakatani and Hirschberg (1994) and Shriberg (1994)	fresh starts, modification repairs, abridged repairs	Neutral with respect to intention
Savova (2002)	Medical dictations	Prosodic, Computational	Nakantani and Hirschberg (1994); Shriberg (1994; 1999)	exact repetitions, exact substitutions, repetition and substitution, repetition and insertion, repetition and deletion	None – but implements reliable prosodic cues in spoken speech for a computer "listener"
Allwood, Nivre and Ahlsén 1990	Swedish corpus of different genres of conversation	Pragmatics	Own Communication Management	pauses, prolongations, self-interruption, filled pauses	Implies OCMs have communicative function

Furthermore, it assumes that speakers depend not only on a model of what s/he intends to say but also on a model of what the listener has understood from the previous conversation. As

Pickering and Garrod (2004) argue, it is resource intensive for the speaker to maintain both a model of his or her own perspective and a model of the listener during speech production.

Clark and Wasow (1998) hypothesize differently by emphasizing the role of the speaker in their model of disfluency. Clark and Wasow view disfluency as a strategic signal to the listener of the speaker's commitment to the utterance (and therefore that the listener should not interrupt). This view assumes the listener's ability to perceive the signal in addition to the speaker's ability to produce such an accurate signal given the rapid pace of conversation. Furthermore, it assumes that speakers depend not only on a model of what s/he intends to say but also on a model of what the listener has understood from the previous conversation. As Pickering and Garrod (2004) argue, it is resource intensive for the speaker to maintain both a model of his or her own perspective and a model of the listener during speech production.

In Levelt (1983), it is possibly more difficult to ascertain just what sort of function speakers and listeners play. This is largely because Levelt (1983; 1989) asserts that speech production makes collateral use of an internal monitor that interlocutors possess. This internal monitoring loop shares at least part of the perceptual unit which processes incoming speech, as a listener. In the event that the interlocutor speaks disfluently, then the same mechanism that perceives errors in another person's speech is also responsible for perceiving errors made during his or her own speech. In a sense, speakers are listeners of their own speech. Although Levelt (1983) also establishes a criterion for classifying disfluencies according to their cognitive motivations, he also explicitly states that speakers have little or no access to the speech production mechanism.

Throughout this section I have used the controversial term disfluency to refer to the span of speech under investigation where disfluency means a section of speech is not fluent. By using this term, I have not assumed any sort of global communicative function implicit in the error correction process though research by Allwood et al. (1990) and Clark and Wasow (1998) may assume such a communicative function. Since psycholinguistic and cognitive research have yet to uncover how much of the speech production process is under our intentional control, it is best to suspend such assumptions. For that reason, the remainder of this thesis will employ the atheoretical disfluency surface classification system of Lickley (1998) and will continue to use the term disfluency to refer to these regions. This is done out of an attempt to remain theory-neutral and to allow the experimental data to depict the phenomenon appropriately.

2.3 Psycholinguistic Models of Collaborative Dialogue

Within the psycholinguistic research community there is considerable debate between at least two separate hypotheses about collaborative dialogue. The two theories are divided with regard to the amount of effort a speaker puts into modelling the listener during the course of the conversation. According to the first hypothesis, speakers employ intentional tactics during conversation and constantly check a listener model, or a model of what the listener could know, during speech production (Clark, 1994; Clark & Wilkes-Gibbs, 1986). This hypothesis will be referred to as the Strategic-Modelling view here. According to the second hypothesis, the Cognitive Burden view, the process of intentionally modelling a listener and adapting utterances to the listener is taxing (Horton & Keysar, 1996). Speakers may choose to ignore cognitively taxing feedback during a conversation, if the effort is too great (Horton & Keysar, 1996).

Both hypotheses have been the subject of considerable experimental research using a variety of techniques. As this research has progressed over time, it has become possible to outline the ideal profile of a speaker as viewed by the two theories of collaborative dialogue. In subseqent sections, I will outline these ideal profiles as portrayed in experimental results obtained in the field.

2.3.1 Strategic-Modelling View

Throughout this thesis, I will refer to the view that speakers regularly model their listeners as 'strategic-modelling'. In actuality, the Strategic-Modelling View is an amalgamation of hypotheses, most of them originally proposed by Clark and colleagues. As discussed in Chapter 1, Clark et al. (1983) propose the *Principle of Optimal Design* which suggests that speakers form their utterances by referring to a mental model of the listener. To resolve the 'Mutual Knowledge Paradox', Clark and Marshall (1981) suggest that speakers do not require full mutual knowledge but instead can rely on common ground information based on their *physical, linguistic,* and *community copresence* with their listener. In order to share common ground with a listener, however, the speaker must have a model of the listener. The speaker refers to this model of the listener when designing utterances or 'collateral signals' during dialogue (Clark, 2002).

In an alternative view, Clark and Wilkes-Gibbs (1986) propose the *Principle of Least Collaborative Effort* to suggest that dialogue is a joint activity in which speakers and listeners share responsibilities to ensure the success of the dialogue. According to this principle, speakers and listeners should try to minimise the effort required to establish an understanding. This is taken to mean that speakers are fully capable of attending both visually and auditorily to their listener's feedback throughout the entire dialogue. If the speaker encounters difficulty during the course of the dialogue, it is the responsibility of both participants to resolve the conflict. To do this, speakers may also ‘signal’ his or her intentions to the listener and may mean something by their choice of signal (Clark, 1996, Chapter 6).

> “The logic here is based on a *principle of choice*: Whenever speakers have more than one option for part of a signal and choose one of the options, they must **mean something by that choice, and the choice is a signal**. ” (Clark, 1996, p. 261)

Grice (1957) distinguished between *non-natural meanings* (e.g. A glance at a watch which in a certain circumstance means ‘We’re late’) and *natural meanings* (e.g. red spots on the skin meant that Brad had the measles). Clark uses the term *signal* to refer to what Grice calls non-natural meanings and the term *symptom* to refer to Grice’s natural meaning. That is, a symptom has a natural meaning while a signal is used by a speaker to mean something in the current circumstance (Clark, 1996). Signals may be linguistic (Paul saying ‘I’m hungry’ to mean that he is hungry and wants to eat some food) or non-linguistic (Angelina points to a bowl with food to mean that Paul can eat the food in the bowl) in nature. As Clark states, a signal does not have a meaning behind it unless a speaker uses it to accomplish a conversational goal; nor can a speaker utter anything meaningful without using some sort of linguistic or non-linguistic signal.

Clark (1996) outlines three strategies that the speaker may pursue during the discourse. Speakers who employ the *stop-and-continue* strategy will present their utterances phrase by phrase and may pause between phrases in order to formulate the next chunk. Speakers may also use the *commit-and-repeat* strategy which involves initiating an utterance before it is fully formulated, stopping to finish the formulation, and then upon resuming begin by repeating the previously uttered word for two reasons: 1) to show commitment to the utterance and 2) to provide continuity for the listener. Alternatively, the speaker may employ the *commit-and-repair* strategy if she changes her mind about what to say mid-utterance. In this case, the speaker would substitute one word for another, insert a new word, or delete a word and start afresh. By doing this the speaker once again signals to the listener that she is attending to both the listener and the utterance (Clark, 1996; Clark & Krych, 2004; Clark & Wasow, 1998; Fox Tree & Clark, 1997). The most important thing according to the strategic-modelling view is that there must be coordination between speaker and listener so that any disruptions are efficiently handled. This requires that the speaker must pay attention at the critical moment, when the listener needs attention most. For this reason, speakers constantly monitor their listeners and send collateral signals of this attention. Speakers who cannot monitor their listeners at all are predicted to encounter difficulties during the dialogue (Clark & Krych, 2004).

The theory of *audience design*, or the notion that utterances are designed for the listener, has been further developed in work by Brennan and colleagues (Brennan, 2004; Brennan & Clark, 1996; Brennan & Schober, 2001; Kraljic & Brennan, 2005). Although Brennan's work tends to differ in its conclusions from Clark's conclusions, I will classify the notions of *joint action* and *audience design* as part of the Strategic-Modelling View. A major difference, however, is that Schober and Brennan (2003) describe some processes in dialogue as automatic, and therefore not under the intentional control of the speaker whilst other processes do seem to be strategic. As an example of a strategic process, speakers establish "conceptual pacts" with their listeners over the course of the dialogue (Brennan & Clark, 1996). When establishing a conceptual pact, either the speaker might behave in an egocentric manner and require the listener to adopt this perspective as well or the speaker might behave altruistically and refer to items according to the listener's point of view. As an example of an automatic process, Bard and Aylett (2001) showed that while speakers adapted the definiteness of their referring expressions, their articulation (measured in terms of phonetic duration) did not change upon second mention. Likewise, Kraljic and Brennan (2005) find that prosodic lengthening does not seem to be part of audience design: a speaker's choice to lengthen at a prosodic boundary seemed to depend on the speaker's processing of the syntactic structure, rather than on the listener's needs.

Brennan (2004) reports the results of an experiment done without eye-tracking. In this experiment, she asked subjects to participate in a 'car parking task', where subjects moved icons around a computer screen with their mouse. In a visual evidence only condition, subjects could see their partner's icon as well as provide instructions verbally; in a verbal-only condition, subjects could not see the other icon and could only give verbal instructions. Brennan found that subjects were most efficient in the visual-only condition: fewer words were required to accomplish the task in the visual condition compared to the verbal-only condition. As evidence of mutual responsibility, subjects who had visual feedback would sometimes interrupt their own utterances (e.g. "And park right in Memor-…right there, that's good") if they could see that the icon had already reached the desired location (Brennan, 2004; Brennan & Lockridge, 2004). Interruptions like these present an interesting case to watch out for since a purely structural classification of disfluency might necessarily classify them as disfluent.

While reporting the results of their experiments, Brennan and Clark (1996) and Schober (1993) discuss how lexical entrainment can be used as evidence for conceptual pacts. Lexical entrainment is the use of a single name or style of name in the expressions which refer to an entity over the course of the conversation. An experiment in which interlocutors were assigned a card-matching task found that referring expressions were used based on their recency, or what the expressions referred to the last time around (cf. Garrod & Anderson, 1987), frequency and informativeness in the discourse (Brennan & Clark, 1996). In Brennan and Clark's card-matching task, speakers were given cards with pictures of everyday objects (i.e. shoes, dogs). In each set of cards, there were often more than one type of shoe (penny loafer, high heel, tennis shoe) or dog (Scottish terrier or cocker spaniel). Speakers referred to objects by using the same referring expression they had used in a previous trial about 81% of the time (Brennan & Clark, 1996) showing that recency does have an effect on lexical entrainment. As Brennan and Clark (1996) show, referring expressions simplify with frequency of use: speakers used more specific terms *(pennyloafer)* 69% of the time, significantly more often, after a series of four trials than they did after only a single trial. These results are in line with other experiments which have shown that interlocutors develop routines in collaborative dialogue for selecting items from the discourse and that over time these referring expressions simplify with the frequency of mention (Ariel, 1990; Bard et al., 2000; Bard, Aylett, & Bull, 2000; Bard & Aylett, 2001; Brennan & Clark, 1996; Clark & Wilkes-Gibbs, 1986; Gundel, Hedberg, & Zacharski, 1993; Krauss, Vivekananthan & Weinheimer, 1968; Krauss & Weinheimer, 1966).

Haywood (2004) tested audience design by asking participants to describe to a confederate a sequence of cards, organized by either pattern first and then colour or vice versa, to a confederate. According to this study, speakers were capable of recalling a conceptual pact (i.e. *the upside down T*) when using referring expressions in a tangram description task with previous partners if doing so would help the success of the dialogue. In further experiments, Haywood showed that speakers are capable of designing their descriptions of the cards in a helpful manner (i.e. optimal design) after a period of being the addressee. In subsequent experiments, however, Haywood showed that syntactic priming effects were stronger than a speaker's tendency to participate in audience design. Overall, Haywood concluded that speakers are capable of participating in optimal design by adjusting the word order, referring expressions and syntactic forms of their descriptions according to the listener's needs. Optimal design is a complex process, however, and involves establishing a balance between what the speaker can easily produce and what will be easy for the listener to understand according to Haywood (2004).

In two early experiments, Krauss and Weinheimer (1964; 1966) studied concurrent feedback and confirmation in dialogue. Concurrent feedback is defined as feedback from the listener that occurs simultaneously with the speaker's utterance. Confirmation is defined as the listener's behaviour as a result of the speaker's message. Krauss and Weinheimer predict that by restricting the amount of feedback the speaker receives from the listener, one can shorten the length of referring expressions that the speaker uses to refer to objects in the display. In Krauss and Weinheimer's experiments, subjects described novel graphic shapes present on cards to listeners who were seated in another booth. Listeners had the same cards as the speaker and had the task of determining which card the speaker was describing. In the 'concurrent feedback' (CF) situation, the listener could provide verbal feedback as in an everyday conversation. In the 'nonconcurrent feedback' (NCF) situation, the speaker described the card over an intercom and the listener indicated the choice of card by pressing a button on a box. The listener provided confirmation in both situations by pressing a button, which appeared as either a correct or incorrect on the speaker's box after experimental manipulation. Speakers either received 50% correct confirmation or 100% correct confirmation. From this experiment, Krauss and Weinheimer observed that speakers shortened their referring expressions in the CF condition. Referring expressions were also shorter in the 100% confirmation condition than in the 50% confirmation condition. Thus, Krauss and Weinheimer conclude that both concurrent feedback and confirmation affect the speaker's planning of the utterance.

Following Krauss and Weinheimer, a number of studies have found that the establishment of conceptual pacts seems to be a joint action (Brennan & Clark, 1996; Clark & Wilkes-Gibbs, 1986; Haywood, 2004; Metzing & Brennan, 2003; Schober & Clark, 1989; Wilkes-Gibbs & Clark, 1992) and furthermore that conceptual pacts are formed in a gradual process (Brennan & Clark, 1996; Metzing & Brennan, 2003). In an experiment performed by Schober (1993), speakers were assigned the task of describing the location of objects to a listener positioned at different angles from themselves. Schober controlled for time-pressure, the angle of separation between listener and speaker, and whether or not speakers were participating in a monologue or a dialogue. When partnered with a listener, speakers tended to adopt the perspective of the listener and solo speakers took only an egocentric perspective. Time-pressure had no significant effect on accuracy of the object description; speakers did not adopt a different perspective or describe locations more precisely when under time-pressure compared to when they had unlimited time allotted to the task (Schober, 1993). Other research has shown that interlocutors develop routines in dialogue (Clark & Wilkes-Gibbs, 1986; Brennan & Clark, 1996; Bard & Aylett, 2001).

As with any theory, the Strategic-Modelling View seems to vary in its strength. The strong version of the Strategic-Modelling View supported by Clark (2002), Clark and Fox Tree (2002), Clark and Krych (2004) and Fox Tree and Clark (1997) suggests that speakers design utterances and disfluencies as collateral signals for their listener. A weaker version supported by Brennan (2004) and Lockridge and Brennan (2002) suggests that speakers engage in audience design for some processes (e.g. referring expressions) but not for others (e.g. articulation and prosodic lengthening). This notion has been proposed before by Bard et al. (2000), Brown and Dell (1987) and Horton and Keysar (1996). Where does disfluency fall with respect to these two theories? I address these issues in Chapters 3, 4 and 5 along with a comparison of the Strategic-Modelling View and the Cognitive Burden View.

2.3.2 Disfluency as Signal

As part of the Strategic-Modelling View, Clark and Wasow (1998) propose the *Commit-and-Restore Model* of speech repair. This model takes the view that Levelt's (1983) model of repair is limited in scope and following Schegloff et al. (1977) argue that repair is an interactive process brought about through mutual participation of both participants. Here 'interactive process' means that speakers are jointly present and provide feedback in an attempt to communicate effectively. Plauché and Shriberg (1999) extend Clark and Wasow's (1998) proposal by presenting prosodic evidence of strategy in repetitive repair.

Schegloff et al. (1977) analyse disfluencies, or speech repairs as they refer to them, in conjunction with turn-taking in conversation. According to Schegloff et al., a repair can be initiated by the speaker (i.e. a 'self-repair') or by the listener (an 'other repair'); in both cases, the initiator seeks to correct the 'trouble source'. Typically, repairs are performed as soon as possible in the span of the dialogue. If the speaker initiates the repair, it is usually within the same turn, as observed by Schegloff et al. If the other person initiates the repair, the repair usually occurs in the turn after the trouble source and no earlier. A self-repair usually contains cut-offs (a.k.a. word fragments), sound stretches (a.k.a. prolongations) and *uhs* (a.k.a. filled pauses). Other repairs typically contain a Wh-question about the trouble source or a repeat of the 'trouble source', with or without an added question word. Schegloff et al. observe that generally self-repair in speech is the preferred method of correction for both participants because other-repair generally requires more turns to complete and therefore more work for both participants.

Repetition repairs are by far the most frequent type of repair in dialogue (Lickley, 1999; Shriberg, 1994), a fact which makes them interesting to study. Fox Tree and Clark (1997) conducted a study on one particular type of repetition, the repetition of the determiner *the*. In English, the word *the* can be pronounced in two ways as either *thiy* or as *thuh*. Fox Tree and Clark (1997) hypothesise that speakers can consciously choose which pronunciation they use. Moreover Fox Tree and Clark hypothesise that when a speaker says *thiy,* this choice is a signal that the speaker is encountering difficulty in speech production and is signalling this to be considerate of the listener. Fox Tree and Clark analysed 461 tokens of *thiy* and a matched set of 461 tokens of *thuh* to determine whether or not *thiy* was a signal when made during a repetition repair. Results showed that *thiy* preceded a suspension of speech 81% of the time whereas *thuh* only preceded a suspension of speech 7% of the time. *Thiy* also preceded more filled pauses, silent pauses and speech repairs than *thuh* did. From this evidence, Fox Tree and Clark (1997) conclude that *thiy* signals a major problem whereas *thuh* signals a minor problem. Fox Tree and Clark argue that speakers are conscious of their choice of *thiy* versus *thuh* in the same way that back channel commentaries like *uh-huh* and *yeah* are signals.

The Commit-and-Restore model posited by Clark and Wasow (1998) extends the views of both Fox Tree and Clark (1997) and Schegloff et al. (1977) and applies them specifically to repetition repairs. According to Clark and Wasow, the basic insight of the Commit-and-Restore model asserts that a repetition repair occurs either because of a) some problem pertaining to the grammatical complexity of the utterance, b) the speaker's desire to maintain continuity or c) out of an attempt to uphold a prior syntactic commitment. Repairs that result from grammatical complexity are accounted for under the 'complexity hypothesis' which suggests that a speaker is more likely to suspend the flow of speech prior to a grammatically complex unit. Clark and Wasow measure complexity via 'grammatical weight' which is calculated in terms of the number of syntactic, word and phrasal nodes (Wasow, 1997; Hawkins, 1994). Clark and Wasow present evidence in support of the complexity hypothesis by comparing the frequency of function words with the frequency of content words per thousand words. Content words were repeated only 2.4 times per thousand while function words were repeated 25.2 times. Furthermore, a function word was reiterated more frequently if it appeared in a more syntactically complex NP (i.e. in topic positions) than if it appeared in a less complex NP (i.e. the object of a preposition). From such evidence, Clark and Wasow (1998) argue that one can gauge the likelihood that a particular word will be repeated by referring to its word type (eg. content vs. function status) and its syntactic position in the sentence. Speakers hesitate prior to constructions with greater grammatical weight that cause uncertainty and as such tend to repeat the function words leading into these constructions, at the points of greatest complexity when the lexical words are being chosen. A similar result was presented by Maclay and Osgood (1959).

In addition to the complexity hypothesis, Clark and Wasow (1998) posit the continuity hypothesis to explain why speakers repeat error words in a verbatim restart rather than simply commencing from the trouble source. It is important to note the similarity of repetitive restarts to the C-Repairs of Levelt (1983); Figure 5 (page 29) shows that the example C-Repair from Levelt's corpus appears as a repetition on the surface. Levelt suggests simply that restarts might signal the presence of a C-Repair, a notion to which Clark and Wasow present objections.

Creation of a C-Repair does not explain why the speaker chooses to restart rather than start from the trouble spot. Instead, as Clark and Wasow go on to suggest, a speaker repeats a portion of the utterance in an attempt to maintain continuity between the reparandum and the beginning of the repair. Analysis of the location of filled pauses (*um, uh, ah)* in relation to determiners (*the, a)* from the Switchboard corpus provides evidence in support of the continuity hypothesis predictions that speakers will be more likely to pause prior to a constituent than after, that repetition is more likely to occur when a constituent has been more severely disrupted and finally that once a disruption has occurred, a speaker should strive to maintain continuity. As evidence for these claims, Clark and Wasow note that filled pauses in the Switchboard corpus occurred before the determiner (*um the*) 64 times per thousand as compared to a significantly higher 198 times per thousand times after the determiner (*the um*) .

Finally, the third hypothesis of Clark and Wasow (1998) proposes that speakers make commitments to their utterance at major phrase boundaries. The commitment hypothesis attempts to explain why a speaker might interrupt an utterance to which he or she is committed. Clark and Wasow argue that speakers make a preliminary commitment to an utterance with the full expectation of suspending it later. They claim that temporary suspensions can be tracked in the prosody of the phrase. Selkirk (1995) states that frequently mono-syllabic function words are cliticized or attached onto a content word that follows it, unless the function word was intended to be spoken in isolation or occurs at the end of the phrase. Therefore, Clark and Wasow continue, one can detect a 'phonological orphan', a case of non-cliticized function words, when adjacent function and content words are not resyllabified into one phonological word. To take Clark and Wasow's example, one would normally expect a speaker to syllabify *I'm employed* as *I.mem.ployed*, where the coda consonant of *I'm* is pronounced as the onset of the following word. A phonological orphan, evidence of preliminary commitment, is exemplified in the pronunciation of the same phrase as *Im.employed* where a pause occurs between the function word and the verb form. Clark and Wasow contrast phonological orphans with fragments. Clark and Wasow suggest that phonological orphans like *Im.employed* constitute evidence of a syntactic commitment: the speaker has interrupted on a syntactic level and has made a commitment to continue on a syntactic level:

> "When a speaker interrupts themselves on the syntactic level, as in 'Im.employed', they are making preliminary commitments both to the words themselves (I'm) and to the constituents they initiate." (Clark and Wasow, 1998, p. 227)

Clark and Wasow contrast this example of syntactic commitments with articulatory commitments, or interruptions on a phonological or articulatory level, commonly evidenced by word fragments (eg. *The ma-*). In an articulatory commitment:

> "They are committing themselves to going on with their speech … [creating] the illusion of a continuous delivery" (Clark and Wasow, 1998, p. 231)

Articulatory commitments occur because the speaker interrupted on a phonological level; Syntactic commitments occur because the speaker interrupted on a syntactic level. Syntactic commitments occur to indicate the speaker's intention to utter a certain phrase or clause. Articulatory commitments occur as means of indicating the speaker's intention to speak in a continuous fashion.

Similarily to Fox Tree and Clark (1997), Clark and Fox Tree (2002) propose a signalling function for English filled pauses *uh* and *um.* According to Clark and Fox Tree, there are three views surrounding filled pauses. The first view, the filler-as-symptom view, proposes that filled pauses are automatic items used in speech that are not under the voluntary control of the speaker (Levelt, 1989; O'Donnell & Todd, 1991). According to the second view, the filler-as-nonlinguistic-signal view, filled pauses are a signal to listeners that the speaker wishes to hold the floor while they think of what to say next. This view was originally proposed by Maclay & Osgood (1959). Finally, according to the filler-as-word view, the third view, filled pauses are equivalent to interjections like *oh* or *well.* Clark and Fox Tree go on to develop the filler-as-word view to suggest that filled pauses like *um* and *uh* can be considered lexical items with their own meanings. They hypothesise that *um* signals that speaker expects a major delay before he or she can continue speaking while *uh* signals that a speaker expects a minor or shorter delay before resuming speech. Clark and Fox Tree test these hypotheses by analysing corpus evidence from the London-Lund corpus (Svartvik and Quirk, 1980), the Switchboard corpus (Godfrey, Holliman, and McDaniel, 1992), an answering machine corpus and the Pear stories corpus (Chafe, 1980). Pause duration following filled pauses for the majority of their data is measured by trained coders in perceptual units, not in terms of any temporal duration. A unit consists of "one stress unit" and a brief pause (0.5 units) consists of "one light foot" (Clark and Fox Tree, p. 80). Clark and Fox Tree find support for their hypotheses about *um* and *uh*: *um* occurred more often before a longer delay than *uh* did (61% > 29% of the time). *Um* (0.68 units) also occurred prior to significantly longer pauses than *uh* (0.25 units) did.

To investigate whether speakers actually *plan* their filled pauses, Clark and Fox Tree (2002) analyse three prosodic locations where planning loads differ to determine whether speakers delay for the same amount of time at each location. As shown in the example below taken from Clark and Fox Tree (p. 94), location (I) is at the prosodic boundary, location (II) is after the first word and location (III) is later in the sentence.

> and then uh somebody said, [I] but um – [II] don't you think there's evidence of this, in the twelfth [III] and thirteenth centuries?

At location (I), the speaker has to plan what they want to say, plan the syntax and create appropriate prosody. Speakers are therefore predicted to pause the most in location (I). Speakers should have less need to pause in location (II) because they've already planned their message and its syntax and prosody. Speakers should pause the least often in location (III) because they have completed most of the processing by that point. After analysing the corpora, Clark and Fox Tree found that speakers used *um* more often in location (I) compared to the other locations. This finding is used to support the claim that filled pauses are planned like other words during speech. Clark and Fox Tree conclude that filled pauses *uh* and *um* should be considered words in a prosodic, syntactic, and semantic sense. Prosodically, *uh* and *um* can be cliticized onto other words (e.g. *an.duh*, *bu.tum*) and this would not be possible if they were non-linguistic sounds. Syntactically, *uh* and *um* are used to predict upcoming delay and the following speech.
Semantically, the meaning of *uh* differs from that of *um: um* denotes a major delay where *uh* denotes a minor delay. Finally, Clark and Fox Tree conclude that speakers are able to plan their preparation of *uh* and *um* just as they are able to plan the rest of language production.

O'Connell and Kowal (2005) provide empirically measured evidence refuting Clark and Fox Tree's (2002) claims. O'Connell and Kowal used Praat software to measure the duration of pauses surrounding *uh* and *um* in radio and television interviews of Senator Hilary Clinton by well-known individuals (Television: Barbara Walters, David Letterman, Katie Couric, Larry King; Radio: Juan Williams and Terry Gross). O'Connell and Kowal used instrumental methods to measure pauses because they argue that the perceptual method used by Clark and Fox Tree is highly suspect. Research by Spinos, O'Connell and Kowal (2002) found that while 85% (206/241) of the filled pauses in the London-Lund corpus were perceptually coded, there was a false positive rate of 25% (51/206) suggesting a low reliability rate from perceptual coding.

According to O'Connell and Kowal's instrumental results from the Senator Hilary Clinton data, 44% of all *uhs* and 33% of all *ums* ranged from 0.12 to 0.24 seconds in duration. Goldman-Eisler (1968) would have considered these data to be fluent because they fell beneath her 0.25 second minimum pause duration. Only 3% (4/147) of all *uhs* and 14% (9/69) of all *ums* were longer than 0.77 seconds. As O'Connell and Kowal admit, this finding does offer some support to

Clark and Fox Tree's claim that *ums* precede longer delays but *um* can hardly be considered a reliable signal of impending delay since the mean duration for all pauses following *um* was 0.44 seconds, only 0.12 seconds longer than the mean duration of all pauses following *uh* (0.32 s). Furthermore, Clark and Fox Tree relied on *perceptual* measurements of pauses and not instrumental measurements.

O'Connell and Kowal also argue against Clark and Fox Tree's conclusion that *uh* and *um* should be considered interjections in their own right. Interjections are used with an emphatic sense and can constitute a conversational turn on their own. *Uh* and *um*, on the other hand, are non-emphatic and are rarely used as a turn. Clark and Fox Tree support the notion of filled pauses being like interjections because it lends support to their theory of ideal delivery in dialogue. As Blackmer and Mitton (1991) observed people can plan their speech while they are talking without using silent pauses after a filled pause. O'Connell and Kowal conclude in line with Maclay and Osgood (1959) that filled pauses help sustain fluency but they are not signals of major and minor delays according to instrumental measurements.

The theories of disfluencies reviewed in this section have presented arguments which suggest that disfluency is used as a signal to a listener. Clark and Wasow (1998) present evidence for the Commit-and-Restore Model which advances three hypotheses about repetitive repair as a strategic signal. The problem with such a theory is that it is based only on ambiguous linguistic evidence which could also be used to support the cognitive burden hypothesis. Recall from Chapter 1 that the Cognitive Burden hypothesis of collaborative dialogue views disfluency and speech repairs as errors of a taxed production system. Clark and Wasow suggest that function word repetitions are evidence that the speaker is undergoing planning difficulties prior to a grammatically correct object. The fact that the speaker has a difficulty prior to a complex object is in line with the cognitive burden view which argues that psycholinguistic resources must compete for time. During the repetition of content words, the speaker could simply be reapportioning cognitive resources or stalling for more time. Moreover, the fact that a speaker consistently repeats words verbatim in a restart does not automatically entail that the speaker intends the action for the benefit of the listener. As argued by Barr and Keysar (2002), all mutual knowledge for both participants is also knowledge for a single participant. It could, therefore, be the case that repetition helps the speaker get back on track. Because Clark and Wasow's evidence is always composed of linguistic forms, it does not show that the speaker actually attends to the presence of a listener or intends his or action as a signal. It is merely assumed to be the case.

Clark and Wasow (1998) and Clark (2002) have suggested that a speaker employs disfluencies as a signal and indeed designs them for the listener. Underlying this proposal is the assumption that a listener is able to reliably detect disfluency in speech and furthermore that the listener is capable of detecting the speaker's intention. In the next section, I will review the literature for how one can detect a disfluency, that is what cues do speech technologists and phoneticians use to determine whether disfluency has occurred. Following that, I will review perceptual psycholinguistic literature about whether listeners are always capable of detecting genuine disfluency in speech. Finally, I will explain the theory of intentionality in speech to understand what is implied by the notion 'intentional signal'.

2.3.2.1 Modelling and Automatic Detection of Disfluencies

The answer to the question 'How do you know when a disfluency has occurred?' depends on the definitions of the word *when* in the field that poses the question. In phonetics and speech technology, emphasis is largely on those acoustic or prosodic characteristics of reparandum and repair which can be detected prior to the recognition of the linguistic string. To answer this question, we turn first to a review of the literature in automatic disfluency detection.

Disfluencies create numerous difficulties for engineers and researchers working in the automatic speech recognition (ASR) field for a number of reasons. A major goal behind many ASR applications is to produce error-free reports or transcriptions without a lot of extra manual editing. Disfluencies, where the speaker repeats or restructures the utterance, can create a problem for the ASR system because the naïve machine cannot tell what to edit and what to keep, if the system actually was able to recognize the often garbled and abruptly cut off speech in the first place (Pakhomov, 1999). In order to develop better automatic speech recognition systems, a great deal of research has been dedicated to disfluency detection (Bear, Dowding, & Shriberg, 1992; Bell et al., 2003; Hindle, 1983; Liu, Shriberg, Stolcke, & Harper, 2005; Oviatt, 1995; Oviatt, MacEachern, & Levow, 1998; Plauche & Shriberg, 1999; Shriberg, 1994, 1995, 2005). Some of these studies have focused solely on acoustic properties to detect disfluency (Plauche & Shriberg, 1999; Shriberg, 1995) whilst others have used one or more sources of knowledge, such as acoustic information, part-of-speech tagging, Hidden Markov Models or specific language models, to aid the search (Bear et al., 1992; Hindle, 1983; Liu et al., 2005; Shriberg, 2005). Still others have conducted studies of human-computer interaction to detect disfluencies (Oviatt, 1995).

In one paper focused on the prosodic aspects of speech, Plauché and Shriberg (1999) classified repetition disfluencies in which the speaker repeats *the* (*the…the*). Repetitions in this work were classified into three groups based on their prosodic attributes. Plauché and Shriberg (1999) examined pause length, word duration, the presence of non-modal (i.e. 'creaky') voicing, and pitch patterns. Each of these prosodic cues was then normalized and 'binned' according to their values. For example, the fundamental frequency of the first repeated word could be classified as either a falling, rising or complex pattern. Three clusters of repetitions emerged from this operation and each cluster was assigned an independent role, either as a *retrospective* repetitions or *prospective* repetitions as defined by Hieke (1981). A retrospective repetition acts like a connecting bridge between utterances while a prospective repetition allows the speaker to stall during lexical retrieval (Hieke, 1981; Plauché & Shriberg, 1999). In Set A repetitions like (1), the first token is often characterized by a longer than fluent duration (denoted with '+'), a rising intonation, a long pause in the interregnum, and no pause after the second token of *the*.

(1) ([pause] *making all of* **the** + + [long pause] ***the*** *family* [pause] *things work)*

The prosodic cues for Set A repetitions were found to correspond to the authors' judgment of a canonical repetitions with a retrospective function. Set B repetitions, as shown in (2), are characterized with tokens that are both slightly longer than usual in duration and a falling intonation. Instead of a pause in the interregnum, there was often creaky voice or glottalization on the end of the first token.

(2) (*I I think* [pause] ***the*** **+** [creaky] ***the*** **+** *thing is though, I I guess)*

The prosodic cues to Set B repetitions were labelled covert self-repairs by the authors. The first token of Set C repetitions were characterized by a slightly longer than fluent duration while the second token was much longer than fluent. Both Set C repetition tokens had a falling intonation. A possible pause could occur in the interregnum between the tokens.

(3) ([pause] *don't have* ***the*** **+** [pause] ***the*****+ + +** *special tools or* [pause])

The prosodic cues to Set C repetitions were classified as prospective or stalling repetitions (Plauché & Shriberg, 1999). According to this research, prosodic cues offer speech applications some insight into the speaker's strategy during the dialogue. The implications of this research are limited, since Plauché and Shriberg looked exclusively at repetitions of the first person pronoun (*I*) and the definite article (*the*) in English; in order to claim definitively that speakers employ such strategies one would want to extend the results to other repeated tokens first. Moreover, one would want to pursue other psycholinguistic or perceptual tests to confirm that the speaker's strategy was actually conveyed to a human listener before claiming that a certain set of prosodic cues signal a certain type of repair.

Prosodic cues such as fundamental frequency, duration and glottalization have assisted in disfluency detection (Plauché & Shriberg, 1999; Shriberg, 1995; Shriberg, Bates, & Stolcke, 1997). The majority of studies, however, have concluded that disfluencies are best detected by applying a variety of approaches including language models, part-of-speech tagging, prosodic cues and even semantic features (Baron, Shriberg, & Stolcke, 2002; Bear et al., 1992; Liu, Shriberg, & Stolcke, 2003; Liu et al., 2005; Savova & Bachenko, 2002; Shriberg, 2005). Liu, Shriberg and Stolcke (2003), for example, find that their disfluency prediction model works best when it uses a specially designed language model, prosodic cues and a corpus tagged for part-of-speech. In contrast, Liu, Shriberg, Stolcke, and Harper (2005) compare the performance of an HMM (Hidden Markov Model) to a CRF (conditional random field model). In an HMM, a disfluency is predicted by looking at the surrounding independent words or states. In a CRF, the probability of a disfluency is predicted for a particular sequence of words in a conditional state. The results indicate that the CRF model detects the disfluency without the use of rules as required by the HMM. The CRF model uses part-of-speech tags and information about a speaker's turn (i.e. whether or not the turn has ended) as cues to detecting disfluency. For example, the CRF will be detect that two first person pronouns have occurred in *I I have to go* and use this as a cue to detect disfluency.

Other work on human-computer interfaces has found that a very reliable indicator that disfluency will occur is the length of the utterance (Bard et al., 2001; Oviatt, 1995). In Oviatt's (1995) study, utterances tended to be longer when the presentation format was unconstrained and when the speaker had to impose their own structure. Disfluency seems to be indicative of planning difficulties (Bard et al., 2001; Bell et al., 2003; Clark & Wasow, 1998; Gregory, Joshi, & Sedivy, 2003) so speakers can be predicted to produce increased numbers of disfluencies whenever they are under cognitive load. Yet, at the moment, the speech recognition and multi-party dialogue system fields have limited knowledge about what causes a speaker to be placed under stress. For the time being then, we turn to a complete review of the prosodic cues to disfluency in order to understand what sorts of cues speakers employ and whether these cues are at all likely to be used consciously and systematically by speakers.

2.3.2.2 Prosodic Cues to Disfluency: Fundamental Frequency

Plauché and Shriberg (1999) are not the only researchers to suggest that prosodic cues to speech repair exist. Hindle (1983) claims that a computational algorithm for deterministic parsing in a repair editing system should be able to distinguish between fluent and disfluent utterances on the basis of "phonetic evidence". Following Hindle's claim, Lickley, Shillcock, and Bard (1991) conducted a gating experiment to repeatedly elicit listeners' perceptual judgments about whether a disfluency had occurred while presenting incrementally enlarged chunks of the utterance. Listeners did not perceive a single phonetic cue, but rather seem to attend to a variety of prosodic cues in order to detect disfluency. In fact, there is a substantial literature dedicated to the detection of prosodic cues to repair. It is the goal of this section to review this body of literature.

The status of fundamental frequency as a cue to repair has been the subject of much debate. There is evidence which suggests that f0 might be reset when a repair begins (Lickley, 1994; Savova, 2002), further evidence which suggests f0falls over the course of a repair (Shriberg, 1995) and rises (Nakatani & Hirschberg, 1994; Stifelman, 1993). Therefore no conclusive simple claim can be made. Arguing in support of f0 fall, Shriberg (1995) suggests that it is possible to distinguish between prospective and retrospective repairs (cf. (Hieke, 1981)) with reference to f0. Retrospective repairs exhibit a tendency towards falling f0values at the reparandum offset, while prospective repairs tend to exhibit a continuous f0 fall throughout the repair (Plauché & Shriberg, 1999; Shriberg, 1995). Hieke (1981) suggests that retrospective repairs fill a 'bridging function' to connect the repair with the reparandum after the interruption of fluency. Prospective repairs were predicted to fill a 'stalling' function to gain additional time for the speaker.

Shriberg (1995) finds evidence for falling f0, but both Nakatani and Hirschberg (1994) and Stifelman (1993) report evidence for a rise in fundamental frequency. A small reliable rise of +4.1 Hz was detectable for the absolute f0 of the nucleus of the last accented syllable in the reparandum as compared to the first accented syllable of the repair (Nakatani & Hirschberg, 1994). Stifelman reports that average f0 values tend to increase also by about +4.1 Hz for exactly repeated words. For partially repeated words, however, she observes only a 1% increase in f0 values (Stifelman, 1993). While such results might aid automatic detection of disfluencies, it is unlikely that human speakers are capable of perceiving such discrete changes in speech because it is such a small change and is likely to happen rapidly.

Finally, Lickley (1994) tests the Reset hypothesis (Levelt & Cutler, 1983; Pike, 1945), which holds that speakers reset the f0 value of the repair so that it matches that of the reparandum prior to the interruption point. Pike (1945) proposed that normal intonational downdrift is stalled if a disfluency occurs and it should be possible to excise the reparandum portion of the disfluency and create a fluent sounding version of the original utterance. Since the speaker will begin the repair at a normal sentence-initial intonation, Pike (1945) predicts that the repair onset will have a slightly higher pitch than the onset of the reparandum. Lickley extracted f0 values from both before and after the interruption point. Though he does not find a significant difference between the pre-IP and post-IP onset portion, the Reset Hypothesis cannot be dismissed, because repair type could have been a confounding factor. Partial support of f0 reset is observed for false starts but f0 patterns differently for repetitions, i.e. most of the repetitions exhibited the same f0 pattern on both tokens but in some cases f0 was lower and in one token there was an f0 fall (Lickley, 1994). Savova (2002) also finds only partial support for the Reset Hypothesis: repair onsets were higher than reparandum onsets but repair onset values depended on the values of the reparandum offsets.

Since there is so much variation within the literature, one cannot conclude anything at all about f0 as a cue to repair. Further investigation that controls for repair type, reparandum length and syntactic structure might clarify the situation. Lickley (1994) conducted a perceptual study of repair on low-pass filtered stimuli and concludes that the cue to repair is likely to be prosodic in nature because listeners were capable of detecting disfluencies in low-pass filtered speech.

2.3.2.3 Prosodic Cues to Disfluency: Duration

A somewhat more reliable cue, duration may manifest itself as cue to repair in two prosodically different ways. Overall word duration on can be used to compare identical words occurring on either sides of the IP (Bard & Aylett, 2000; Bear et al., 1992; Shriberg, 1999; Stifelman, 1993) or in the event that identical words do not exist, an overlong duration compared to a 'standard' token (i.e. prolongation) may serve as cue (Eklund, 2001). Although not a distinguishing characteristic of disfluencies, the first token of a repetition is often much longer than the second repair token (Bear et al., 1992; Clark & Wasow, 1998; Shriberg, 1999; Stifelman, 1993). The same observation has also been made for fluent speech to signal 'new' and 'given' information in referring expressions (Bard et al., 2000; Fowler & Housum, 1987). Bard et al. (2000) and Fowler and Housum (1987) find that the second mention of a referring expression (e.g. *the parked van*) was shorter in duration than the first mention.

Counterevidence against the general trend of longer disfluent first tokens is reported by both

Plauché and Shriberg (1999) and Nicholson, (2002). Plauché and Shriberg (1999) find that during a stalling repetition (i.e. Set C repetitions: [pause] *don't have* ***the*** [pause] ***the*** *special tools*) the second token is longer than the first. Their definition of a stall, however, is left somewhat vague. According to their description, the speaker is still suffering difficulty during the second token. It is possible that this conception of a stalling repair differs for that proposed elsewhere in the literature. Nicholson (2002) compared the duration of mispronounced first tokens with their satisfactorily articulated second tokens. All mispronounced pairs were carefully controlled to ensure that they contained the same phonological segments (i.e. a long vowel in the case of /ðai/ versus /ði/.). She observed that contrary to the general trend, mispronounced first tokens were on average 37 ms shorter than the repair versions. It seems that the speaker spends more time amending the second token after unsatisfactorily pronouncing it the first time around. However, as Nicholson (2002) admits, this observation was made on the basis of a small data sample and furthermore makes no predictions for other modes of speech such as monologue or other-initiated repairs in dialogue.

Although several studies have compared word durations in repetitive repair, only few studies have examined the prolongation of a portion of the word. Eklund (2001) conducted a comparative analysis of disfluent prolongation in Swedish and Tok Pisin. Tok Pisin is a language spoken in Papua New Guinea. As he observed, prolongation may be a language specific trait or at least subject to the phonological rules present in the language. Swedish speakers exhibited a preference for prolonging word final continuant segments while speakers of Tok Pisin tended to prolong word final labial and velar nasal consonants. Although the tendency in both languages was to prolong segments more often at the end of a word, the segments differed considerably in the degree of lengthening.

Lengthened words and segments can also occur in fluent speech. When studying this phenomenon, as Wightman, Shattuck-Hufnagel, Ostendorf, and Prince (1992) note, it is important to take the speech rate into consideration. Wightman et al. (1992) developed a technique to analyse normalized durations involving linear scaling in a gamma distribution. In line with Ladd & Campbell, (1991), they demonstrate that pre-boundary lengthening (eg. lengthened segments co-occurring with a syntactic boundary) can be used to distinguish among four levels of prosodic constituents, namely the foot-initial stressed word, all segments between the foot-initial vowel and the last vowel before the boundary, coda consonants before the boundary, and the vocalic nucleus before the boundary. Of these, coda consonants and vocalic nuclei exhibit the greatest pre-boundary lengthening (Wightman et al, 1992). As Wightman et al. (1992) point out it is difficult to tell whether this pre-boundary lengthening in fluent speech is under the volitional control of the speaker. Furthermore, pre-boundary lengthening could be related to some other phenomenon, for example discourse prominence. This suggests that the prosodic phenomena that some researchers have labelled as a cue to disfluency, actually occurs in fluent speech: something that is perceived as a disfluency effect may just be a prosodic boundary effect.

This section has reviewed the literature about potential prosodic cues to disfluency and discovered that neither fundamental frequency nor duration are exclusive cues to disfluent speech. Furthermore, research by Heeman (1997), Lickley, Shillcock and Bard (1991), Nakatani and Hirschberg (1994), Savova and Bachenko (2002) and Shriberg (1994) has suggested that there is no single "phonetic signal" to disfluency as originally suggested by Hindle (1981). Instead, it seems that disfluency is detected by a combination of prosodic cues, if at all (Heeman, 1997; Lickley et al., 1991; Nakatani & Hirschberg, 1994; Savova & Bachenko, 2002; Shriberg, 1994). I have now explained how one might be aware that a disfluency occurred prosodically and in terms of automatic detection. The next section will discuss how well listeners are able to perceive disfluency and whether there are any perceptual consequences as a means of evaluating the predictions of the Strategic-Modelling View that listeners are capable of this task in dialogue.

2.3.3 Perception and Processing of Disfluency

The Strategic-Modelling View predicts that listeners are capable of perceiving and processing disfluencies as a signal from the speaker. The purpose of this section is to review the literature on disfluency perception to see whether listeners can perceive disfluencies. By looking at perceptual studies of disfluency, one can determine what effect disfluency has on the listener and whether this effect is the same as what the speaker might have intended. This section will also review some disfluency processing studies to review whether disfluency is helpful or a hindrance to the listener.

Lickley (1994) investigated the perceptibility of disfluency by conducting gating experiments. His stimuli were gathered from a corpus of spontaneous face-to-face casual conversations that Lickley himself collected. Participants in the perceptual experiment heard portions of disfluent utterances each one 35ms longer than its predecessor. At each 'gate', subjects were asked to report the words they have heard and to decide whether the utterance is about to become (or has become) disfluent. From this study, Lickley observed that listeners were able to perceive disfluency before they recognize the first word in the repair. Subjects were only capable of perceiving disfluency after the disfluency had begun; they could not perceive when a disfluency was about to occur. Since listeners are not given the opportunity to listen to normal speech in 35ms portions, one could conclude that an average listener would not perceive disfluency as accurately in everyday dialogue.

In other related research, Lickley (1995) tested whether or not native Dutch speakers were capable of detecting disfluencies in normal spoken Dutch. Subjects were given a written, edited version transcript of a set of spoken instructions describing how to build a house from coloured pieces of card. Disfluencies were edited out of the written transcript. Subjects were requested to mark the transcript at any point when what was said differed from what had been transcribed. Subjects were simultaneously asked to follow the instructions and build the house from pieces of card. Lickley observed that subjects perceived filled pauses correctly 55.2% (69/125) of the time but were only capable of perceiving single-token repetitions 27% (4.7/16) of the time and single-word false starts 39.3% of the time (11.4/29). This research confirms previous observations by Martin and Strange (1968) that listeners were very poor at perceiving disfluencies when requested to do so in an online task. Listeners in Martin and Strange's experiment were better at perceiving filled pauses than they were at perceiving false starts. As Lickley argues, there may be prosodic reasons to explain why filled pauses are more easily discernable than other types of disfluencies: filled pauses have pitch features that vary across contexts (Shriberg & Lickley, 1993).

Fox Tree (1995) also conducted a similar experiment to Lickley (1994) in which subjects were asked to participate in the *identical word monitoring task* designed by Marslen-Wilson and Tyler, (1980). In this task, subjects keep a word in mind and press a button once they've heard it. Marslen-Wilson and Tyler showed that subjects were faster at identifying words when the utterance was comprehensible and less fast when the utterance was incomprehensible. Fox Tree extends this task to testing word monitoring in two types of disfluent speech: repetitions or false starts. Mid-sentence false starts caused the slowest word identification times, sentence-initial false starts the next slowest while repetitions caused no difficulty at all (Fox Tree, 1995).

In an online study of perception of disfluency from the listener's perspective, MacGregor, Corley, and Donaldson (2005) found that disfluency (in this case, a filled pause) has an immediate effect on language comprehension. Listeners heard sentences which either ended in predictable or unpredictable words. ERPs (Event Related Potentials) measure the electrophysiological response on the scalp to a certain event and have been used extensively in psycholinguistic research (van Berkum et al., 2002; van den Brink, 2004; Hagoort et al., 1999). When an unpredictable word had been uttered, ERP measurements revealed a N400 effect, indicating that the listener was indeed surprised by the word. However, if a filled pause occurred prior to the word, MacGregor et al. observed a reduced effect of the N400 effect suggesting that the filled pause had some effect on signalling the unpredictability of the upcoming word. MacGregor et al. report that control materials consisted of both highly predictable and less predictable words as determined by a cloze probability pre-test. Furthermore, half of the utterances preceding the un/predictable word were disfluent; half were fully fluent.

One must wonder though whether it is actually disfluency or just time that causes this effect. Other work has suggested that any noise (dog barking, cough, car horn) would achieve the same effect (Bailey & Ferreira, 2001). More recent work has corroborated this finding for both native and non-native speakers of Japanese: filled and silent pauses behaved in the same manner (Watanabe, Den, Hirose, & Minematsu, 2005) suggesting that listeners respond to the extra time and there is nothing particular about filled pauses.

Brennan and Schober (2001) conduct a further perceptual test of how listeners handle disfluencies in spontaneous speech. Listeners were given spoken instructions in which they were asked to press coloured squares on a keypad as quickly and as accurately as possible. The spoken instructions consisted of disfluent and fluent controls. Disfluent stimuli came in three types: between word disfluencies *(Move to the yellow- purple square*), mid-word disfluencies (*Move to the yel- purple square*) and mid-word with filler disfluencies (*Move to the yel- uh purple square*). The reaction time of the listener's key press was considered to be a measure of the helpfulness of disfluency. Listeners were found to press the correct coloured square fastest after a disfluent stimulus with a long edit interval (*Move to the yel- uh purple square*), suggesting that disfluencies contain information which help the listener resolve any processing issues (Brennan & Schober, 2001).

Brennan and Schober tested whether the phonological form of the disfluency helped the listener or whether the listener was simply aided by additional processing time. The disfluent stimuli (e.g. *Move to the yel- uh orange square)* in this experiment were electronically replaced with a pause so that there were Filler removed (*Move to the yel- orange square)*, Word removed (*Move to the uh orange square*), Disfluency replaced by a pause (*Move to the [pause] orange square)* and Entire Disfluency Excised (*Move to the| orange square*) versions. A Fluent version was used as a control. Listeners responded in the same amount of time and as accurately when a pause replaced the disfluent portion of speech as when the original filler or fragmented word was left in tact. From this evidence, Brennan and Schober (2001) conclude that the listener benefited only from the additional time and not from any particular phonological cues in the disfluent stimuli.

Bailey and Ferreira (2003a) find similar results to Arnold, Altmann and Tanenhaus (2003) after investigating disfluency, gaze and listener perception of ambiguous constituents. Bailey and

Ferreira tease apart two theories: Clark and Wasow's (1998) theory that disfluencies are used in a signalling function and a second theory that suggests that any noise (e.g. dog bark, cough) would fulfil the same function. As part of the signalling theory, Clark and Wasow predict that a filled pause following a definite determiner *the* would signal that the following noun phrase is the subject of the clause rather than the object of the old one. Recall that according to Clark and Wasow (1998), a noun phrase in topic position is considered to be 'grammatically heavier' or 'syntactically more complex' than a noun phrase in the object position. In contrast to this theory, Bailey and Ferreira suggested that any noise, speechlike or not, would fulfil the same function.

To test these theories, Bailey and Ferreira manipulated sentences that were syntactically ambiguous. In a so-called 'garden path sentence', for example *the horse raced past the barn fell* there is a temporary ambiguity between a main clause and reduced relative clause reading (Bailey & Ferreira, 2003; Pinker, 2000; Trueswell & Kim, 1998). Bailey and Ferreira (2003a; 2003b) showed that both disfluencies and environmental noises affect the time course of ambiguous constituent processing. Thus, if the listener hears an interruption, albeit a filled pause or a dog bark, when they hear *While the man hunted the* **uh** *deer ran into the woods*, they should interpret *the deer* as the subject of the clause rather than as the object because a disfluency is more likely before a syntactically 'heavy' constituent. Bailey and Ferreira showed that this prediction is indeed met. Listeners judged the sentences grammatical regardless of whether a filled pause or an environmental noise interrupted the sentence.

In a later article, Bailey and Ferreira (2005) tested a listener's gaze reaction after hearing an ambiguous sentence which instructed them how to move objects in front of them. They found that while interpreting these instructions listeners looked at the target object sooner when the presence of disfluency (um, uh) biased the participant in that direction. Listeners heard the sentence *Put the frog on the towel in the box* in an ambiguous context when two towels and related distractor items were visually present, eg. a frog sat on one towel (target at location), a frog by itself (distractor) and another towel in a box (destination at location). The speaker could be instructing the listener to put either the frog that sat on a towel into the box or to place the frog by itself into the box that also contained a towel. So, when the filled pause preceded *frog* (*Put [the uh frog on the towel] in the box*), the listener looked at frog-on-towel sooner. When the listener heard a filled pause before *towel* (*Put the frog on [the uh towel in the box]*), the speaker looked at the target object (i.e. the frog on the towel) later, indicating that s/he was entertaining the modified goal reading.

Snedeker and Trueswell (2003) show similar results for fluent speech suggesting that it is rather a different matter whether the speaker actually intends these cues as a signal to the listener. Snedeker and Trueswell (2003) show that speakers employ prosodic cues when they are necessary in order to parse an ambiguous sentence (eg. *Tap the frog with the flower*). When the context was unambiguous for the speaker, the need to signal a distinction disappeared, as did the prosodic cues. Thus, the online time course of ambiguous speech processing seems to be sensitive to what a speaker might in theory intend by including particular cues. Of course, one cannot even be certain that the speaker actually *intends* to use the disfluency in a strategic manner. If the predictions of the cognitive burden view presented at the beginning of this chapter are correct, it could also be the case that the listener would assume that the speaker emitted a filled pause because he was experiencing cognitive difficulties in producing the utterance. One could certainly predict that the results that Bailey and Ferreira (2003; 2005) get would look the same as if they plotted the complexity of the referring expression against disfluency.

This section has presented mixed results on the perceptibility of disfluency. Lickley (1995) and Martin and Strange (1968) suggest that listeners performed poorly when asked to detect disfluencies in spoken speech. In both experiments, the stimuli consisted of genuine disfluencies (e.g. filled pauses, repetitions and false starts). As both Fox Tree (1995) and Lickley (1995) observe, subjects had more difficulties with false starts compared to repetitions. As Lickley suggests, this could occur because subjects just did not notice the repetition as easily as the false start. An ERP study by MacGregor et al. (2005) suggested that listeners were less surprised by unpredictable words when a filled pause occurred prior to the word. This result suggests that listeners are capable of using at least filled pauses as a signal. As Lickley (1995) and Martin and Strange (1968) point out, listeners are better at perceiving filled pauses than any other type of disfluency. Shriberg and Lickley (1993) find that filled pauses have pitch features which explain why this may be so. Research by Bailey and Ferreira (2001) suggests that listeners responded to non-linguistic noises (e.g. dog barking, coughs, etc.) inserted into speech in the same manner they responded to disfluent speech noises. The same finding was observed for Japanese speakers by Watanabe et al. (2005) suggesting that all listeners perceive is the extra time and not the disfluency itself. Brennan and Schober (2001) observed that listeners benefited simply from having extra time to process the disfluent utterance; there was no specific phonological cue which helped them process the disfluency.

Thus far, the previous section on detection and cues to disfluency suggested that a) there is no single cue to disfluency but there may be a combination of cues and b) a number of cues to disfluency are also cues to fluent phenomena in speech (e.g. phonological boundaries). The current section reviewed whether listeners are capable of perceiving and processing disfluency. Lickley (1994) showed that listeners were only capable of detecting disfluency once it had begun. It seems that on the one hand that listeners are better at perceiving and processing certain types of disfluencies than they are others. Moreover, listeners seem to achieve the same benefits by simply having extra time to respond when asked to process disfluent stimuli. If speakers use disfluency as a signal, as suggested by The Strategic-Modelling View, then the speaker must also intend to make such a signal. The next section reviews the psycholinguistic and philosophical literature on intention in speech for an understanding of what would be entailed by the notion of an 'intentional disfluency'.

2.3.4 Intention and Speech

Section 2.3.1 distinguished between symptoms, or natural meanings, and signals, or non-natural.meaning. Part of the difference between symptoms and signals has to do with the speaker's intention in speech: a speaker uses a signal, for example the wave of a hand which might otherwise not have its meaning naturally, in a certain circumstance to mean the speaker wishes to say good-bye. Speakers use signals intentionally to have the meaning that they have. Sections 2.3.2 and 2.3.3 discussed the strategic use of repair and potentially intentional cues to disfluency. Each of the repair models presented in Section 2.1 assumes that the speakers are at least partially aware of their desire to communicate before they begin the speech production process. How do speakers devise these plans and what is known about communicative intention in speech production? These questions will be the focus of this section.

Speakers need not be acutely aware of what they are going to say before they say it, but it is generally thought within the linguistic and psychological communities that speakers produce utterances to fulfil some goal or purpose (Austin, 1962; Grice, 1957, 1968, 1989; Levelt, 1989; Levinson, 1983; Searle, Kiefer, & Bierswich, 1980; Sperber & Wilson, 1995). This particular goal or purpose is known as a *communicative intention* (Clark, 1996; Clark & Carlson, 1982a, 1982b; Grice, 1957, 1968, 1989). For example, a speaker may wish to ask a question about something someone else said; they may wish to tell the other person how to do something or they may wish to express a feeling about a particular topic. The intended goal of an utterance is its *illocutionary force* (Austin, 1962). According to Grice, a communicative intention differs from an intention to inform (or what Sperber and Wilson call an *informative intention)*. In a communicative intention, the speaker's intention is specifically to have their intention to inform recognised, that is a speaker wants the listener to know that the speaker wants the listener to know something. Any utterance that a speaker makes that has an illocutionary force is known as a *speech act* (Austin, 1962).

Levelt (1989) proposes that the process of realizing one's communicative intention occurs during the conceptulization phase (see Section 2.1.1) of speech production in a phase known as *macroplanning*. During macroplanning, a speaker breaks the communicative intention into individual speech acts. During a second sub-phase of conceptualization known as *microplanning*, a speaker will decide upon all the language-specific requirements necessary for producing various speech acts. Once these have been determined, the message is then passed on to the formulator for grammatical encoding. Levelt (1989) points out that macroplanning need not have entirely finished for an utterance before the phase of microplanning can begin. Thus, intentions should be linguistically coded at various parts of an utterance at the same time.

How do speakers know what a valid sort of communicative intention in speech is? Philosophers of language suggest that speakers rely upon *mutual knowledge* in order to communicate (Clark & Carlson, 1982a, 1982b; Schiffer, 1972; Sperber & Wilson, 1995). Mutual knowledge, according to Schiffer (1972), is an infinite regression of interpersonal knowledge that two individuals possess and according to some is required by both speaker and listener in order for communication to occur (Clark & Carlson, 1982a, 1982b; Clark & Marshall, 1981). For example, suppose Angelina and Bryce were in the same room, seated in front of a television. Angelina would know that there is a television in the room because she can see it. Bryce would also know that there is a television in the room because he can also see it. Angelina would also know that Bryce knows that there is a television in the room because Angelina can see that Bryce is watching the television. Likewise, Bryce knows that Angelina knows that Bryce knows that there is a television in the room because Bryce can sense that Angelina sees Bryce watching the television. The list of possible states of knowledge could go on forever without termination. Schiffer, a philosopher, suggested that the infinite regression of knowledge is "perfectly harmless" (Smith, 1982). According to Clark and Marshall (1981), the human mind cannot process such infinite regression during speech. Furthermore, according to Clark and Carlson, there is no need for the infinite regression, and so they suggest that a "mental primitive" exists *'A and B mutually know that p'* and then propose a recursive inference rule, *If A and B mutually believe that p, then: (a) A and B believe that p and believe that (a)* stating that only a few iterations (i.e. A must know that B knows that A knows and A knows that B intends for A to know) are necessary in order to establish mutual knowledge. By stating that only a few iterations are necessary, one removes the infinite regression of possible knowledge states. Given this inference rule, speakers can *inductively infer* the mental primitive *p* and the mutual belief. Clark and Carlson suggest that mutual beliefs can vary in strength from weak to strong. For example, suppose Angelina told Bryce that she wanted to watch 'ER' at 10:00pm. At 10:01pm, Bryce is still watching a programme on another channel, giving Angelina a reason to suspect that Bryce forgot about Angelina's desire to watch 'ER'. Thus, the grounds for mutual belief are weak between Angelina and Bryce.

Clark (1996) suggests that there are actually three representations for mutual knowledge: Common Ground-iterated, Common Ground-shared and Common Ground-reflexive. Clark refers to the infinite regress of mutual knowledge, already discussed above, as "Common Ground – iterated" or "CG-iterated". He suggests that humans cannot possibly rely on this type of mutual knowledge because human mental capacity cannot process it (Clark & Marshall, 1981; Clark, 1996). Instead, speakers can have CG-shared according to Clark (1996). This idea was first proposed by (Lewis, 1969) and was called "common knowledge":

Common Ground (shared basis)

p is common ground for members of community C if and only if:

1. every member of C has information that basis *b* holds;
2. *b* indicates to every member of C that every member of C has information that *b* holds;
3. *b* indicates to members of C that *p*.

In the proposition above, C represents any community of at least two people and *b* represents a basis for some piece of common ground that proposition p holds (Clark, 1996, p. 94). To take an example, Angelina and Bryce form a community because they are co-present in the same living room and can see the same television set. Therefore, it is common ground to both Angelina and Bryce that there is a television in the room. Once Angelina and Bryce have established CG-shared, they can derive the third type of mutual knowledge, Common ground (reflexive):

Common Ground (reflexive)

p is common ground for members of community C if and only if:

(*i*) the members of C have information that *p* and that *i*. (Clark, 1996, p. 95)

Note that this type of common ground is reflexive because it contains a reference to itself via the proposition (*i*). Clark suggests that CG-reflexive allows individuals to derive the belief that they both share the same information. Therefore, Angelina and Bryce can deduce that they both share the proposition *i* that they both share the same proposition *p* 'there is a television in the room'. Clark concludes that CG-shared is the basic form of mutual knowledge (or common ground as he refers to it) and that Sperber and Wilson (1987) were wrong to dismiss it because it is a logically acceptable form of mutual knowledge that does not require complicated infinite regression.

Clark proposes the *Principle of Least Effort* to suggest that people seek efficiency and sufficiency in when they act intentionally.

> *Principle of Least Effort*: All things being equal, agents try to minimize their effort in doing what they intend to do. (Clark, 1996, p. 224)

As a corollary of the Principle of Least Effort, Clark (1996) suggests that mutual knowledge need only be good enough for current purposes. He suggests that a speaker seeks information that a certain act will achieve completion. This proposition is formalized as:

> *Principle of Opportunistic Closure:* Agents consider an action complete *just as soon as* they have evidence sufficient for current purposes that it is complete (Clark, 1996, p. 224)

When deciding that whether an action will achieve completion, Clark suggests that people seek evidence that is “valid, cheap and timely enough for current purposes” (Clark, 1996, p. 224) to indicate that an action will achieve completion. He calls this type of evidence *Holistic Evidence*:

> *Holistic evidence:* Evidence that an agent has succeeded on a whole action is also evidence that the agent has succeeded on each of its parts (Clark, 1996, p. 225)

An action, for example calling an elevator to take Clark’s example, can be broken down into individual parts. To call the elevator, a person needs to press either the ‘up’ or ‘down’ button. If the button is working properly, it will usually light up and thus provide evidence that the elevator has been summoned and is on its way to collect the person. This is evidence that the action of summoning an elevator will achieve closure. If the light is not functioning properly, it won’t light up and the person may continue to press the button because people need closure on events. If the elevator arrives or the light does light up, the person can use this as holistic evidence that the action has succeeded. This means that the person does not need to verify each of the individual actions involved in pressing the button (i.e. extending one’s arm, extending one’s finger, feeling the button depress underneath one’s finger, etc.) because the light has turned on or the elevator arrived.

Above I explained Clark's proposal for individual actions. Clark makes the same assumptions about individuals behaving in joint actions like conversation:

> *Principle of Joint Closure:* The participants in a joint action try to establish the mutual belief that they have succeeded well enough for current purposes. (Clark, 1996, p. 226)

According to the *Principle of Least Collaborative Effort* described in Chapter 1, Section 1.2 individuals will pursue the easiest and cheapest method possible in order to complete an event. In conversation, people will look for signs of uptake from their listener that what they have said has been understood or accepted. For this reason, conversations tend to be broken down into *local projects* or *adjacency pairs* (Clark, 1996; Schegloff and Sacks, 1973). To take Clark's example:

Adjacency Pairs	**Example**
1. Summons	Jane: (rings Kate on the telephone)
2. Response	Kate: Miss Pink's office –
1. Greetings	Hello
2. Greetings	Jane: Hello
1. Question	Is Miss Pink in?
2. Reply	Kate: Well, she's in but she's engaged at the moment

By breaking down joint actions like conversation down into two parts, Clark suggests that people solve problems in an efficient manner. The first part of an adjacency pair is usually a signal of some type (e.g. Jane ringing Kate's telephone) and requires uptake of some sort in the second part (e.g. Kate answering the telephone). Clark suggests that people engaged in conversation can use communicative acts to fulfil two purposes: 1) simple communicative acts tied to the official subject of the conversation and 2) meta-communicative acts which are acts about the communicative acts at hand. To take an example:

Utterance	**Communicative Act**	**Meta-communicative Act**
A: it was uh it was a lovely day	1. [I assert] it was a lovely day	1. [Do you understand this?]
B: yes	2. [I ratify your assertion]	2. yes [I understand that]

(Clark, 1996, p. 243).

In the example above, B says *yes* not because B is agreeing that *Yes, it was a lovely day* but because B wants to show that she understands what A means by his assertion that it was a lovely day. Thus, we have an example of a joint action in conversation.

Notice that A was disfluent in his assertion ***it was uh*** *...it was a lovely day*. Clark proposes that disfluencies, silent pauses and filled pauses can be considered "signals" of the meta-communicative sort. Smith and Clark (1993) and Clark and Fox Tree (2002) have suggested that speakers choose to say *uh* when they expect a short delay and *um* when they anticipate a longer delay. A filled pause is in this way a signal to the listener about how long the speaker expects to delay. A speaker is also sending a signal when they suspend their utterance according to Clark.

> "The logic here is based on a *principle of choice*: Whenever speakers have more than one option for part of a signal and choose one of the options, they must mean something by that choice, and the choice is a signal. By this logic, [a] word-cut off is a signal: Speakers could have chosen to complete the word as formulated. To cut it off is to signal they have changed their minds about it." (Clark, 1996, p. 261).

In the example above, A repeats *it was* after the suspension and the filled pause. Clark and Wasow (1998) suggest that the choice to repeat *it was* is a meta-communicative act. According to the Commit-and-Repeat Strategy proposed by Clark and Wasow and described previously in Chapter 1, a speaker chooses to repeat portions of an utterance in order to maintain a continuous utterance that will assist the listener. In this way, a disfluency can be an intentional signal according to Clark and colleagues. In order for disfluency to be a signal, however, there must be mutual knowledge that it is a signal, speaker and listener must be working jointly towards a common goal in the most efficient manner possible as described above.

According to a number of critics, however, mutual knowledge need not necessarily exist (Johnson-Laird, 1982a; Sperber & Wilson, 1995) and certainly couldn't be used reliably by speakers and listeners to comprehend a communicative intention. In order for a speaker to communicate an intention, he or she needs not only to communicate the simple information in the intention *It's cold in here* but also what is intended by uttering the sentence in the current context (*It's cold in here because the window is open so I really want you to close the window*). In order to derive the intended inference, a listener needs to derive an infinite set of assumptions before he can know for certain that the speaker wants the window closed, to use the previous example. It is of course possible that the listener could derive a different assumption given the utterance *It's cold in here*. Say for example that the listener knows that the speaker chose to spend the winter conducting global warming research in Antarctica. The statement *It's cold in here* given the current context and previous knowledge may strike the listener as rather ironic and that compared with Antarctica the room is not cold. For the reasons illustrated by the previous example, psychologists like Johnson-Laird suggest that there is no way a listener could generate all the possible assumptions, and since a listener never needs to generate all assumptions, mutual knowledge cannot exist (Johnson-Laird, 1982, pg. 41; Sperber & Wilson, 1995). Therefore, mutual knowledge is not required for communication as Clark and Carlson (1982a) suggest (Johnson-Laird, 1982a, 1982b, 1983), nor can it be relied upon because it is always possible that a listener could infer an assumption which the speaker did not intend.

Sperber and Wilson (1995) propose an alternative to mutual knowledge. They suggest that although all human beings exist in the same world, perceptual abilities differ from person to person, concepts may differ between cultural and linguistic groups and each person has different experiences and memories of those experiences from which to drawn upon during interaction with another person. Thus, even though people share the same physical world, their *cognitive environments* differ (Sperber & Wilson, 1995). A cognitive environment is defined as the entire set of facts that are *manifest* to an individual (Sperber & Wilson, 1995, p. 39). They define the concept of *manifest* thus:

> ([1]) A fact is *manifest* to an individual at a given time if and only if he is capable at that time of representing it mentally and accepting its representation as true or probably true.

For something to be manifest, a person only has to be able to perceive or infer it. In the event that two individuals share the same cognitive environment, and the same facts of that environment are said to be manifest to both of them, then those facts can be said to be *mutually manifest*.

As the reader may observe, the concept of manifestness is similar to but weaker than the concept of mutual knowledge. Sperber and Wilson (1995) argue that “it is weaker in just the right way” (p.43). Firstly, the concept of mutual knowledge was rejected because it was psychologically and cognitively implausible: there is no way the human mind can process the infinite regression necessary to say that something, a communicative intention say, is truly mutually known. Mutual manifestness does not have this problem because it makes claims only about the cognitive environment, not about cognitive processes (Sperber & Wilson, 1995). Secondly, if humans really did possess true mutual knowledge, there would be no explanation for why misunderstandings and misinterpretations occur at such a frequent rate. Mutual manifestness abandons the need to be infallible and correctly predicts that misunderstandings should and will occur naturally during interaction (Sperber &Wilson, 1995).

This section has examined in brief the philosophical, cognitive and psychological perspectives on intention in speech production. For a full appreciation of the debate in the field, see Smith (1982) and a peer-reviewed commentary in The Behavioural and Brain Sciences (Sperber & Wilson, 1987) .

Why discuss the issue of mutual knowledge or mutual manifestness in a thesis about disfluency in dialogue? Since the Strategic-Modelling view, largely proposed by Clark and colleagues, outlined at the beginning of this thesis argues that speakers use disfluency in an intentional way, they are in a sense proposing that a disfluency is akin to a special sort of speech act. Clark (1996) and Fox Tree and Clark (1997) certainly suggest that a listener will have access to enough mutual knowledge to be able to derive the speaker's intended meaning when the speaker utters *thee….uh the*. If mutual knowledge does involve this indefinite regression of states as Schiffer (1972) proposed, then much explanation is necessary by proponents of mutual knowledge to explain how a speaker is capable of planning the intention first and foremost in a rapid disfluency and secondly in normal fluent speech. I would tend to argue that the predictions of Sperber and Wilson are closer to the mark: cognitive environments can be mutually manifest where interlocutors have access to information about these environments but that this information is not required for planning in speech production.

In his studies of dialogue, Clark infers that his speakers are using disfluency in an intentional manner and furthermore, he infers the intentions of his speakers. Likewise, Sperber and Wilson do not explain how one could test empirically whether something is mutually manifest, therefore making it as intractable to the experimental purpose of this thesis as testing intentionality. Brennan and Schober (2003) suggest that a speaker's intentions are best reached by conducting online experiments where the experimenter performs a role. For this reason, I will investigate the function of disfluencies by first constraining the possible intentions that a speaker could have had by asking them to participate in an online experiment. Later, I will contrast the possible functions of structural disfluency types when speakers face different levels of difficulty and are more motivated to perform well.

To summarise this section, Strategic-Modelling View proposes that speakers employ disfluencies as intentional signals when they encounter difficulty in speech production and that listeners are capable of detecting and correctly processing these disfluencies. Section 2.3.2 showed how speech technologists typically employ a number of criteria in order to detect disfluencies. In terms of prosodic cues to disfluency, section 2.3.2 reviewed that some prosodic cues are not used only to signal disfluent speech but also fluent speech. Section 2.3.3 showed that listeners are capable of detecting a disfluency once it has begun. In terms of processing disfluency, a number of studies have shown that a listener may simply benefit from extra time presented when a speaker pauses or is disfluent and not necessarily a phonological signal in the *um* or *uh*. Finally, Section 2.3.4 discussed intentionality and what is denoted by this notion in dialogue: in order for something to be an intentional signal, interlocutors must be engaged in a joint action in which they pursue joint closure. The speaker must design the disfluency according to his or her model of the listener and the listener must understand the meta-communicative intent of the disfluency. In the next section, I will review the Cognitive Burden View which argues that Strategic-Modelling View is unnecessarily demanding on what is required in dialogue.

2.3.5 Cognitive Burden View

The Strategic-Modelling account states that speech production is taxing but asserts that speakers are still capable of using disfluencies as signals of difficulty to the listener. The Cognitive Burden view states that speech production is cognitively burdensome and as a result of this, speakers are unable to model the listener throughout the entire dialogue so a disfluency is an actual error, not an intentional signal. According to the Cognitive Burden view, a speaker not only has to decide what needs to be said, he must also plan the utterance in a syntactically correct way whilst attending to what the listener is likely to need to know. Because speech production is such a complicated process, Brown and Dell (1987) argue, the production system apportions its resources to avoid over-burdening itself. The Dual-Process Hypothesis distinguishes between costly and inexpensive production processes: low cost processes include the sort of rapid response processes like priming whilst more costly processes cover the slower processes like reasoning. Listener modelling is thought of as particularly taxing because the speaker must constantly update the listener's model (Bard & Aylett, 2001). For example, a speaker might need to keep track of what the listener can see, hear or otherwise has access to during a conversation and keep this in mind when designing his utterance. For this reason Keysar et al. (2000), Barr and Keysar (2002) and Bard and Aylett (2001) argue that listener-modelling is a cognitively taxing process. It puts an additional and possibly unnecessary burden on the production system. Rather, the amount of modelling between listener and speaker that occurs during conversation depends on the available resources (Brown & Dell, 1987). Moreover, as Bard and Aylett (2001) and Barr and Keysar (2002) argue, speakers may not even need to model their listeners in order to have a successful dialogue. Basic set-theory logic guarantees that mutual knowledge in dialogue is also knowledge that the speaker alone knows (Barr & Keysar, 2002). It could then be the case that speakers often formulate referring expressions on the basis of purely egocentric knowledge and hence do not need to utilise a listener model at all according to the Dual Process Hypothesis of Bard and Aylett (2001).

Collaborative theories of discourse wishing to prove the importance of mutual knowledge should also show that there are occasions on which speakers refrain from using knowledge that is not mutual (Barr & Keysar, 2002). By so doing, one proves that the speaker relies exclusively on mutual knowledge. Furthermore, a collaborative theory of dialogue should also examine listener expectation, as Barr and Keysar set out to do in a series of experiments involving referring expressions and their linguistic precedents. Listeners depended on linguistic precedents, or established forms of reference more often when the entity was either unmentioned or unconventional; such a result argues that listeners relied upon a precedent only because it was available. One could hypothesize from this that listeners resort to mutual knowledge only in periods of difficulty (Pickering & Garrod, 2004).

In an experiment that tested partner specificity (i.e. a listener's expectation that the speaker will use a referent if that referent is shared knowledge), speaker identity (i.e. whether their presence was strongly determined or entirely independent) and partner-independence in a referential communication task, Barr and Keysar (2002) found strong evidence that speakers do not rely on partner specificity as previously claimed by Brennan and Clark (1996). As with the Brennan and Clark experiment, participants in Barr and Keysar's experiment, i.e. Matchers were asked to move objects in an array around until the array matched that of their (confederate) partner, the Director. Barr and Keysar show that although Matchers are able to retrieve conceptual pacts established with a partner, they did not rely on this knowledge to establish the referent. Listeners heard instructions from two confederate directors, one Director who arrived on time and began the experiment and a second Director who the listener was led to believe had arrived late. Matchers were shown to expect linguistic precedents by fixating on the target object when the Director used a precedent (e.g. *carnation*) even when the listener was looking at basic-level objects (e.g. a car and a flower). This suggests that listeners expected precedents regardless of whether they had shared knowledge with the speaker and even when mention of such a precedent was over-informative. This suggests that contrary to the results obtained by Brennan and Clark (1996), there is no evidence that linguistic precedents were used in a partner-specific manner (Barr & Keysar, 2002; Pickering & Garrod, 2004). Participation in a dialogue does not *ipso facto* prove there should be less reliance on egocentric knowledge. One cannot deduce simply because speakers interact with their listeners that they therefore depend upon the collaborative representation without with the conversation is doomed to failure.

In a later experiment, Metzing and Brennan (2003) argue that Barr and Keysar (2002) failed to test the condition most likely to determine whether listeners retain partner-specific information. Metzing and Brennan tested what happened when the Director abandoned a conceptual pact in preference of a new referent. Their results show that listeners were slower to process new referents if the speaker had first abandoned a conceptual pact in the process of mentioning the new referent. Metzing and Brennan (2003) argue that this is strong evidence that listeners employ partner-specific information.

Bard, Anderson, Chen, Nicholson, and Havard (2004) report the results of the MONITOR Project, explained in further detail in Chapter 3. Subjects in the MONITOR Project participated in a map task experiment in which they received either visual, verbal or both visual and verbal feedback from a confederate listener. No feedback trials were used as controls. If speakers do design utterances according to audience design, then one would expect a speaker to attend to the listener's feedback at all times especially if feedback indicated that the listener had difficulties. The MONITOR Project simulated this situation by providing sequences of visually 'correct' or on-route and 'wrong' or divergent feedback. Positive and negative verbal feedback was also provided. A record of the speaker's genuine gaze at the visual feedback was kept. Bard et al. observed that speakers did not pay direct attention to their listeners who showed signs of difficulty. Instead, speakers tended to gaze at what was easiest for them: the correct feedback which they would have to look at anyway because the feedback hovered over the next landmark on the route. Speakers attended more often when negative verbal feedback indicated that the listener required assistance. On the basis of these results, Bard et al. conclude that speakers seemed to operate on the principle of *joint responsibility* (Carletta & Mellish, 1996; Clark & Wilkes-Gibbs, 1986). According to this view, interlocutors in dialogue share the responsibility for communicating in an effective manner. It is not solely the responsibility of the speaker to assist the listener; the listener must also reveal their need for help in a salient manner.

Bard and Aylett (2001) ask whether referring expressions and articulation cater to the knowledge of the listener by manipulating the 'givenness' of knowledge in a Map Task experiment, or whether the speaker or listener had encountered the object earlier. 'Newness' is used to describe entities that are novel to at least one interlocutor. Previous research has shown that both the syntactic and articulatory form of a referring expression simplify over time (Ariel, 1990; Bard et al., 2000; Bard et al., 2000; Gundel et al., 1993). In four experiments, Bard and Aylett controlled for the givenness of particular landmarks on a map. Either both the speaker and the listener have said, seen and heard a mention of a particular entity or some one of them lacks some part of that knowledge. Word duration, a concomitant of articulatory precision, suggests that a listener model is not consulted: the amount of reduction on repeated mention depended only on what the speaker had heard. However, referential form results demonstrate evidence of listener modelling: referring expressions failed to simplify during mention to new listeners even though the referent was given for the speaker, unlike word duration which did reduce on repetition even if the listener was new. Bell et al. (2003) found that reduction of word duration could be affected by utterance position, predictability of the word and whether the word was disfluent or not.

Horton and Gerrig (1996) argue that listener modelling does not occur simultaneously with utterance planning, a process they call 'Initial Design' but rather during the monitoring phase when speakers assess whether the utterance was correct. This process is termed the 'Monitoring and Adjustment Model' by Horton and Gerrig (1996). Support was found for this model in experiments testing whether speakers referred to objects (for example, picture cards with different sized circles on them) in a different manner when speakers and listeners either shared or did not share a context (i.e. had privileged contexts) for the object (Horton & Keysar, 1996). For example, in a shared condition, the speaker would describe a picture card with two circles on it, where the smaller circle was above the larger circle. In a 'privileged' condition, the listener would only see one of the circles and not be able to determine the referent "small circle" in this instance. Results showed speakers who had no time-limit relied more on the shared context in planning their utterances; speakers who were under time-pressure, however, relied on both the shared and the privileged contexts to the same degree. As Horton and Gerrig argue the Initial Design or Monitor and Adjust Model both account for this data. The Initial Design Model would argue that speakers designed their utterances to consider the common ground and thus rely on the context more when it was shared rather than privileged. The Monitor and Adjust Model would describe these data by arguing that common ground is used only when monitoring and correction have occurred. The Monitor and Adjust Model is attested only as an effect of the time-limited case, however, while the Initial Design Model argues that time-pressure should not matter. Thus, Horton and Gerrig (1996) conclude that the Monitoring and Adjustment Model makes the best predictions and that speakers working under a time-limit treated the shared versus the privileged contexts equally because their initial utterance plan did not take common ground into account; common ground was brought to bear only later, during the monitoring process.

The Initial Design Model predicts that speakers access a listener model when designing utterances for an audience. This contrasts with the Monitoring and Adjustment Model which says that listener modelling is a cognitively costly process and provides convincing evidence to show that speakers are not necessarily modelling their listeners, although it may appear that they are.

So, the Initial Design Model predicts that speakers rely on a model of the listener and the Monitoring and Adjustment Model predicts that they do not. Although not entirely the same, the Strategic-Modelling view has similarities with the Initial Design Model in that both suggest that speakers rely on a model of the listener during speech production. The Initial Design Model differs from the Strategic-Modelling view in that the Initial Design Model makes no predictions about whether speakers use disfluencies as strategic signals. Likewise, the Cognitive Burden view has similarities with the Monitoring and Adjust Model: both models predict that speakers only rely on common ground knowledge in order to make corrections. The strong version of the Cognitive Burden view differs from the Monitoring and Adjust Model by making a more strigent claim that speakers never need to rely on listener knowledge during collaboration because anything that is deemed common ground is also known only to the speaker (Barr & Keysar, 2002).

Horton and Gerrig (1996) raise an interesting point: it may appear as though the speaker is modelling the listener when in fact he or she really is not. I propose to address this issue in this thesis by comparing the predictions of the Strategic-Modelling view where speakers gaze at a listener throughout a dialogue whilst also planning pertinent utterances that cater to the listener's needs with the predictions of the Cognitive-Burden Hypothesis which suggests that glances at a listener's feedback will be tempered by the availability of cognitive resources and that difficulty will arise when these resources are not available.

2.3.6 Disfluency as Difficulty

Section 2.3.5 outlined the hypotheses of the Cognitive Burden View. According to this view, a speaker does not need to rely on a model of the listener at any point because the speaker has his or her own model of the conversation. Furthermore, modelling a listener during a conversation is a taxing, burdensome process for the speaker who is also engaged in language production. I have shown in section 2.3.1 that the Strategic-Modelling View proposes that disfluency is a signal to a listener. In contrast to this position, the Cognitive Burden View argues that disfluency is an indication of difficulty. This section will review the literature which supports the disfluency as difficulty view.

Bard et al. (2001) conducted an empirical observation on data from the HCRC Map Task corpus to determine whether disfluency was induced by difficulty. As measures of difficulty, Bard et al. analysed task difficulty, interpersonal factors, order effects and effects such as length and the number of referring expressions from the prior utterance and effects of the current move (i.e. length, number of referring expressions, and proximity to the nearest conversational boundary). Their dependent variable was the number of disfluencies per Conversational Move. A Conversational Move is a sub-goal of the dialogue, i.e. an instruction, question or reply (see Carletta et al., 1997). Bard et al. predict that if disfluency is affected by difficulty, then there should be noticeable similarities between Inter-Move Interval (IMI), or the time between the offset of the speaker's utterance and the other speaker's reply, and disfluency. Bard et al. (2000) found that IMI tended to be longer earlier in the dialogue or at the beginning of an utterance when the speaker had a greater planning burden. IMI can also be affected by interpersonal factors like the amount of familiarity between speakers or whether they can make eye contact. Bard et al. found that disfluency behaved quite differently from IMI: a multiple regression analysis revealed that disfluency was only affected by production processes such as length of the Move, referential complexity and the conversational role in the dialogue. Previous research by Clark and Wasow (1998) and Oviatt (1995) also observed higher disfluency rates in longer utterances. Unlike IMI, disfluency was not affected by interpersonal factors. Somewhat surprisingly task difficulty (i.e. IMI) was not a significant predictor of disfluency. Bard et al. suggest that perhaps this indicates "a separation rather than sharing of processes" (Bard et al., 2001, p. 100).

Oviatt (1995) compared two measures of task difficulty to determine whether disfluency was associated with planning difficulty. The first measure was the length of the utterance in words. The second measure was the structure of the task: subjects received either a constrained task, in which the speaker had relatively little to prepare or an unconstrained task, in which the speaker had to plan more of their utterance. In the constrained task, the speaker saw a screen which gave them detailed instructions about what to say or do. It was predicted that more disfluencies would occur in the unconstrained task and that disfluency would be an indicator of difficulty. Results showed that utterance length was a clear predictor of disfluency: 77% of the variance for the rate of disfluencies could be predicted simply by knowing the utterance length. Oviatt controlled for utterance length when determining whether task format (constrained vs. unconstrained) was a predictor. She found that even with the control for length, disfluencies were more common in the unconstrained task format. In fact, 70% of disfluencies could be avoided simply by asking subjects to use a constrained task format. Thus, length of an utterance and asking a speaker to speak extemporaneously seem to be clear predictors of disfluency.

Section 2.3.1 and 2.3.2 found that according to the Strategic-Modelling view disfluency can be indicative of a speaker's intention to signal commitment to the listener. In this section, I reported results which support the Cognitive Burden View to suggest that disfluency is an indication of difficulty and cognitive load. These issues are at the very heart of this thesis. Can disfluency fulfil both of these roles at once? In the next section, I will outline proposals for a middle ground and how disfluency would be expected to pattern according to this view. In subsequent chapters, I will test these predictions and describe the results and implications.

2.3.7 Can there be a Middle Ground?

In this section, I have reviewed the literature on two theories of collaborative dialogue, the Strategic-Modelling and the Cognitive Burden Views. These views present opposing ideas regarding the nature of the speaker's responsibilities and capabilities during a dialogue. The strong version of the Strategic-Modelling View argues that speakers design even disfluencies as strategic signals for their listeners by referring to a listener model while the Cognitive Burden View suggests that speakers do not need to rely on a listener model nor are they always cognitively capable of doing so during dialogue. In their review of these theories, Schober and Brennan (2003) present a middle ground between two extremes, entirely altruistic modelling on the one hand and purely egocentric motivation on the other.

> "The evidence so far suggests that adaptation doesn't seem to be an all-or-nothing phenomenon at any level: people can be shown to adapt under some circumstances and not to adapt under others at virtually every level of language use –from higher discourse-level functions to articulation. Thus we propose, the more fruitful research agenda is to explore the factors that affect conversationalists' adaptations in particular circumstances – the sorts of tasks, individual ability differences, discourse goals, and so on that affect when and how partners can adapt to each other." (Schober & Brennan, 2003, p. 155)

Furthermore, Brennan and Schober (2003) argue that in order to observe dialogue in its natural state one must conduct online experimental investigations, rather than simply relying on instances from a corpus of spontaneous speech. The reasoning behind this suggestion is that a researcher is like a third-party listener on any pre-recorded corpus conversation and will not have access to a speaker's intentions. A task-oriented experiment with a map to traverse or a parking lot to park in constrains the possible intentions that a speaker might have had and the pertinent facts that a speaker might know, making them more readily accessible to the experimenter. Of course, the experimenter can never be entirely certain of a speaker's intentions, since they are known only to the speaker. Furthermore, the experimenter cannot be sure that the same speaker would behave the same way unobserved but the experimenter can still assume that the speaker was using his or her normal language faculty when he or she participated in the experiment.

Chapter 1 and Section 2.3.1 explained the notion of audience design (Brennan & Lockridge, 2004; Clark, 1994; Clark & Carlson, 1982a, 1982b; Clark & Marshall, 1981; Clark & Wasow, 1998; Schober & Brennan, 2003). Studies of audience design, or the process of formulating a particular utterance on the basis of mutual knowledge (Brennan, 2004; Clark, 1996; Schober & Brennan, 2003), suggest that speakers and listeners have a joint responsibility for achieving a successful conversation. Joint responsibility means that each participant has different but equally important roles to play during the conversation (Carletta & Mellish, 1996). The speaker is not solely responsible for modelling the listener and adjusting his or her utterance contributions to the knowledge of the listener. Instead, it is the responsibility of the listener to indicate his or her knowledge and needs for clarification. Joint responsibility comes close to the Principle of Optimal Design (Clark, Schreuder, & Buttrick, 1983) and the Principle of Least Collaborative Effort (Clark & Wilkes-Gibbs, 1986), which suggest that partners in dialogue should find the most cost-efficient means of collaborating. In a map task dialogue, for example, it would be the responsibility of the speaker to communicate the route instructions in an understable fashion according to the common ground. The listener's responsibility, on the other hand, is to indicate when the instructions don't make sense or to ask for further clarification.

Pickering and Garrod (2004) proposed a model of interactive alignment for collaborative dialogue. This model is another possible Middle ground view of collaborative dialogue. According to this view, a speaker is thought to use a situation model, a model of the listener and his knowledge of the common ground when attempting to align with his 'audience' or listener; any time a speaker does this, he is said to be participating in audience design. Speakers are said to align with another when both speakers have the same representation at some linguistic level. During a dialogue, a speaker is proposed to have access to a 'situation model', or a representation of the interaction which encodes space, time, causality, intentionality, and reference to the main individuals or objects that are discussed (Johnson-Laird, 1983; Sanford & Garrod, 1981; van Dijk & Kintsch, 1983; Zwaan & Radvansky, 1998). The term 'situation model' has also been described for Semantics by Barwise & Cooper (1991) and Cooper (1992). Pickering and Garrod define 'interactive alignment' as alignment which is brought about in dialogue as the result of two speakers providing verbal feedback in an attempt to communicate effectively. Pickering and Garrod (2004) propose that during the course of a successful dialogue, speakers will develop aligned situation models, but that by no means is a fully aligned model necessary for communication to occur. Nor is feedback always necessary because alignment processes are automatic. Horton and Gerrig (1996) found that speakers and listeners do not always attend to information presented in feedback. Instead, speakers rely on their own knowledge of the situation. The main thrust of Pickering and Garrod's paper is to argue in favour of an interactive model of discourse processing in which speakers *align* with one another on a number of different levels of linguistic knowledge.

Pickering and Garrod (2004) argue that because speakers develop aligned situation models, there is no need for the speaker to 'model the listener' during dialogue. Previous research by Bock (1986) and Levelt and Kelter (1982) found that speakers show evidence of priming in monologue. Bock (1986) showed evidence of syntactic priming in monologue in an experimental setting. Levelt and Kelter (1982) found that when speakers were asked *What time do you close?* in Dutch, they tended to reply with a congruent answer like *Five o'clock*. Similarly, when they were asked *At what time do you close?* speakers preferred to respond *At five o'clock*. As evidence for their claim that speakers interactively align with one another on a number of linguistic levels, Pickering and Garrod cite Branigan, Pickering, and Cleland (2000), as having shown that there is clear evidence for syntactic priming or alignment on a syntactic level during dialogue. Speakers consistently responded with appropriately matched syntactic constructions to a confederate speaker's initial description during a picture matching task (Branigan et al., 2000). Subjects were shown a card with a picture of a robber and a ballerina in which the robber was handing a banana to the ballerina. English syntax permits two constructions to describe this action: either *The robber handing a banana to the ballerina* or *The robber handing the ballerina a banana*. A confederate participant presented one of these constructions to the naïve participant (*The robber handing a ballerina a banana*). Branigan et al. found that the naïve participants could be 'primed' to describe a subsequent card to their partner using the same syntactic construction that they had just heard (i.e. *The police officer handing the dog a horse*). Pickering and Garrod cite this as evidence that interlocutors align on a syntactic level with their partners.

Alignment also occurs on an articulatory level, as evidenced by Bard et al.'s (2000) study of repeated mentions of referring expressions. Though it is common for articulation to be reduced during second mention as shown by Fowler and Housum (1987), Bard et al. (2000) demonstrated that the amount of articulatory reduction was as acute whether the original speaker or the conversational partner a uttered the expression a second time. This evidence argues directly for a 'middle ground' between the extremes of Strategic-Modelling and the Cognitive Burden View.

As previously discussed (Section 2.3.1) interlocutors engaged in a dialogue to establish referring expressions, perspectives and a common ground in order to be sure that they are working towards a mutual understanding (Barr & Keysar, 2002; Brennan & Clark, 1996). However, it is less clear what sort of activity replaces interaction in monologue and what, if such a thing exists, the nature of the speaker's addressee model is like. As Barr and Keysar (2002) argue, it is not guaranteed that interaction in dialogue is an essential prerequisite.

Section 2.3.1 described the importance of common ground with respect to intentionality. The notion of common ground is also a crucial part of determining the differences between dialogue and monologue. In dialogue, one goal might ostensibly be to agree on what constitutes the common ground and model the amendments to this common ground over time. The goals of a monologue might simply be to inform an unknown audience (listening at an unknown time) about a particular state of being or topic. The goals of the separate speech modalities (i.e. dialogue and monologue) differ as do the method by which the goals are processed. A monologue speaker does not need to (in fact, cannot in some cases) respond to visual and verbal feedback from any members of the audience whereas a speaker engaged in a dialogue has this opportunity. Pickering and Garrod (2004) differentiate these goals from the goals of a dialogue speaker. Instead, the monologue speaker is forced to devise her production plan based on what knowledge she believes her audience to hold. Schober (1993) asked subjects to describe a spatial arrangement to an imaginary listener and that they could do anything necessary to get the task done. They were told nothing else about their imaginary listener. Evidence from Schober (1993) suggests that monologue speakers were more likely to take a listener's spatial perspective than they were their own frame of reference. However, this might be an artefact of Schober's experimental task: speakers were always told the spatial orientation of an imaginary listener and therefore had something specific to crosscheck against. It is as yet unclear what perspective a speaker without such specific listener information would take during a monologue.

Pickering and Garrod (2004) argue for an interactive dialogue account partially on the basis that when alignment of situation models does not occur, there is a frequent occurrence of repetition speech repairs. Instead of constantly modelling a listener at a high production cost, a notion presented as controversial in Section 2.3.6, speakers refer to a listener model only when the speaker suspects a misunderstanding. This is done in two stages: first the interlocutor checks to see whether the input matches his or her own representation and second, if this check fails, the interlocutor will reformulate the utterance in order to re-establish the common ground (Pickering & Garrod, 2004). The same process is argued to occur when the interlocutor is checking for whether incoming information is new or given. Pickering and Garrod propose that repairs of unaligned representations commence when the speaker checks the other interlocutor's understanding of the discourse and continue with a modification of the original utterance. Interactive alignment in dialogue is crucial at the point of the greatest difficulty and quite possibly worth the production costs. Schegloff et al. (1977) have proposed a similar notion as previously mentioned in Section 2.1.1. According to this view, either a speaker or a listener can initiate a repair but typically self-repair is the preferred method for both speakers and listeners because it requires fewer turns to complete and therefore less work by both participants.

In this sense, the Interactive Model of Alignment and the view of Joint Responsibility propose similar predictions as far as disfluencies are concerned. According to the Interactive Alignment Model, disfluency is symptomatic of an unaligned dialogue. By being disfluent, the speaker is fulfilling their responsibility to re-align and redirect the collaboration. Since these views make the same prediction for disfluencies, there is no need to test between them. I will henceforth refer to this notion at the Middle Ground view of disfluency.

The Middle Ground view of disfluency differs in some respects from the Cognitive Burden view in that the Cognitive Burden view suggests that disfluency implies the speaker encountered difficulty and an error occurred as a result. On the other hand, the two views make similar predictions in that neither predicts that disfluency is under the intentional control of the speaker. In this respect, the Cognitive Burden and Middle Ground view make similar predictions. Therefore, it is not necessary to test between all three models, namely the Strategic-Modelling, Cognitive Burden and the Middle Ground views. Rather, one can test between the Strategic-Modelling view and the Cognitive Burden view by devising a collaborative experiment that makes it possible to analyse the functions that structural disfluencies play during dialogue. If the speaker uses disfluency in a signalling function, then there is support for the Strategic-Modelling view. If the speaker uses disfluency during periods of difficulty, then this could be an indication of cognitive burden overload. If neither or both of these scenarios eventuate, the one has support for a view somewhere in the middle of these two views.

In order to determine the function of disfluency, the experimental paradigm must be as controlled and yet as interactive as possible. By incorporating eye-gaze tracking technology into the paradigm one has a time-stamped record of the speaker's gaze which can be cross-referenced with the speech record. The focus of the next section will be to review the findings of collaborative dialogue and eye-gaze experimentation.

2.4 Collaborative Dialogue and Eye-Gaze

Apart from the words and silences that occur during a dialogue, interlocutors have other sensory tools which they can use to learn things about their other interlocutor and the world around them. One such tool is eye-gaze. This section will review what is known about the use of eye-gaze in collaborative dialogue experiments. For a complete review of the state of language production and gaze research, consult Griffin (2004).

With the advent of eye-tracking technology a number of psycholinguistic studies have been run which tend to show that people tend to gaze at the objects they talk about (Griffin, 2004; Meyer, Sleiderink & Levelt, 1998) or consider talking about (Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995; Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995). Moreover, gaze is a powerful social tool since people have been shown to be sensitive to the direction of another person's gaze (Langton, Watt, & Bruce, 2000) and tends to take this as evidence of what the other person is thinking about (e.g. Goodwin, 1981).

Typically, psycholinguistic studies have employed the eye-tracking techniques in studies known as the 'visual world paradigm'. The visual world paradigm typically involves an array of ordinary objects placed in a grid and an eye-tracker which keeps a record of the participant's gaze. The visual world paradigm has been used to study the form of referring expressions (Brown-Schmidt & Tanenhaus, In Press; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), and object naming (Griffin, 2005; Griffin & Bock, 2000). In studies of dialogue, the visual world paradigm has been used to test theories of common knowledge and audience design (Brown-Schmidt & Tanenhaus, In Press). Dialogue studies involve 2 interlocutors, a Director and a Matcher, whose goal it is to move the objects to different locations in the grid. Typically, one or more items may be visible only to the Director of the task.

According to the Monitoring and Adjustment view put forth by Keysar, Barr, and colleagues, speakers are egocentric from time to time when planning a referring expression in a dialogue with another individual (Hanna, Tannenhaus, & Trueswell, 2004; Horton & Keysar, 1996; Keysar, Barr, Balin, & Brauner, 2000). Evidence from face-to-face conversations suggests that speakers avert their gaze when engaged in cognitively taxing processes (Glenberg, Schroeder, & Robertson, 1998). Further research by Keysar et al. (2000) used the visual world paradigm in a referential communication task where the subject had to retrieve a named object. Speakers gazed at privileged objects visible only to them when considering possible references made by other participants. Occasionally, participants even grabbed for an occluded object, thus committing an error. Keysar et al. (2000) suggest that while there is a cost associated with taking the egocentric perspective (because this sometimes led participants to grab the wrong and occluded referent), there is also a cognitive cost associated with using mutual knowledge. Thus, there was a trade-off between the cost of taking an egocentric perspective and possibly choosing the wrong object and the cost of constantly modelling the perspective of the other participant. Keysar et al. (2000) conclude that occasionally speakers find the cost of mutual knowledge too high and instead take a risk and consider an egocentric perspective. Hanna et al. (2003) ran a similar study which in part replicated the findings of Keysar et al. (2000) in finding that speakers did not tend to ignore salient objects even though these objects were occluded from the other participant's view. Hanna et al. (2003) suggest that while participants do employ egocentric knowledge, they failed to find evidence that participants ever ignored knowledge of the common ground as predicted by the Monitoring and Adjustment view. For this reason, Hanna et al. (2003) propose that perhaps viewing common ground as a constraint-based model is the most parsimonious approach. According to their view, while participants are capable of using their cognitive resources to monitor another participant closely, this strategy may not be the ordinary one.

Alternatively, Clark and Krych (2004) provide a "bilateral account" of collaborative dialogue, suggesting that speakers do indeed monitor other participants during interactive dialogue. In this view, bilateral accounts of dialogue in which speakers complete a joint act, contrast with "unilateral" accounts in which speaking and listening are argued to be autonomous processes. Clark and colleagues note that speakers monitor their listeners' faces (Bavelas, Black, Lemery, & Mullett, 1986) and gaze direction (Argyle & Cook, 1976; Goodwin, 1981; Langton, Watt, & Bruce, 2000) in order to establish common ground (Clark & Krych, 2004; Clark, 1996). Clark and Krych set out to determine how speakers monitor listeners for both vocal and gestural information. They predict that speakers who cannot monitor listeners "should make more errors[2], take longer or both" than speakers who can monitor listeners. In order to test this prediction, Clark and Krych recorded participants who were engaged interactively in a Lego assembling task. The videos of their joint actions were then analysed for instances of gestural and deictic references both when participants could see each other's faces and when they could not. Since participants were interacting face-to-face, it was not possible to track their gaze. Results showed that although participants did attend to gaze and head gestures, the overall success of the task did not depend on these cues. An increase in performance was found if speakers could see the listener's workspace. Speakers used more deictic references (e.g. *this, that, here, there*) if the workspace was visible than if it was hidden from view.

In terms of hand gestures, Clark and Krych identified four distinct types of gestures: exhibiting, manifest actions, postponement, and negative manifest actions. The participant who was assigned to build the Lego model would often hold up the block in an exhibit gesture. A

8/5/078/5/07

[2] Here Clark and Krych seem to refer to actual errors, not speech errors or disfluencies.

manifest action occurred when the builder positioned a block or somehow moved the block in a way that was visible to the speaker. In a postponement, the builder held the block in mid-movement and often waited for confirmation from the speaker that s/he was moving the correct block. Finally, a negative manifest action is when the builder detaches a block that they've placed, perhaps incorrectly, onto the model. Results showed that most of the gestures were also jointly construed as signals between speaker and builder. Clark and Krych conclude that conversation is a jointly orchestrated act in which conversationalists attend to a host of gestural, vocal and facial cues in order to ground the utterances of the other speaker.

Argyle (1990) has argued that speakers tend to look no more than 50% of the time at their partner during conversation. Kendon (1967) showed that speakers gazed for short periods of time at a listener while they were speaking. Gaze levels did not rise about 22% of total conversation time in Kendon's study. Similarly, Watts and Monk (1996) found that participants in a video link conversation gazed at the listener less than 25% of the time. Anderson, Bard, Sotillo, Newlands, & Doherty-Sneddon (1997) found that subjects who participated in a map task experiment matched the mutual gaze levels of other studies (Argyle, 1990; Argyle & Cook, 1976; Kendon, 1967): subjects made mutual gaze with their listeners on roughly 2.7% of all words. Furthermore, Anderson et al. (1997) suggest that having access to another person's mutual gaze may have had a detrimental effect on task performance. They conducted a measure of intelligibility, defined here as "the proportion of listeners able to identify the word token correctly over all the experiments in which it was used", in a condition with and without gaze (Anderson et al., 1997, p. 588). Anderson et al. examined whether intelligibility was changed between first and later mentions of the same landmark name (e.g. *site of the forest fire)*. Intelligibility loss, or the proportion of correct identifications when the speaker changes from the citation form to a spontaneous instance of the same word, in the condition without gaze was 10%, a significantly smaller figure compared to the intelligibility loss for the condition with gaze (23%). From these results, Anderson et al. conclude in line with Argyle, Alkema, and Gilmour (1972), Argyle and Graham (1977) and Krantz, George, and Hursh (1983) that when an object is involved in conversation, subjects will spend more time looking at the object than they will making eye contact. The results from Anderson et al. also show that speakers pronounce spontaneous words much less intelligibly when a partner's face is visible.

Anderson, Bard, Dalzel-Job, & Havard (submitted) and Bard et al. (2004) report the results of gaze at a simulated visual feedback in the MONITOR Project. As explained in Chapter 1, Section 1.3 subsequent chapters of this thesis will report additional findings on disfluency and gaze from the MONITOR Project. The MONITOR Project used a simulation of a "listener's" gaze in a map task experiment. As with previous map task experiments (Anderson et al., 1991; Brown et al., 1983), subjects performed the role of an 'Instruction Giver'. In the MONITOR Project, subjects were asked to provide route descriptions on the map to an 'Instruction Follower', whose gaze in this case was actually represented by simulated visual feedback in the form of gaze. Anderson et al. (submitted) found that while speakers did respond to the visual cues given by the simulated feedback, speakers tended to avoid gazing at the visual feedback when it showed that the listener did not follow the route instructions (i.e. the listener went to a 'wrong' landmark). Subjects did not alter their gaze patterns at visual feedback when presented with added time-pressure. Instead, the addition of verbal feedback caused speakers to gaze more often at their listener as did increased task motivation to perform well.

Vertegaal and Ding (2002) tested whether subjects were more likely to speak in a) a 'sync' condition in which a partner synchronised their gaze at the speaker to co-occur with the speaker's turns or b) a random condition in which the partner gazed randomly at the speaker but with the same overall frequency as in the sync condition. Participants participated in a collaborative language puzzle task in which each subject was given a fragment of the same sentence (3 fragments made one sentence). Subjects were asked to collaborate to think of as many syntactically permissible permutations of the sentence as possible (6 correct answers for each sentence). Each permutation had to be grammatically correct, meaningful and subjects were not allowed to change the order of words within a sentence fragment. Vertegaal and Ding observed that while task performance was 46% higher in the sync condition, overall results showed that subjects were no more likely to speak when gaze was synchronised with their turns than when gaze was random. Subjects spoke more often when they received more gaze. Thus, Vertegaal and Ding conclude that models of conversational agents or avatars which employ random gaze will suffice in situations where task performance is not critical.

In a further study, Monk & Gale (2002) divide human gaze awareness into three separate groups. They use the term *Full gaze awareness* to denote a person's ability to discern what object another person is gazing at. This contrasts with *partial gaze awareness* or a person's ability to discern in what direction another person is gazing. *Mutual gaze awareness* is a person's ability to know when another person is looking at them in the eyes. Mutual gaze is only possible when both individuals make direct eye contact. Historically, mutual gaze has been difficult to achieve over video link conversations because of problems presented with the positioning of the camera. In order to make eye contact with the other person, an individual must gaze directly into the camera. If the individual does this, however, it appears to the other person that the individual is gazing at his or her abdomen because of the position of the camera. Buxton and Moran (1990) devised a method known as a ‘video tunnel’ to overcome this problem. A video tunnel uses half-silvered mirrors to put the camera in the same position as the monitor, thus making mutual gaze between the two participants possible. Monk and Gale developed a ‘full gaze awareness’ display which used a video tunnel arrangement to provide mutual gaze but also provided both participants with an actual size version of the object. As a control, subjects also experienced a ‘video-tunnel only’ condition and an ‘audio only’ condition. In the video-tunnel only condition, subjects could monitor each other’s faces, make mutual gaze and see a version of the receiver’s display but they only saw a reduced version of an expert’s display. In the audio only condition, subjects only heard verbal feedback from their partner; they did not have full gaze awareness or the ability to make mutual gaze. Results suggested that it was less important for participants to have mutual gaze and more important that participants see what they were meant to be discussing. Monk and Gale observed a reduction in the number of turns required to complete the task in the full gaze awareness display condition compared to the other two controls. This corroborates previous results found by Doherty-Sneddon et al. (1997). Doherty-Sneddon et al. tested participants in a video-mediated condition that enabled mutual gaze and a condition that did not. When mutual gaze was available, participants made more turns and overlaps than in the control condition.

Thus, according to both Monk and Gale (2002) and Doherty-Sneddon et al. (1997), there is substantial evidence that participants do not benefit from mutual gaze in video-mediated conversations. Instead it seems to be the case that participants tended to gaze at the object of discussion rather than at the other person in a face-to-face situation (Anderson et al., submitted; Anderson et al., 1997; Argyle et al., 1972; Argyle & Graham, 1977; Krantz et al., 1983). Face-to-face studies of conversation have suggested that while gaze is an important tool in conversation, speakers do not monitor their listeners any more than 50% of the time (Argyle, 1990; Kendon, 1967; Watts & Monk, 1996). This speaks directly to the predictions of the Strategic-Modelling View which suggest that speakers will constantly monitor their speakers for signs of uptake during conversation. These results also suggested that one does not necessarily need an experimental paradigm which supports mutual gaze in order to guarantee effective dialogue because a) subjects tend to gaze at the object in task-oriented dialogues more often than at their partner and b) subjects only make limited use of mutual gaze and c) some studies have observed that subjects “over-gaze” when provided with the novelty of mutual gaze in a video-mediated conversation (Doherty-Sneddon et al., 1997).

2.5 Disfluency and Gaze

Studies of collaborative dialogue have used eye-tracking techniques to study the relationship between gaze and disfluency. Such studies have focussed on listeners' perception of disfluent speech and what may be occurring in the speaker's mind when a disfluent utterance is produced.

2.5.1 Gaze in Perceptual studies of Disfluency

For fluent speech, Dahan, Tanenhaus, and Chambers (2002) observed that pitch accents used to signal new from given referring expressions can be used in real-time processing. Dahan et al. showed that subjects were sensitive to the discourse status (focussed or non-focussed) as well as mention (new vs. given) when interpreting de-accented or accented referents. De-accented referents were associated with previously mentioned and focussed entity and accented referents were associated with previously mentioned but unfocussed entities. For disfluent speech, gaze research has shown a relationship between the time course of disfluent speech and listeners' gaze immediately following a filled pause (Arnold, Altmann, & Tanenhaus, 2003; Arnold, Fagnano, & Tanenhaus, 2003; Arnold, Tanenhaus, Altmann, & Fagnano, 2004). Arnold et al. (2003) found that during comprehension of disfluent speech listeners looked at old information following a fluent introduction, thus providing support for previous research (Dahan et al., 2002). Following disfluent speech (i.e. a filled pause), listeners tended to gaze more at new objects in a visual display (Arnold et al., 2003; Arnold et al., 2004). Arnold and colleagues suggest that these results show that disfluency is used as a cue to signal new information. Previous work has shown that speakers may signal the difficulty of uttering an upcoming constituent by inserting a filled pause (Clark & Fox Tree, 2002; Fox Tree & Clark, 1997). Arnold et al. (2003) and Arnold et al. (2004) hypothesise that speakers signal the difficulty of uttering the name of a new object (*the salt shaker*), as compared to a given object (*the grapes*), in the same way, by inserting a filled pause. Since listeners gazed more at new objects, as compared to given objects, following a filled pause, this suggests that listeners are sensitive to this information and can utilize it while processing an utterance (Arnold et al., 2004). Arnold et al. propose this prediction based on the fact that speakers did tend to insert a filled pause more often prior to the name of a new object (*thee uh candle*) (Arnold et al., 2003) and based on the reasoning that a speaker will require more lexical search time in order to name a new object than a given one.

We have already reviewed the results found by Bailey and Ferreira (2005) in Section 2.3.3:

listeners who hear a filled pause prior to an ambiguous garden path are biased towards looking at the target object because a filled pause prior to a noun phrase is likely to be interpreted as the subject of the clause rather than the object of the clause. Such a result is similar to the results found by Arnold et al. (2003). When the listener heard a filled pause before *towel* (*Put the frog on [the uh towel] in the box*), the speaker looked at the target object (i.e. the frog) later, indicating that s/he was entertaining the modified goal reading (Bailey and Ferreira, 2005). Thus, the online time course of ambiguous speech processing seems to be sensitive to what a speaker might intend by uttering a filled pause.

The perceptual studies of disfluency reviewed here and in Section 2.3.3 suggest that listeners are capable of processing disfluencies as an indication of difficulty (Arnold et al., 2003, 2004; Bailey & Ferreira, 2003a, 2005; Brennan & Schober, 2001; Brennan, 2004; MacGregor et al., 2005). These results are in line with the Strategic-Modelling View which suggests that speakers use disfluencies as signals. Most of the studies who suggest that listeners are capable of using disfluencies as signals only tested filled pauses or prolongations as disfluent stimuli and did not test genuine disfluencies 'in the wild' (i.e. repetitions, deletions and substitutions). Thus, it could be the case that listeners would find processing complicated disfluencies like repetitions and deletions more difficult as studies like Fox Tree (1995) and Lickley (1995) seem to indicate for false starts. Moreover, as previously mentioned, a number of studies have suggested that listeners benefit only from additional processing time and not from the actual phonological form of a disfluency (Bailey & Ferreira, 2001; Brennan & Schober, 2001; Watanabe et al., 2005).

2.5.2 Gaze in Production studies of Disfluency

Although the results from studies of listener perspective tend to favour the Strategic-Modelling view, results of speakers' gaze during the production of disfluent speech suggest an association with cognitive difficulty. Gaze aversion in face-to-face dialogues during periods of difficulty is a well-documented phenomenon (Anderson et al., 1997; Glenberg et al., 1998; Griffin, 2005; Griffin & Bock, 2000). Female speakers who had access to their partner's eye contact in a map task were found to be less disfluent than when the speakers could not make eye contact with their partners (Branigan, Lickley, & McKelvie, 1999). Branigan et al. (1999) argue that since females are generally more socially aware, the absence of eye contact incurred great difficulty, thus causing disfluency.

Griffin (2005) found that when a speaker makes a substitution error, the speaker tends to look at rather than talk about the objects they intend to speak about. As Griffin (2005) shows, speakers gaze at items that they name correctly for the same amount of time as they look at objects they name disfluently. This result is important because it shows that errors stem from purely linguistic problems rather than the extralinguistic factor of how much time the speaker gazed at the object prior to naming it. Speakers have been found to look at objects for less time before fluently uttering the name of the object than before disfluently naming an object (Griffin & Bock, 2000). Speakers also tend to gesture more often when disfluent (Beattie & Shovelton, 2003; Kendon, 1967). Beattie and Shovelton (2003) specifically found that hand gestures seemed to coincide with gaze aversion, disfluency and mental effort. Although one might predict that disfluent regions of speech would coincide with gestures in the eye region (i.e. blinks, eyebrow raises, changes in direction of gaze), so far there is no support for this prediction (Yasnik, Shattuck-Hufnagel, & Veilleux, 2005).

Other work done in the area of gaze and disfluency production suggests that disfluencies have a strategic role in buying the speaker time (Brown-Schmidt & Tanenhaus, In Press). Speakers who didn't notice the size contrast between large and small objects were more disfluent, suggesting that disfluency was one way to buy the speaker additional time. This is in line with other work suggesting that disfluencies are associated with planning difficulties (Gregory et al., 2003). Gregory et al. asked subjects to move items which contrasted in colour, material or scalar property. Only the eye-gaze of the speaker was recorded. Gregory et al. found that subjects produced more disfluencies (e.g. silent or filled pauses and repetitions) prior to or following a scalar adjective than they did surrounding a colour or material adjective. Since subjects had to compare objects in order to produce a scalar adjective (i.e. to denote a size contrast), Gregory et al. argue that scalar adjectives are more difficult to produce than colour or material adjectives, and therefore that disfluencies are associated with production difficulty. Both Brown-Schmidt and Tanenhaus and Gregory et al. used pairs of naïve subjects and only recorded eye-gaze of one subject (i.e. the speaker). Subjects did not have gaze awareness of where their partner was looking. If speakers had had gaze awareness, they would have been aware of the fact that their partner had or had not noticed the size contrast, and thus, this knowledge might have affected the speaker's tendency to produce disfluent referents when a size contrast was present.

2.6 Summary of Literature Review

As has been demonstrated in this chapter, studies of collaborative dialogue have dedicated abundant research to common ground, spatial and conceptual perspective-taking, referring expressions and disfluency (Bard & Aylett, 2001; Clark & Wasow, 1998; Lickley, 1994; Schober, 1993). Two competing psycholinguistic hypotheses have emerged. One view explains interaction in dialogue as an intentionally strategic process during which the speaker updates a model of the listener (Brennan & Clark, 1996; Clark & Wasow, 1998; Clark & Wilkes-Gibbs, 1986; Lickley, 1994; Schegloff et al., 1977; Schober, 1993). The other hypothesis argues that speech production is a cognitively burdensome task and that the amount of interaction that speakers can engage in is decided based on the available resources (Bard et al., 2000; Bard & Aylett, 2000; Brown & Dell, 1987).

According to the proponents of the Strategic-Modelling View, an ideal speaker will establish conceptual pacts with their listeners (Brennan & Clark, 1996) and adopt the spatial-perspective of their interlocutor (Schober, 1993). An ideal speaker will also signal their commitment to both the utterance and the listener during a repetitive repair (Clark & Wasow, 1998; Fox Tree & Clark, 1997). Clark and Krych (2004) suggest that attentive speakers will monitor their interlocutors' faces (Bavelas et al., 1986) and direction of gaze (Gale & Monk, 2000) in order to establish common ground. The speaker will be capable of doing all of this whilst planning the upcoming utterance (Clark, 1996).

The Strategic-Modelling View predicts that listeners will be able to reliably detect disfluency in spontaneous speech during a dialogue in order to employ them as signals. In Section 2.3.2, I reviewed the literature on disfluency detection in speech technology and phonetics in order to determine whether there are any reliable cues to disfluency. Word duration, fundamental frequency, pause duration, glottalisation and coarticulation are all potential cues to signal the presence of an error (Eklund, 2001, 2004; Howell & Young, 1991; Lickley, 1994; Nakatani & Hirschberg, 1994; Plauché & Shriberg, 1999; Shriberg, 1994, 1999). The problem remains that each potential prosodic cue is not restricted only to disfluent speech. Prolongation (eg. segment lengthening) is a common process before a syntactic boundary (Local, Kelly, & Wells, 1986; Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992). Speakers may routinely glottalise a segment prevocalically or when their air supply is low in fluent speech (Lickley, 1996). Fundamental frequency, which is an inconsistent cue to disfluency, exhibits falls, rises and resets in fluent speech as well. Over the course of a fluent utterance, the pitch level commonly decreases in the downtrend and downstep phenomena (Liberman & Pierrehumbert, 1984). However, when listeners were put to the task of predicting whether disfluency was about to occur in a word-gating task, their performance suggested that they could not reliably predict an upcoming disfluency (Lickley, 1994).

Within psycholinguistics, considerable research has been dedicated to determining whether hearing a disfluency affects a listener's language comprehension in any measurable way (Arnold et al., 2003; Arnold et al., 2003; Arnold et al., 2004; Bailey & Ferreira, 2001, 2003a, 2003b, 2005; Lickley, 1994, 1995, 1996; Lickley, McKelvie, & Bard, 1999; Lickley et al., 1991; Nicholson et al., 2005; Nicholson et al., 2003). After hearing a filled pause, listeners exhibited sensitivity by gazing earlier at target objects (Bailey, 2003b) or at new objects (Arnold et al., 2003) or by showing a reduced ERP affect after hearing an unpredictable word (MacGregor et al., 2005). While it is certain that prosodic cues exist and that listeners are capable of utilising prosodic cues in both fluent and disfluent speech, it is less clear that listeners do this all the time or that they can access what the intentions of the speaker might be (Lickley, 1994; Snedeker & Trueswell, 2003). Bailey and Ferreira (2001), Brennan and Schober (2001) and Watanabe et al. (2005) show that there did not seem to be anything particular about the phonological form of a disfluency which aided listeners: listeners seemed to be sensitive only to the fact that there was a delay. Thus, it is uncertain whether listeners can really utilise disfluencies as signals as the Strategic-Modelling view suggests.

The alternative view, the Cognitive Burden View argues that an ideal speaker will avoid attending to information when the cognitive cost of attention is high (Horton & Gerrig, 2005; Horton & Keysar, 1996). Keysar et al. (2000) found that speakers occasionally opted for an egocentric perspective, when the demands of gazing at mutual information were too high. In terms of responsiveness, speakers are predicted to respond according to their own needs and not as the result of modelling their listener. Evidence for this claim comes from Anderson et al. (1997) who demonstrated that speakers who had the ability to make eye contact during a map task did so rarely. As reported in Section 2.4, studies of gaze in object-oriented dialogue have found that when an object pertinent to the task is present, speakers will look at the object more often than at the person. In terms of disfluency, proponents of the Cognitive Burden view argue that disfluency is classifiable simply as the output error of an overburdened system (Bard et al., 2003; Bard et al., 2001; Nicholson et al., 2005; Nicholson et al., 2003). Therefore, an ideal speaker according to the cognitive burden view will avoid gazing at the interlocutor when information is costly and will only respond when cognitive resources permit.

Barr and Keysar (2002) argue that it is insufficient to provide only positive evidence of interaction when studying collaborative dialogue. Studies should also attempt to give negative evidence to show when certain processes (eg. Inference on the basis of either mutual or individual knowledge) do not take place. As such, it seems that there is much to be learned about the potentially intentional processes occurring in dialogue by actively engaging in a comparative study of monologue. Pickering and Garrod (2004) point out some of the ways in which monologue differs from dialogue in their proposal of an interactive model of alignment. According to this model, it is only necessary to model the listener when the alignment process has been derailed. Interactive repair alignment occurs automatically out of the overlap in knowledge between speaker and listener (Pickering & Garrod, 2004). In monologue, however the solo speaker is unable to align to with an interlocutor and as such has no recourse to develop a routine of checking that the message is being understood. Schober (1993) reports that monologue speakers tended to adopt the perspective of their hypothetical listener but that this process was costly. Thus, it could be the case that speakers only implement costly inferential processes in times of difficulty (Pickering & Garrod, 2004).

A variety of research has been dedicated to common ground, referring expressions, disfluency and how such things pattern in collaborative human dialogues. Of these, disfluency is an interesting avenue for further research because it is unclear whether the speaker intentionally caused the disfluency to happen or whether it was simply an error of an overburdened system. For this reason, the function of disfluency is a valuable metric when testing two competing hypotheses within psycholinguistics, the strategic-modelling view on the one had and the Cognitive Burden View on the other. A substantial amount of research has investigated whether listeners respond to filled pauses as a cue (Arnold et al., 2003; Arnold et al., 2003; Arnold et al., 2004; Bailey & Ferreira, 2001, 2003a, 2003b, 2005). However, as it is still unclear what the intentions of the speaker might be (Lickley, 1994; Snedeker & Trueswell, 2003), more effort should be directed at observing the speaker and the instances in which she or he is disfluent during a collaborative dialogue task.

Accordingly, I propose to investigate why disfluency occurs, that is, when it occurs and for what reason. What other dialogue behaviours or gaze patterns did the speaker exhibit when she was disfluent? If one knows what the speaker was gazing at during disfluency, one can determine whether this situation induced disfluency because of its difficult nature or whether the speaker could have used disfluency as a signal. Knowledge about the type of utterance that caused the speaker to be disfluent is important because it tells us what sort of dialogue goal the speaker was trying to fulfil when s/he became disfluent. Was the speaker more disfluent when giving simple route instructions or when attempting to interact directly with the listener by providing a clarification or acknowledgment?

In order to specifically address the outstanding questions between the Strategic-Modelling and Cognitive Burden Views, I propose to analyse the speaker's eye-track record in order to determine whether the speaker attends to their interlocutor's feedback or not. If speakers are strategically modelling their listeners and signalling their commitment, then speakers should gaze at the listener throughout and particularly when the listener is lost and needs assistance. If speakers are suffering a cognitive burden on the other hand, then speakers should only gaze when it is feasible to do so and avoid looking during times of difficulty in their task. If speakers are doing either both or neither of these things consistently, then we have a possible case for the Middle Ground view, that is the Joint responsibility view and Pickering and Garrod's (2004) model of interactive alignment.

Since gaze has been shown to have social implications (i.e. it is considered rude to stare at someone for too long) (Argyle & Cook, 1976) which could interfere with a controlled experiment, I will use a simulation of gaze, rather than face-to-face eye contact between listeners. Anderson et al. (1997) showed that when speakers are engaged in a map task and they have full view of their partner's face, they tend to gaze more at the map and less at the partner's face. By using a simulation of gaze in conjunction with real gaze from the speaker, one can also time align the speech record with the gaze record and from this data have a finely detailed account of the disfluency event and the speaker's attention. Finally, interlocutors will perform a map task in order to guarantee spontaneous and collaborative dialogue.

CHAPTER 3 – DISFLUENCY AND VISUAL FEEDBACK

3.1 Introduction

Studies of disfluency in dialogue have suggested that speakers may be disfluent for strategic, communicative and even intentional reasons (Fox Tree & Clark, 1997; Clark & Wasow, 1998). According to these accounts, when a speaker produces a repetition disfluency, she or he wishes to signal commitment to listeners and to utterances (Clark, 2002; Clark & Wasow, 1998; Fox Tree & Clark, 1997). In Chapter 2, I attributed these predictions to the Strategic-Modelling View. In contrast to this view, the Cognitive Burden View proposes that listener modelling cannot be carried out in a consistently altruistic manner. Instead it competes with the demands of language production so that speakers' altruism is limited by the available cognitive resources (Anderson et al., submitted, Bard et al., 2004; Horton & Keysar, 2004). According to this prediction, disfluency is associated with a cognitive cost for producing speech under cognitive load (Bard et al., 2001).

Hence, there are at least two possible predictions for the sources of disfluency. The Strategic-Modelling view, supported by Fox Tree and Clark (1997) and Clark and Wasow (1998), suggest that disfluency originates out of speaker modelling and strategizing during dialogue. The Cognitive Burden view, on the other hand, describes disfluency as the result of an overburdened system (Bard et al., 2001; Horton & Keysar, 1996; Pickering & Garrod, 2004). Further testing of these hypotheses is necessary and it will be the focus of Chapters 3, 4 and 5 to report three experiments which address this issue.

3.2 Rationale & Predictions

One aim of this thesis is to investigate why disfluency occurs. I will approach this issue by attempting to tease apart two explanations already extant in the literature for the origin of disfluency: the Strategic-Modelling and the Cognitive Burden view. If disfluency is the result of strategic modelling, one possible prediction is that disfluency will increase in the presence of feedback from a listener. The Strategic-Modelling View may predict speakers will be more disfluent in an interactive setting because a speaker is more likely to signal commitment in the presence of a listener. Fox Tree and Clark (1997) argued that repetitions fulfil a specific function in speech: the speaker repeats to signal that they are having difficulty finding a word but are committed to the noun phrase they've started and to preserving continuity for the listener. Similarly, Clark and Fox Tree (2002) predict that filled pauses fulfil a specific signalling function. The filled pause *uh* signals a minor delay while *um* signals a major delay is to follow. Accordingly, the Strategic-Modelling View may instead predict no change in disfluency rate when feedback is present because filled pauses and repetitions fulfill such specific functions. Thus, there are two possible predictions for the Strategic-Modelling View alone and to be able to rule out one prediction, we need to be sure that the manipulation of feedback used in this experiment permits the speaker to do at least two things. Firstly, speakers must have a task which allows them to make complex noun phrases so that they can ostensibly stop and repeat function words while performing a lexical search in the manner described by Fox Tree and Clark (1997). Secondly, speakers must have the opportunity and a reason to choose between minor and major delays in their language production, if filled pauses are indeed signals in the manner suggested by Clark and Fox Tree (2002).

According to Barr and Keysar (2002) and Horton and Keysar (1996), feedback from a listener may make dialogue more difficult because the speaker has to manage both speech production and occasionally monitor the listener. According to Pickering and Garrod (2004), dialogue is easier than monologue precisely because the speaker can align with a listener. Therefore, there are two possible predictions for the Cognitive Burden view as well. If feedback makes the dialogue more difficult, then we would predict an increase in disfluency rate when feedback is present. If on the otherhand, the speaker is facilitated by feedback, then one would predict a rise in disfluency rates only when the feedback itself is difficult, that is when it shows signs that the listener is lost or confused.

Since both theories make two of the same predictions, namely that disfluency rates will rise in the feedback condition, the difference between a feedback and a no feedback situation is not sufficient to pinpoint the source of disfluency. Another factor is necessary. The Cognitive Burden hypothesis predicts that under the pressure of time, a speaker will be less capable of modelling the listener (Horton & Gerrig, 2005; Horton & Keysar, 1996). In contrast, the Strategic-Modelling view makes no predictions about time pressure: a speaker who is capable of monitoring a listener should be capable of doing so under varying circumstances or deadlines as long as dialogue can be conducted at all. Thus, the Cognitive Burden view predicts that when the level of difficulty increases, so too will the cognitive cost and, therefore, one would predict higher rates of disfluency in interactive, time-limited trials than in interactive, time-unlimited trials or in non-interactive trials regardless of time pressure. As it makes no predictions about time-pressure, the

Strategic-Modelling view predicts roughly the same disfluency rates in timed versus untimed trials.

It has long been known that repetitions are the most common type of disfluency (Shriberg, 1994; Lickley, 1994; Maclay & Osgood, 1959). For this reason, Clark and Wasow (1998) attempted to determine why speakers retrace or repeat a portion of an utterance rather than simply resuming where they left off. According to their continuity hypothesis, speakers repeat themselves in order to maintain a continuous utterance. Clark and Wasow (1998) propose three potential reasons to explain why a speaker might prefer a continuous utterance and thus choose to repeat rather than just begin from where they left off. First, repetition may benefit the speaker as it may be easier to repeat what one just said from the beginning. Secondly, the speaker may strategically repeat an utterance in order to make the task of comprehension easier for the listener. Thirdly, a speaker may want to present themselves as a fluent and organized. As Clark and Wasow (1998) predict there is no way to distinguish the three potential sources of continuous repetition in natural circumstances.

Nevertheless, if the speaker repeats simply because it is easier to produce an utterance from the start, then repetition rates might be equal in an interactive setting and in a trial with no feedback as long as the speaker's needs did not change. An interactive setting is defined for present purposes as a dialogue situation where a speaker receives feedback from a listener. If, on the other hand, the speaker repeats for the benefit of the listener or simply to present herself to her audience as an organized individual, then presumably one would predict that repetition rate would increase in an interactive setting. Thus, the Strategic-Modelling view predicts that repetitions could occur for intentional and strategic reasons.

Since it is primarily a hypothesis about difficulty in speech production, the Cognitive Burden hypothesis makes no predictions about particular types of disfluencies in any situation. One version of the Cognitive Burden theory does, however, predict higher disfluency rates with greater difficulty. If the listener shows signs of misunderstanding or confusion, then presumably the speaker has to expend extra effort to re-establish the conversation and this extra effort could result in disfluency arising as an indicator of difficulty. This Cognitive Burden view would predict higher disfluency rates associated with such difficult patches in the dialogue, because the speaker has had to expend extra effort in re-aligning with the listener. Once again, however, the Strategic-Modelling view and the Cognitive Burden hypothesis make similar predictions: just as the Cognitive Burden view predicts that higher disfluency rates would arise out of the difficulty of realigning with a listener after a difficult period, the Strategic-Modelling view would predict that disfluency will arise in these circumstances, as both an indicator of difficulty necessarily and as a strategic signal. Moreover, the Strategic-Modelling view predicts that the speaker should have no trouble in constantly monitoring the listener in either a visual or auditory sense. When a listener experiences difficulty, the Strategic-Modelling view predicts that the attentive speaker will track the listener until the difficulty is resolved.

Thus, a number of sub-goals arise out of the simple comparison of the Strategic-Modelling view and the Cognitive Burden hypothesis. The first sub-goal is to establish whether speakers attend to the simulation of visual feedback in a natural manner. It is important to investigate the associations between disfluency and gaze because it is one way of teasing apart the predictions of the Strategic-Modelling and Cognitive Burden Views. Table 3 lists the predictions made for speaker gaze at IF feedback with regards to both the Strategic-Modelling and Cognitive Burden views. According to the Strategic-Modelling View, speakers will monitor their listeners constantly. According to the Cognitive Burden View, speakers will avoid gazing at their partners when it is costly to do so. The advantages of using a simulation of visual feedback are that a) one can align this feedback with a record of the speech and the speaker's gaze and b) one can control the nature of the feedback to be either on-track or divergent and by so doing gain insights into which type causes more disfluency. Since the simulation of gaze is slightly unnatural, there is a need to ground the current visual feedback paradigm to be certain that it achieves similar effects to other face-to-face dialogues. The method used for simulating the visual feedback will be described in more detail in Section 3.4 and 3.5.

A second sub-goal for this experiment is to investigate that the amount of speech produced during the experiment. For example, I will test the number of words per trial because previous research by Oviatt (1995), Bard et al. (2001) and Haywood (2004) suggests that lengthier trials are associated with difficulty. If this is the case, then it would be useful to know under which circumstances speakers used the most words before the results for disfluency rate per words are given. I will test the number of transactions per trial as a baseline measure of speaker responsiveness to the visual feedback simulation of gaze. In addition to an analysis of gaze per feedback episode described above, an analysis of Transactions could reveal the responsiveness of the speaker by showing how often the speaker bothered to retrieve the visual feedback when it went awry. Finally, I will test speech rate and the temporal duration per trial as dependent variables as a means of ruling out any possible artefact for the core disfluency rate analyses. One could argue that a speaker might be more disfluent simply because he or she was speaking too quickly under time-pressure. To rule out this possibility, I will look at speech rate and temporal dialogue length per trial. It should be made clear that these measures of raw speech are not included in the predictions listed in Table 3 because their outcome is not centrally linked to the difference between Strategic-Modelling view and the Cognitive Burden view. Rather, I will report results as a 'health check' of the experimental paradigm to be sure that the central tests of disfluency rate do not contain speech-related artefacts.

Table 3. Table summarising the predictions for the Cogntive Burden and Strategic-Modelling Views

	COGNITIVE BURDEN		**STRATEGIC-MODELLING**	
Dependent Variable	**Feedback**	**Time-Pressure**	**Feedback**	**Time-Pressure**
Disfluency Rate	Increase with feedback	Increase with time pressure	Increase with feedback	No prediction
Disfluency Types	No prediction	No prediction	The rate of repetitions and filled pauses will increase with feedback	No prediction
Disfluency Rate by Conversational Move Type	Higher rate in Instruct Moves; Lower rate in Interactive Moves	Increase with time pressure	Increase in disfluency rate regardless of Move Type	No prediction
Disfluency and Gaze within a Feedback Episode	Expect disfluency when the feedback is difficult to process; Avoid gazing when costly	Expect increase in disfluency with time-pressure	Expect the Giver to look most and be disfluent during 'Wrong' feedback	No prediction
Function of Structural Disfluency Type	No prediction	No prediction	Repetitions and Filled pauses fulfil a signalling function Deletions are a sign that the speaker is opportunistic	No prediction

Next, we can begin to test the central manipulations present in this experiment, namely feedback and time-pressure. Both the Cognitive Burden and the Strategic-Modelling hypotheses predict increased disfluency rates in the presence of listener feedback. This means that we can not test feedback alone because by itself it does not distinguish between the two theories. The Cognitive Burden hypothesis predicts disfluency rate will rise under time-pressure. The strict version of this theory proposed by Horton and Keysar (1996) and Barr and Keysar (2002) predicts that disfluency rate should be at its highest when the speaker is under the most cognitive load, for example when s/he has both listener feedback and time-pressure.

Strategic-modelling makes specific predictions about two types of disfluency, disfluencies in which repetition has occurred and filled pauses, suggesting that they are signals. Strategic-Modelling also discusses how ‘self-interruptions’ support the notion that speakers are opportunistic and will take advantage of the opportunities that arise during the course of a dialogue (Clark & Krych, 2004). According to a structural classification system (e.g. Lickley, 1998), these self-interruptions would be classified as deletions (i.e. *and put it on the right-hand half of the- yes the green triangle*). The Strategic-Modelling View includes only repetitions, filled pauses and deletions in order to support the view that disfluencies occur for strategic reasons. Table 3 indicates that the Strategic-Modelling View predicts a higher repetition rate in periods of Follower feedback. Since it is difficult to predict when a speaker might be opportunistic, I have omitted this prediction from Table 3. I have also omitted any predictions for the Cognitive Burden View with regards to disfluency and function because this view makes no specific predictions about specific types of disfluencies.

By testing the effects of different types of disfluency we can distinguish between the functions each disfluency type fulfils in dialogue to see whether the Strategic-Modelling View makes the correct predictions. Are all disfluency types ‘signals’ in the strategic sense? A test of individual disfluency type is needed to answer this question. This will be conducted by calculating the rate of individual disfluency types, specifically repetitions, substitutions, insertions, deletions and filled pauses, per fluent word to see whether any individual types are sensitive to the manipulations of visual feedback or time-pressure tested in this Experiment. Other types of disfluencies (e.g. silent pauses and prolongations) will be omitted from this analysis because of their relationship to fluent prosodic boundaries (Goldman-Eisler, 1972; Wightman et al., 1992). Furthermore, I will not analyse the difference between *uh* and *um* because this difference has already been disputed by O’Connell and Kowal (2005). Instead, I will investigate what the function is for a particular type of disfluency, for example a deletion or repetition. If deletions are opportunistic as the Strategic-Modelling View predicts, then one would expect an association between deletions and an planning function, for example the movement of the visual feedback.

For an understanding of the functions disfluency plays in dialogue, we need to investigate disfluency rate in conjunction with another measure of speaker behaviour. An investigation of this sort will reveal the sorts of behaviours the speaker was engaged in when he or she became disfluent, and thus we have an insight into why disfluency occurs. Two measures are available to us in the current experiment: Conversational Moves and speaker attention. A Conversational Move is a unit of coding that can be applied to a dialogue to classify individual utterances by form and goal. To understand speaker behaviour during disfluent periods, we can analyze disfluency rate per Conversational Move for an indication of what sort of goal the speaker was trying to fulfil when s/he became disfluent. As shown in Table 3, the Strategic-Modelling View makes no specific predictions about the type of Moves in which disfluency will occur because this view predicts that speakers signal throughout a conversation to their listener regardless of utterance type. As shown by Lickley (2001), Instruct moves, which require the speaker to utilise creativity, planning and to introduce new referencts, were the most disfluent type of move in the Map Task Corpus. In conjuction with this finding, the predictions for the Cognitive Burden View shown in Table 3 suggest that Instruct Moves will also be the most disfluent move type in the MONITOR experiment.

In addition to disfluency and move type, we can investigate whether speaker attention and disfluency are related by measuring disfluency rate per feedback episode. As shown in Table 3 (page 98), the Cognitive Burden hypothesis would predict that speakers avoid gazing at the visual feedback when it is costly and an increase in disfluency during periods of complicated feedback (i.e. the square goes to a wrong landmark). Strategic-Modelling predicts that listeners will check the square throughout the trial regardless of difficulty and an increase in disfluency rate prior to complex noun phrases or complicated syntactic structures.

3.3 Method

In order to test the Strategic-Modelling and Cognitive Burden Views which make different predictions about uptake of visual feedback, Experiment 1 was designed to determine whether Instruction Givers respond to visual feedback from an Instruction Follower. During a trial, an Instruction Giver (IG) provided route descriptions of a map to an Instruction Follower (IF) located in a separate room. The IF's purported focus of visual attention was projected on the IG's version of the map. IG's gaze along the route was genuinely eye-tracked so it was possible to tell when IG gazed at the IF gaze focus and when she gazed elsewhere. To test whether speaking to an active listener increased the difficulty of the task, this condition was compared to one in which the IG did not have access to the IF's visual feedback in half of the trials. To test whether time-pressure made the task more taxing for the IG, in half of all trials the IG was subjected to a 1 minute time limit; in the remaining trials, there was no time-limit and IGs could speak without interruption.

The results in this chapter describe how Givers do in fact attend to the visual stimulus, a red square simulating saccadic gaze fixations, and guide it around the map when it is present. A red square was used instead of genuine eye gaze from a live participant because the surrogate makes it possible to track the IG's attention at the square precisely. Recall from Section 3.2 that the feedback manipulation must meed certain criteria. Furthermore, as discussed in Section 2.4, studies of collaborative dialogue and gaze have found that when engaged in a task-oriented dialogue, speakers pay more attention to the task and less attention to their partner's face. The hypothesis that responding to feedback from a listener incurs a cognitive cost is evaluated in the light of findings from collaborative dialogue. Of course, any results may be dependent on the specific paradigm. This fact will be further discussed in the Discussion section.

3.4 Materials

The four Maps used for this experiment were taken from the HCRC Map Task Corpus (Anderson et al., 1991). Pictures of the maps can be found in Appendix B. As in the Map Task Corpus, the Instruction Giver (i.e. the naïve subject) was given a map of a fictional location with a pre-printed route that traveled from a 'start' point to a 'finish' point. Also on each map were 12 ± 1 labelled cartoon landmarks. Section 1.3 outlined the design of landmarks and maps in the Map Task Corpus. In the HCRC Map Task Corpus, both participants were told that the maps would not always match perfectly. In actuality, the IG might have two occurrences of the same landmark on his map, a landmark which only he has or a landmark that is named differently from the landmark on the Instruction Follower's (IF) map. In the MONITOR Project and the Map Task Corpus, IG maps were identical since there was no actual Follower's map. The Map Task Corpus IF map was used only as a template for the design of the simulated visual 'IF' feedback. An example of a screenshot during the dialogue is shown below in Figure 7.

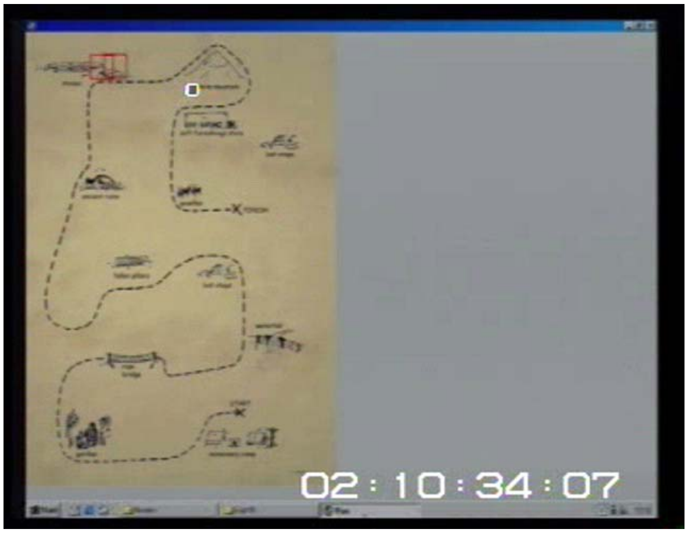


Figure 7. A snapshot of the screen during an ongoing dialogue. The black dot represents the speaker's gaze while the box represents the follower's purported location.

The feedback consisted of the pre-programmed movement of a hollow red square which travelled from landmark to landmark according to a schedule based on the original maps. For example, if the IF map contained only one Great Viewpoint landmark in the north of the map but the IG map contained two Great Viewpoint landmarks, one (the critical one) in the south and one in the north, the feedback square would go to the north Great Viewpoint as a real IF would.

3.5 Experimental Procedure

Naïve participants and the confederate were greeted by the experimenter[3] and taken into the experimental room. The experimenter then explained to the participants in the presence of the confederate that s/he would be describing a map route to the confederate, the Instruction Follower, in another room, and that in some of the trials they would receive visual feedback from the IF. The IG was told that the IF could see a map similar to the IG's map. The IG was warned in advance that some of the landmarks on his/her map might differ from the landmarks on the IF's map, but was given no indication of how or how often they would differ. Subjects were instructed to say whatever was necessary to guide the listener along the route. A copy of the instruction sheet and the consent form that the subject was asked to sign is given in Appendix AA.

8/5/078/5/07

[3] The Experimenter was David Kenicer at the University of Glasgow.

In terms of what the IG believed about the IF, IGs knew that the IF could see a map similar to the one under discussion for that trial. In visual feedback trials, IGs could see a red square which they were told represented where the IF was looking. One might ask whether the IG believed that the IF could also see the IG's gaze and if so, could the IG have tried to use their gaze to direct the IF where to go somewhat like a pointing finger. Firstly, when the experimenter explained the roles of Instruction Giver and Instruction Follower to the naïve participant and the confederate, the experimenter always stressed the fact that the IG could see the IF's gaze and that the IF could not see the IG's gaze. Secondly, just in case an IG mistakenly believed that the IF could see his/her gaze, one might anticipate the use of deictic pronouns (e.g. *Look here*) or other explicit language (*It's right there where I'm looking*) to emphasize the use of gaze as a pointer. The author examined all of the MPEG videos and the transcripts from Experiment 1 for an indication that an IG was using their gaze in a deictic manner. No such indication was found. Finally, following the experiment, IGs were questioned about the naturalness of the experiment. None of the participants suspected that the IF feedback was actually controlled by the experimenter; if a participant did seem suspicious, their data was discarded and a new participant's data replaced theirs.

Participants were then seated in a lounge chair, from which they could see the map projected on a 21" Belinea TFT flat screen monitor 3 feet in front of them. The angle of the chair kept their faces at a constant distance from the screen. The speaker's eye-gaze was calibrated using a nine-point display screen set to 'normal' strength. Eye-gaze was then recorded with table-mounted SMI (Sensory Motor Instruments) non-invasive, infra-red eye-tracking equipment in Iview version 2.0 software so that time-aligned gaze and dialogue comparisons could be made. A Corioscan PRO scan converter was used to combine video signals from the eye-tracker and the subject monitor. These were recorded in MPEG with Broadway Pro version 4.0 software. The speaker and the experimenter could communicate via Asden HS35S headsets with microphone attachments. What the speakers said was recorded in mono on a Mackie micro-series 1202 mixer and an Aiwa tape deck recorder.

All experiments reported in this thesis involved visual feedback, which purportedly represented the Follower's eye-gaze. This visual feedback consisted of a 0.5" x 0.5" hollow red square which was advanced according to a script from landmark to landmark. In effect, this red square was a surrogate for genuine eye-gaze which provided both more information and less information than is available in face-to-face interaction. The red square can be said to provide more information because Givers can see the precise location of the Follower at any time. Likewise, the red square can also be said to provide less information because the Giver cannot see any facial cues or gestures provided by the Follower. These ramifications will be considered in the Discussion section

The experimenter advanced the square by pressing a button after the first mention of each new landmark on the route. The movement of the square was scheduled to be either correct or wrong. When wrong, the square moved to a landmark that had not been mentioned. When scheduled to be correct, the square moved to the landmark named by the speaker. The trial was discarded if the experimenter missed the critical timing for one wrong landmark or more than 30% of scheduled correct landmarks. The square was also programmed to move in a way that represented realistic saccades by a programmer familiar with eye-gaze research: it made brief saccades of random extent and direction, centring on a target landmark. Naïve participants were told that the square would bounce around the screen (i.e. make saccadic movement) and that this was normal gaze behaviour.

Recall that there is a chance that the Strategic-Modelling view would predict no change between a feedback and no feedback trial if repetitions and filled pauses are specific cues in dialogue. In order for these specific cues to occur, our feedback manipulation must meet certain criteria. Speakers must be able to create spontaneously complex noun phrases during the task so that a repetition can occur while the speaker performs a lexical search. Filled pauses *um* and *uh* are thought to signal different degrees in delay. Accordingly, the feedback manipulation must be both challenging and realistic enough so that the speaker can signal a delay while they plan the next utterance, if indeed filled pauses are a signal in this fashion. For this reason, using a Map task with complex landmark names and directional terms (e.g *horizontal, vertical)* is one way to guarantee that speakers had the opportunity to formulate complex noun phrases, and therefore that they could use repetitions in a signalling function, if desired. Although the speaker only had visual feedback in Experiment 1, it is theoretically still possible that they could choose a longer pause by saying *ehm* or *um* and a shorter pause with *eh* or *uh*. The possibility that the feedback manipulation did not permit repetitions and filled pauses will be reviewed in subsequent sections.

Twenty four students of the University of Glasgow participated in the experiment and were paid £5 per hour. All participants had normal or corrected to normal vision. All participants declared themselves naïve to the purposes of the experiment in a debriefing session. Subjects were eliminated if any single map trial failed to meet the criteria for feedback or capture quality. The feedback criterion demanded that the experimenter advance the feedback square between the introduction of the pertinent landmark and the onset of the following instruction in all cases where the feedback was scheduled to be wrong and in 70% of the cases where the feedback was scheduled to be correct. The capture criterion demanded that at least 80% of the eye-tracking data was intact. The loss rate of the table-mounted eye-tracker required that thirty subjects were run before twenty-four remained with valid sessions in all four conditions and with a balanced design in total. No subjects were suspicious about the true nature of the confederate feedback and so no subjects were replaced for this reason.

3.6 Experimental Design

A 2 x 2 Repeated Measures design crossed Feedback (visual feedback and none) and Time-Pressure (timed and untimed). For the Feedback factor, subjects were presented with either visual feedback in the form of the Follower's gaze feedback square or had no feedback. On each map, there were 8 scheduled correct landmarks where the simulated visual feedback was designed to go to the landmark mentioned by the Giver. There were 4 scheduled wrong landmarks where the visual feedback was designed to 'skip' the next landmark on the route when the Giver mentioned it and go instead to a different 'wrong' landmark.

Subjects were told either that they had one minute to complete the route in the 'time-limited' condition or that they had as much time as necessary in the 'time-unlimited' condition. The four maps were rotated through the trials so that each subject saw a different map in each condition and each map was encountered an equal number of times in each condition.

3.7 Data Coding

This section will explain how the data were coded with respect to dialogue units, disfluencies and gaze.

3.7.1 Data Coding – Transactions and Moves

Recorded speech in the MONITOR Project was transcribed[4] verbatim and coded for Transaction and Conversational Move type (Carletta et al., 1997).

Transactions are blocks of dialogue corresponding to task subgoals. Transactions could be

8/5/078/5/07

[4] Under the auspices of the ESPRC MONITOR project, the dialogues were transcribed and coded by undergraduate and graduate students at The University of Edinburgh. The transcription and coding process for Experiment 1 was overseen by Dr. Maria Luisa Flechá-Garcia and in part by Dr. Yiya Chen.

labelled as normal, review, retrieval, overview or irrelevant. During a **normal** transaction, the speaker simply gives instructions on how to get from the current landmark to the next one on the route. In a **review**, the speaker retraces an earlier portion of the route. A **retrieval** transaction occurs when the speaker tells IF how to return to the route from a wrong position. In an **overview** transaction, the speaker gives a broad description of the map at large without giving any specific instructions of the route. **Irrelevant** transactions occur when the IG has to say something to the experimenter; for example the IG's mobile telephone rings. Overall, there were too few overview and irrelevant transactions for analysis purposes, so overview and irrelevant will be summed together and referred to as 'other'.

Table 4. Examples of Transaction types in the MONITOR Project

TYPE:	UTTERANCE:
Normal	"The path then follows the route along the curve of the west lake…"
Review	"Okay, go along the north part of the west lake again…"
Retrieval	"Uh no, no, go down to the other lake … yep, that one there"
Overview	"The map has four quadrants…"

Transactions are subdivided into Moves, which are defined in Carletta et al. (1997) as "simply different kinds of initiations and responses classified according to their purposes". Moves can be divided broadly into two categories: Initiating moves and Response moves. Initiating moves include **instruct**, where the speaker directs their partner to do something, usually to move along the route. In an **explain** move, the speaker spontaneously elaborates on some aspect of the route. This is distinct from an **align** move where the speaker assesses whether their partner agrees with what has been said so far. Finally, Initiating moves include **query-yn** moves in which the speaker asks a yes or no question that does not involve aligning with the partner.

Table 5. Examples of Move types in the MONITOR Project

TYPE:	UTTERANCE:
Instruct	"Go down to the left of the Dead Tree"
Explain	"There's a Dead Tree by the Forked Stream"
Align	"Right, you're at the Dead Tree"
Query-yn	"Where you are right now, have you got a waterfall?"
Acknowledge	"Aye, that's right"
Clarify	"Left underneath the Fallen Pillars"

Table 6. Sample Transaction and Move Coding from the HCRC Map Task Corpus

Transaction 1 normal startpoint 1 endpoint 2	
GIVER	FOLLOWER
Move 1 align right neil ?	
	Move 2 reply-y okay right
	Move 3 query-w where are we going ?
Move 4 reply-w start	
	Move 5 query-w where am i starting ?
	Move 6 explain oh right i've got it yeah i've found the start
Move 7 query-yn have you got the start ... just above?	
	Move 9 reply-y yeah i've found it uh-huh
Move 10 query-yn have you got a camera shop below it ... no ?	
	Move 11 reply-y yes
Move 11.6 check you have ?	
	Move 15.1 reply-y yes

Response moves include **acknowledge** moves, during which the speaker signals that s/he has understood their partner's previous move. Finally, speakers may also respond by **clarifying** on a matter of the route or task.

An example from the HCRC Map Task Corpus in Table 6 shows the relationship between Transactions and Moves. Transactions are completely divisible into Moves. As the example of a real dialogue between two participants from the HCRC Map Task Corpus in Table 6 shows, it was possible for the Giver to make a 'Reply-W' *start* in response to the Follower's Query-w Move *where are we going*. A Reply Move occurs only when one speaker has been asked a question by the other participant. A Reply-W move indicates that the speaker asked a 'wh-question', that is a question beginning with *Who, what, when, where, why* or *how*. This is possible because there were two actual people engaged in dialogue. In the MONITOR Project, there was only one naïve participant who responded to purely visual feedback. For this reason, the MONITOR Project used only a subset of the Move types explained in Carletta et al. (1997). These Move types are shown in Table 5 (page 106).

For ease of calculation in the results reported below, Align, Query, Acknowledge, and Clarify Moves were all classified as 'Interactive Moves'. In an Interactive move, the speaker is not just giving instructions or explaining them to the listener. Instead, the speaker is in some way interacting with the listener, usually to be certain that the speaker understands the instructions, to clarify the instructions more clearly, to ask for information about the Follower's location or to confirm that the Follower is in the right location. Align, Query, Acknowledge and Clarify moves are also grouped together for the purposes of this experiment because they are not expected to occur as frequently individually in the current paradigm. In true monologue, one would not expect to see any Interactive Moves because the speaker would have no one with whom to interact.

3.7.2 Data Coding – Disfluencies

All monologues were coded for disfluency according to the classification system developed by Lickley (1994; 1998). Coding was conducted using Xwaves/Entropic and Xlabel software which makes it possible to refer to spectrograms, insert labels at specific time points, and replay each disfluent area as many times as necessary.

Common disfluency tags included repetitions, substitutions, insertions and deletions. Disfluencies were occasionally deemed 'complex' if one type, say a repetition, was nested within another type, say a substitution (eg. *directly bas- directly we- west of your cattle stockade* where *we- west* is the repetition nested inside the substitution *directly bas- directly west*).

Table 7. Examples of Disfluency types in the MONITOR Project

Original Utterance	**ReparanduM**	**Repair**	**Continuation**
Repetition: strings repeated verbatim with no substitution or deletion			
Just to	My	my	left
Substitution: replacement of a word, fragment or string by a word or string, including repetitions of the original words with shared syntactic features			
Like	to the r-	to the left	of the burnt forest
Insertion: repetition of a string with one or more words inserted before or within a repetition			
Go	Two	ehm about two	centimetres above
Deletion: Interruption and restarting without repetition or substitution			
Oh no	not above the gr-		The line stops at the pirate ship

The disfluency coder also labelled silent pauses and filled pauses (*uh, um, eh*). For the most part, this thesis will focus on disfluencies, specifically speech repairs and filled pauses, rather than silent pauses. Silent pauses were not included in analyses because pauses have been shown to serve two possible functions: a) denotation of a syntactic boundary and b) gain time during hesitation (Goldman-Eisler, 1972; Duez, 1982).

3.7.3 Data Coding – Gaze

Experienced coders[5] coded the videoed gaze data in the Psychology Department at the University of Glasgow. The coders used Observer Pro software, which makes it possible to code the location of the Giver's gaze and the location of the visual feedback frame by frame. The coder typically coded 2 channels of information: on the first channel, the location of the red feedback square was coded with respect to the scheduled landmarks: a square might be 'correct' or

8/5/078/5/07

[5] Gaze coding was done under the auspices of the EPSRC MONITOR Project at The University of Glasgow and overseen by David Kenicer and Lucy Smallwood.

'wrong'. By being wrong, the feedback square would 'skip' a landmark for a period of time and then return to it once guided back. On the second channel of gaze, the location of the Giver's gaze was coded with respect to the landmarks on the route. Tags on the Follower's gaze channel consisted of the landmark name with an indication of correct or wrong, or a 'Travel' tag for frames where the square moved between landmarks. Tags included the focussed landmark name, an 'Away' tag for instances when the Giver's gaze was on the screen and an 'Offscreen' tag for instances when the Giver blinked or looked offscreen. Gaze coding on yet a third level indicated whether the Giver was looking at the Follower or elsewhere on the route.

3.7.4 Coder Reliability

Any research involving coding should have some way of accounting for potentially subjective judgments of the coders. As Carletta (2005) points out, linguistic studies have used a variety of techniques to account for the reliability of their coders. For example, Silverman, Beckman, Pitrelli, Ostendorf, Wightman, Price et al. (1992) asked coders to employ the ToBI system while labelling English prosody. Agreement between coders was the ratio of agreements between coders to possible agreements, taking into account all possible pairings of coders (Silverman et al., 1992). As Carletta (2005) continues, a number of researchers simply relied on the reader's own judgements of linguistic plausibility when presenting the results of their study.

This is no longer an acceptable method, particularly when working with many coders on a large corpus of data (Carletta, 2005; Carletta et al., 1997; Krippendorff, 1980, 1987, 2004; Siegel & Castellan Jr., 1988). Fortunately, statistical methods for computing intercoder reliability exist. One method, suggested by Carletta (2005) for content analysis, is known as the Kappa Statistic or Cohen's Kappa (Cohen, 1960). The Kappa statistic calculates the proportion of agreements among an arbitrary number of coders applying a categorical system to data, accounting for the probability that coders will agree a certain proportion of the time just by chance (Krippendorff, 1980, 2004; Siegel & Castellan Jr., 1988; Weber, 1985).

Another statistic for determining coder agreement is Krippendorff's α. Krippendorff (2004) suggests that Krippendorff's α is a general-purpose means of determining the reliability of a coding system applied by any number of coders. Krippendorff's α can be used to compute the reliability of a coding system with any number of categories or any number of coders. Krippendorff's α calculates the average difference of agreement predictable by chance between all categories, regardless of which coder assigned them and to which units they were assigned

(Krippendorff, 2004).

Krippendorff (2004) argues that the Kappa statistic overestimates reliability by increasing the amount of predictability of the categories that one coder uses compared to the categories used by the other coder. Mathematically speaking, the denominator of the Kappa statistic is similar to chi-square, or a measure of correlation. What this means is that the Kappa statistic is more concerned about the coder agreement and less so about the coding of the agreement.

Since this thesis will rest upon analyses taken from disfluency coding initially performed by the author, I will report the results of two reliability studies in subsequent chapters. The first reliability study was done to ensure that coders agreed on the disfluency coding system as outlined by Lickley (1998). Coders were two PhD. Candidates[6] who were also conducting research on disfluency for their dissertations. At the beginning of the training period, coders were introduced to Lickley's (1998) coding manual and guided through a pre-coded trial. The coders were then given 3 trials to code. The author met on occasion with the coders to resolve disagreements but no judgments were changed as a result of discussion. There were 70 disfluencies about which all three coders agreed, 53 disfluencies about which only two coders agreed and 2 disfluencies about which all three coders disagreed. I will cite Cohen's Kappa because it is the most widely used, as well as Krippendorff's alpha for the reasons explained above. The Kappa results of the disfluency coding reliability test showed that the author and the first coder had a Kappa of .578 at $p < .001$. Agreement between the author and the second coder was $K = .63$ at $p < .001$. Agreement between the two coders was the lowest with $K = .44$ at $p < .001$. When agreement was calculated for all three coders[7], Krippendorff's $\alpha = 0.74$ (i.e. between the $0.67 < \alpha < 0.82$ range at a $p < .05$ confidence level).

3.7.5 Data Analysis

All transcription-related data files were output in XML format for analysis using scripts designed for this purpose[8] by the MONITOR Research Assistants[9]. All experiments had XML files for the dialogue structure, the Giver's gaze, the Follower's gaze, and the Giver's disfluency

8/5/078/5/07

[6] Thanks to Michael Schnadt and Lucy MacGregor for assistance in this regard.

[7] Thanks to Prof. Klaus Krippendorff for assistance in this manner.

[8] Thanks to Henry S. Thompson for his assistance with scripts.

[9] Thanks to Maria-Luisa Flecha García, Yiya Chen and Catriona Havard for assistance in this regard.

record. A dialogue XML file contains time-stamped Transactions and Moves in addition to words and referring expressions. A Giver's gaze file contains time-stamped fixations with respect to the objects that the Giver gazed at. The Follower's gaze XML file includes time-stamped indications of the Follower's movement along the route, with respect to whether these movements were correct or wrong.

Disfluency coding began once a transcript for the trial in question was complete. This transcript was converted into a .words text-file for use in Xlabel using a script to remove extraneous XML tags[10]. Frequently, the disfluency coding process would reveal errors (i.e. missing words, mistranscribed words etc.) in the transcript, which were subsequently amended. Once a trial had been coded for disfluencies in Xlabel, the Xlabel disfluency file was converted into XML using a script especially designed for this purpose[11]. Next, frequency counts were taken for all disfluency types and FPs using the grep function in UNIX. Word counts for entire trials were also taken in the same manner. These counts were then entered into Excel and statistical tests were applied using SPSS v. 11.5 or v.12.0. The results of these tests are described in Chapters 4, 5, and 6.

The next step in the data collection process involved analyzing all XML files for information on the relationship between the dialogue structure, the Giver's uptake of information, Follower feedback and the Giver's fluency. This task was accomplished in one of four ways: 1) by pulling data out using the NITE XML Toolkit[12], 2) by using specific-purpose Perl scripts[13], 3) by manual inquiry by the author if listening was required, or 4) by pulling information out with the MySQL database query language. The subsequent chapters will describe in detail the investigations undertaken and the ensuing results.

In the following sections and chapters, I shall only report the results of by-subject analyses. By-items analyses (i.e. by map or landmark) were not done because a by-item analysis would not generalize over linguistic material. If the difference between the items were due solely to the different experimental conditions, one would benefit from doing a by-item analysis. In the current experiment, since linguistic material differs for each item, be it an entire map or a single landmark, one would not benefit from a by-items analysis precisely because it would not

8/5/078/5/07————————————

[10] Thanks to Cedric MacMartin for his assistance with the trans2xlab script.

[11] Thanks to Ruli Manurung for his assistance with scripts.

[12] Thanks to Jean Carletta for her assistance with NITE. More Information on NITE can be found at http://www.ltg.ed.ac.uk/NITE/

[13] Thanks to Joseph Eddy for his assistance with scripts.

generalize over linguistic material.

3.8 Words and Speech Overall

Table 8 below shows the overall distribution of transactions, words, disfluencies, filled pauses and average time a trial took.

Table 8. Overall distribution of Total Transactions, Total Words, Average Time in Seconds a trial took, Total Disfluencies and Total Filled Pauses in Experiment 1B

MEASURE	FT	FU	NT	NU
Transactions	**278**	**328**	**247**	**308**
Normal	226	240	229	280
Retrieval	36	63	0	0
Others	16	25	18	28
Words	**5763**	**7685**	**5166**	**7760**
Normal	5032	6443	4867	7392
Retrieval	545	897	0	0
Others	186	345	299	368
Time in Seconds	**101.23**	**149.36**	**82.89**	**137.75**
Disfluencies	**204**	**305**	**150**	**249**
Repetitions	80	119	51	101
Substitutions	36	42	19	51
Insertions	51	78	56	75
Deletions	37	65	21	22
Filled Pauses	**176**	**225**	**140**	**236**

Transaction and word counts are broken down into Normal, Retrieval and Other (e.g. Irrelevant, Review and Overview) Transactions to show where the most speech occurs. In Table

8, I refer to the Feedback Timed condition as ‘FT’, the Feedback Untimed condition as ‘FU’, the No-Feedback Timed condition as ‘NT’ and the No-Feedback Untimed condition as ‘NU’. The distributions of disfluencies, filled pauses, words and transactions by subject and trial are shown in Appendix C. The word counts shown include words in reparanda.

3.8.1 Transactions

What type of transactions do speakers make most? To answer this question, the rate of transactions per trial was submitted to a within-subjects ANOVA for Time-pressure (2: Timed vs. untimed) x Feedback (2: Feedback vs. No Feedback). Time-pressure caused the overall transaction rate per monologue to decrease. Speakers produced more transactions in time-unlimited conditions (13.83 transactions per trial) (F_1(1,23) = 9.95*, p* < .01) compared to the time-limited conditions (11.27 per trial). The feedback condition did not contribute significantly (F_1(1,23) = 3.98*, n.s.)* to the transaction total, nor was there an interaction between timing and feedback conditions (*F*(1,23) = 0.305, n.s.). As shown in Figure 8, Normal transactions patterned according to the overall transaction rate: Normal transactions were more numerous in the Untimed condition (11.40 per trial) compared to the Timed condition (F_1(1,23) = 5.77*, p* < .025).

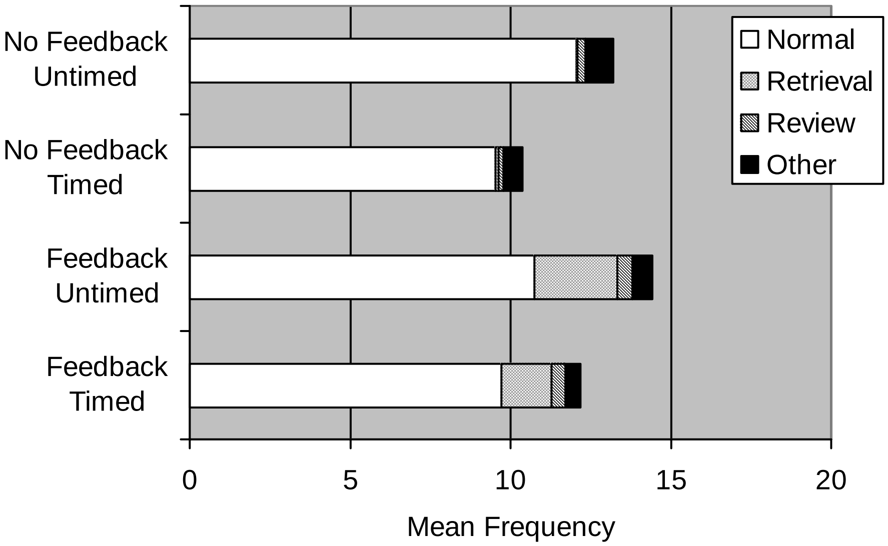


Figure 8. Observed mean transactions with respect to type for each experimental condition.

As is evident from Figure 8, Retrieval transactions occurred only in the two feedback conditions (13% of all Transactions in Feedback-Timed; 18% in Feedback-Untimed) but very rarely otherwise (0.8%[14] of all No Feedback Timed transactions and 0.3% of No Feedback Untimed transactions: by-subjects ANOVA main effect for Feedback, (F_1(1,23) = 25.84, p < .001)). There was a non-significant trend for more Retrieval transactions in untimed conditions (F_1(1,23) = 4.12, p = .054) but only because of the increase in Retrievals in the Feedback conditions (interaction: (F_1(1,23) = 5.40, p = .029). Other transaction types were unaffected by the experimental factors suggesting that only Retrievals and Normal transactions were significant to the effects of the experimental design.

Time-pressure caused the overall transaction rate per trial to decrease. Speakers produced more transactions in time-unlimited conditions (13.83 transactions per trial) compared to the time-limited conditions (11.27 per trial) (Time-Pressure: F_1(1,23) = 9.95, p < .01). The feedback condition did not contribute significantly (F_1(1,23) = 3.98, *n.s.)* to the transaction total, nor was there an interaction between timing and feedback conditions (F(1,23) = 0.305, n.s.). As shown in Figure 8, Normal transactions patterned according to the overall transaction rate: Normal transactions were more numerous in the Untimed condition (11.40 per trial) compared to the Timed condition (F_1(1,23) = 5.77, p < .025).

3.8.2 Words

Before determining how Givers respond in a disfluent manner, it might be helpful to have an inkling of their fluent behaviour during map description. Previous studies have found that longer dialogues, or dialogues with more words, tended to be more disfluent than shorter dialogues (Bard et al., 2001; Oviatt, 1995). According to these studies and to Haywood (2004) one might predict that lengthier trials are symptomatic of difficulty. For these reasons, it is useful to know something about the words delivered per trial (Figure 9).

As previously reported by Bard et al. (2003) and Bard et al. (2004), time-pressure affected only the length of trials. An ANOVA on the total number of words (including words in reparanda) showed that speakers were more loquacious in the conditions without time-pressure (319 words per trial on average) compared to when the IG had a deadline (224 words) (F_1(1,23) = 33.68, p < .001).

8/5/078/5/07

[14] The rates of Transactions in No Feedback trials are just visible in Figure 8 due to the large scale of the graph.

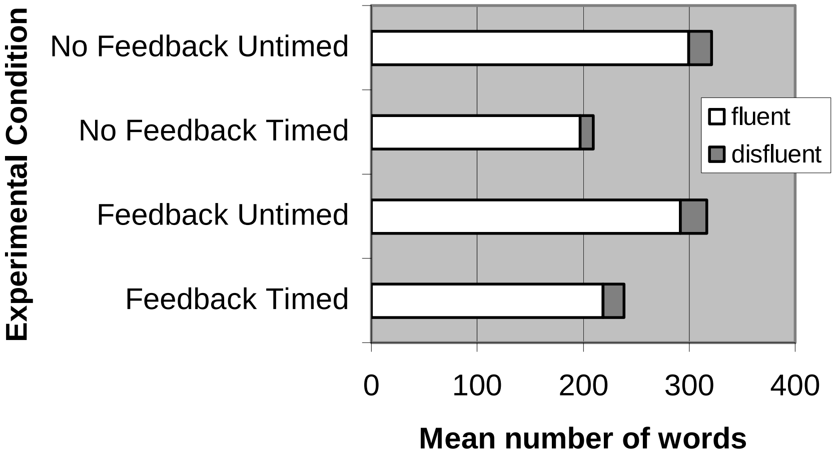


Figure 9. Observed mean fluent and disfluent words per monologue with respect to experimental condition

Feedback had no significant effect for either word count and there was no significant interaction. Thus, the only factor influencing the length of the trials seemed to be time-pressure. This result is in concordance with previous research by Bard & Aylett (2001) and Oviatt (1995) which shows that given more time, speakers will say more.

3.8.3 Temporal Dialogue Length in seconds

As shown in Figure 10, Dialogues tended to be temporally longer in the presence of listener feedback (125.29 seconds on average) than in the absence of listener feedback (110.32 seconds) (by-subject ANOVA, main effect of Feedback: $F_1(1,23) = 11.91$, $p < .01$).

Speakers engaged in temporally longer dialogues when they had the time to do so: the untimed dialogues (143.56 seconds) tended to be longer than timed dialogues (92.06 seconds) with an average difference of 51.9 seconds (Time-pressure: $F_1(1,23) = 58.93$, $p < .001$). A significant interaction of Feedback x Time-pressure was not found. Thus, both a feedback effect and a time-pressure effect are found for dialogue length in terms of seconds, but one was not found for dialogue length in terms of raw words.

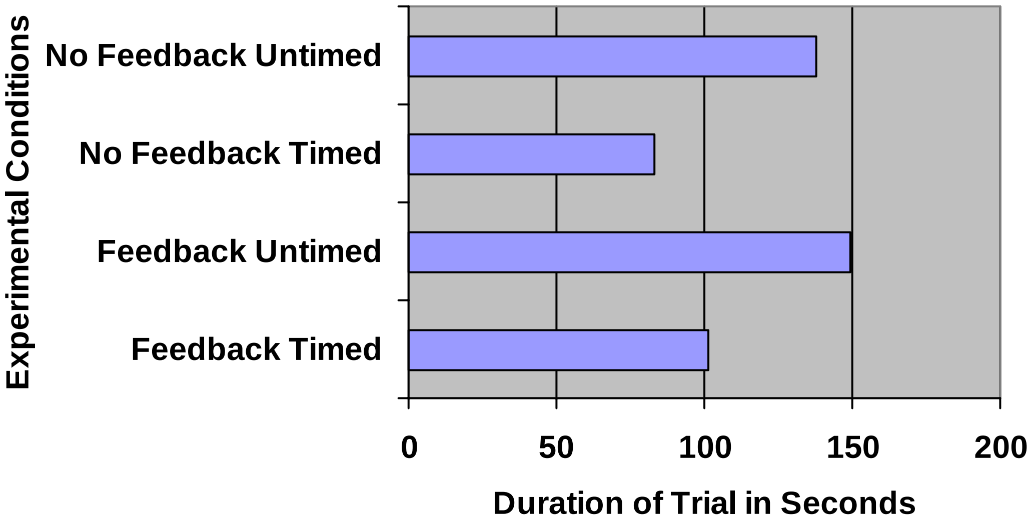


Figure 10. Observed duration of trials in seconds with respect to experimental condition

3.8.4 Speech Rate

Speech rate across experimental conditions was also subjected to a Repeated Measures ANOVA. An analysis of speech rate is important so that we can be certain that IGs were speaking at roughly the same rate in all of the conditions. Once we know that IGs speak at roughly the same rate, we can rule out the possibility that any changes in disfluency were artefacts of an external factor like speech rate. To calculate the speech rate, we divided the total Giver words per map by the total amount of time the Giver spent speaking for that map (i.e. the sum of all conversational moves less the summed durations of silent and filled pause time). There were no significant differences between either Feedback ($F_1(1,23) = 2.24$, $p = .148$), Time-pressure ($F_1(1,23) = .247$, $p = .606$) or the interaction ($F_1(1,23) = .000$, $p = .997$) with respect to speech rate.

3.9 Disfluency Rate

3.9.1 Disfluency Rate Overall

Are speakers more disfluent in interactive circumstances as predicted by both the Strategic-Modelling View and the Cognitive Burden Hypothesis? To answer this question, I analysed disfluency rate per word. Since significant effects were found for both word and transactions in the time-unlimited conditions, the results pertaining to disfluency may be only an effect of the length of the trial and opportunities to be disfluent. Total numbers of disfluencies, that is the total number of speech repairs, are given in Table 8 while disfluency rate per fluent word is depicted in Figure 11. An ANOVA for disfluency rate (disfluency per fluent word) that crossed Time-pressure (timed vs. untimed) and Feedback (feedback vs. no feedback) showed that the rates of disfluency events increased in conditions with feedback ($F(1,22) = 4.45, p < .05$). An ANOVA for disfluency plus filled pause rate per fluent word that crossed Time-pressure (timed vs. untimed) and Feedback (feedback vs. no feedback) failed to reveal any significant results (all $p > .05$).

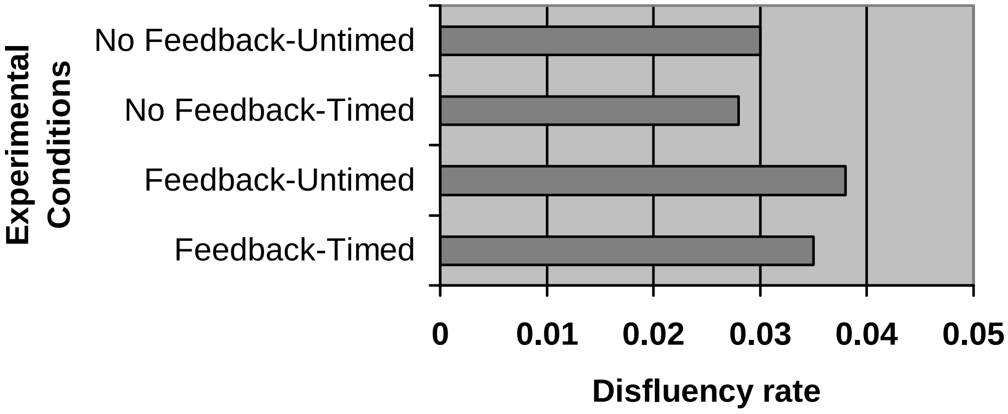


Figure 11. Observed Mean disfluency rate, (i.e. disfluency per fluent words) by experimental condition

Previous research by Oviatt (1995) and Bard et al. (2001) showed that disfluency rate increases as a function of utterance length. In order to control for this effect, an ANCOVA of disfluency rate with the numbers of transactions as a covariate further confirmed the significance of the presence of feedback ($F_1(1,22) = 11.23, p < .01$) without the confounds of word and transaction.

Overall, the results on disfluency rate show that Givers were more disfluent in interactive circumstances. Time-pressure did not affect disfluency rate.

3.9.2 Disfluency Types

Clark and Wasow (1998) predict that repetitions in particular fulfil a signalling function and so in order to test this hypothesis, a test of disfluency rate by disfluency type is needed. One would predict that if repetitions are signals, then repetition rate should be higher in interactive circumstances. As found previously by Branigan et al. (1999), Lickley et al. (1999), (Maclay & Osgood, 1959) and Shriberg (1994), repetitions were the most frequent of the four repair types with a raw total of 351 across all conditions and all speakers. Whilst the majority of work investigating the potentially strategic nature of repairs (cf. Clark & Wasow, 1998) has focused solely on repetitions, no significant effects were found here for either feedback or time-pressure. A nearly significant trend showed that speakers tended to repeat more in the feedback condition (.015) compared to the no-feedback condition (.011) ($F_1(1,23) = 3.89$, $p = .061$). Figure 12 below depicts the disfluency rate breakdown into type of disfluency with respect to the four experimental conditions.

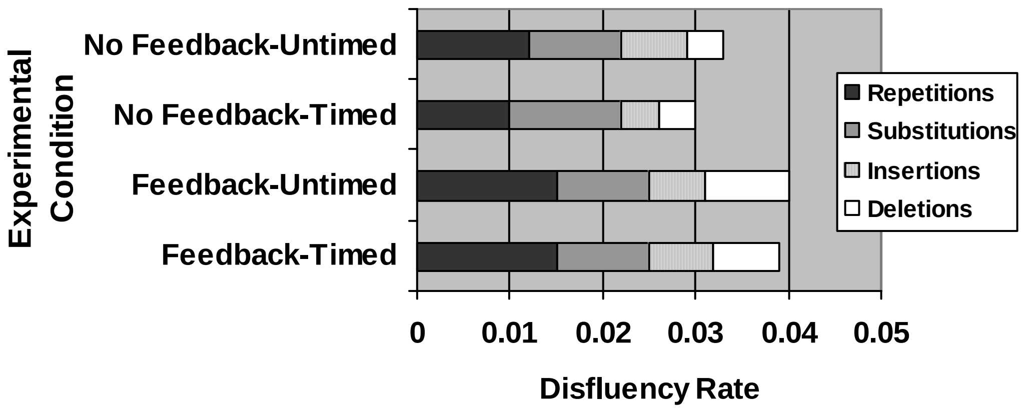


Figure 12. Disfluency rate (disfluencies per total words) by disfluency type in each of the four experimental conditions. Disfluency rate is presented on the x-axis.

Of the disfluency types described in Section 2.1.2, insertions and substitutions tend to correspond to Levelt's (1983) Appropriateness repair type (Lickley, 1994). This means that during an insertion or a substitution, the speaker is attempting to modify the original utterance by either adding or replacing information. Individual ANOVAs for the rate of each disfluency type per fluent word were run. Each ANOVA crossed Time-pressure (timed vs. untimed) with Feedback (feedback vs. no feedback). Two ANOVAs failed to demonstrate significant effects for insertions (N = 148; Means: Feedback Timed = .007, Feedback Untimed = .006, No Feedback Timed = .004, No Feedback Untimed = .007) and substitutions (N = 260; Means: Feedback Timed = .009, Feedback Untimed = .01, No Feedback Timed = .012, No Feedback Untimed = .01).

Deletions (N = 145) occur when the speaker abruptly stops mid-utterance and makes a fresh start. Deletion rate per fluent word was submitted to an ANOVA that crossed Time-pressure (timed vs. untimed) with Feedback (feedback vs. no feedback). The rate of this repair type was the highest in the Feedback Untimed condition compared to any other. Speakers delete most in the presence of listener feedback (.008) compared to the no feedback conditions (.004) (F_1(1,23) = 11.92*, p* < .01).

Filled pauses are quite common in speech and have a role in the Strategic-Modelling View according to Fox Tree and Clark (1997). For this reason, filled pause rate (N =777) per fluent word was submitted to an ANOVA for Time-pressure (timed vs. untimed) and Feedback (feedback vs. no feedback). Filled pause rate alone failed to show any significant results with respect to either feedback (F_1(1,23) = .416, *p* = .526) or time-pressure (F_1(1,23) = .249, *p* =.622). For this reason, further discussion of filled pauses will be omitted from this chapter with the exception of Section 3.10

For non-deletion disfluencies (summed raw totals of repetitions, substitutions and insertions), only the difference between timed and untimed conditions was significant (F_1(1,23) = 14.22*, p* < .001). No effects of any kind were found for non-deletion disfluency rate per fluent words. These results in comparison with those for deletions suggest how feedback influenced the speaker. Deletion frequency rose considerably when the speaker had access to feedback. For other disfluencies raw effects of time-pressure were found. This finding suggests that raw disfluency totals reflect only a measure of the trial length.

Contrary to the predictions made by Clark and Wasow (1998) for dialogue, speakers did not seem to use repetitions to make commitments to their utterances in interactive circumstances. Instead, Givers deleted more often when listener feedback was available. It could be the case that the occurrence of a deletion may depend on the behaviour of the eye-gaze feedback square. When it trails off course, the speaker is likely to abandon the current set of instructions in order to reorient the listener back on course. To draw any conclusions about this matter we will need in depth analysis of eye-gaze with respect to repair onset time.

In this section, we have shown that disfluency rate conforms to the expectations found in the literature, namely that it increases as a function of utterance length. Overall, disfluency rate was greater in the presence of feedback from the listener as anticipated by both the Cognitive Burden hypothesis and the Strategic-Modelling hypothesis. Individual tests of disfluency were conducted to investigate the functions that disfluency might fulfil in dialogue. If repetitions are a signal, they should occur more frequently in the presence of feedback from their recipient. Our results did not show support for this prediction. Instead, results suggested that deletions pattern according to the experimental design: there were more deletion disfluencies in the presence of listener feedback than in the no feedback conditions. What role do these deletions, and disfluencies in general, fulfil in the dialogue? We are now at a point where we can begin to investigate speaker behaviour, in order to understand the functions of disfluency, both generally and individually. First, we will investigate disfluency rate by Conversational Move type in order to understand which dialogue goals the speaker was attempting to fulfil when s/he became disfluent. Then, we shall turn to an investigation of disfluency and gaze in order to understand what the speaker was attending to during a disfluent episode.

3.10 Disfluency Rate by Conversational Move Types

Lickley (2001) reports disfluency rates for different move types in the HCRC Map Task Corpus (Anderson et al., 1991). Lickley investigated the differences between self-repair type disfluencies and filled pauses for every 100 words in moves of different types (See Section 3.7.1 for full classification). His results showed that Reply-W Moves were the most disfluent, if both self-repair type disfluencies and filled pauses were considered. If filled pauses were omitted and only self-repair rates were considered, then Instruct Moves, the bulk of most IG moves, were the most disfluent (Lickley, 2001).

In order to determine whether certain Conversational Move types were associated with disfluency, a by-subjects ANOVA which crossed Time-pressure (2: Timed vs. Untimed), Feedback (2: Feedback vs. No Feedback), Disfluency Type (5: Filled Pauses vs. Repetition vs. Deletion vs. Insertion vs. Substitution) and Move Type (3: Instruct vs. Explain vs. Interactive) was executed. The dependent variable was calculated by counting the number of disfluencies and filled pauses of a certain type, for example repetitions, and then dividing by the number of fluent words in that Move. The rates were then averaged over all the values for that subject in the experimental condition, e.g. Feedback Timed. Since an entire Conversational Move can be abandoned if a deletion occurs resulting in 0 fluent words, 1 was added to all fluent word totals.

Instruct Moves (.078) were again the most disfluency-prone (Move Type: Explain: .055; Interactive: .018, $F_1(2,46) = 17.98$, $p < .001$). Further support for the claim that deletions are more common when feedback is present was obtained. When they had access to feedback, speakers made more filled pauses (.90) than repetitions (.043) or insertions (.026) (Feedback x Disfluency Type: $F_1(4,92) = 6.86$, $p < .05$, $\alpha < .001$; Bonferroni, $t = 3.59$, $p < .003$ $\alpha < .003$; Bonferroni, $t = 4.31$, $p < .003$, $\alpha < .003$). All other Bonferroni t-tests were non-significant. Repetitions, substitutions and insertion type disfluencies tended to occur more frequently in Instruct Moves than in Explain or Interactive Moves (Disfluency Type x Move Type: $F_1(8,184) = 2.62$, $p = .01$;

Bonferroni t-tests, $p < .001$, $\alpha < .001$). Deletions were more common in Instruct Moves (.099) than they were in Interactive Moves (.024) (Bonferroni, $t = 4.35$, $p < .05$, $\alpha < .001$) but no significant difference was found between Deletions in Instruct Moves (.099) and Deletions in Explain Moves (.109) (Bonferroni, $t = -.289$, $p = .775$). Filled Pauses were most common in Instruct Moves (.096) compared to Interactive moves (.032) (Bonferroni, $t = 6.71$, $p < .001$, α $<.001$) matching Lickley's results for Instruct Moves.

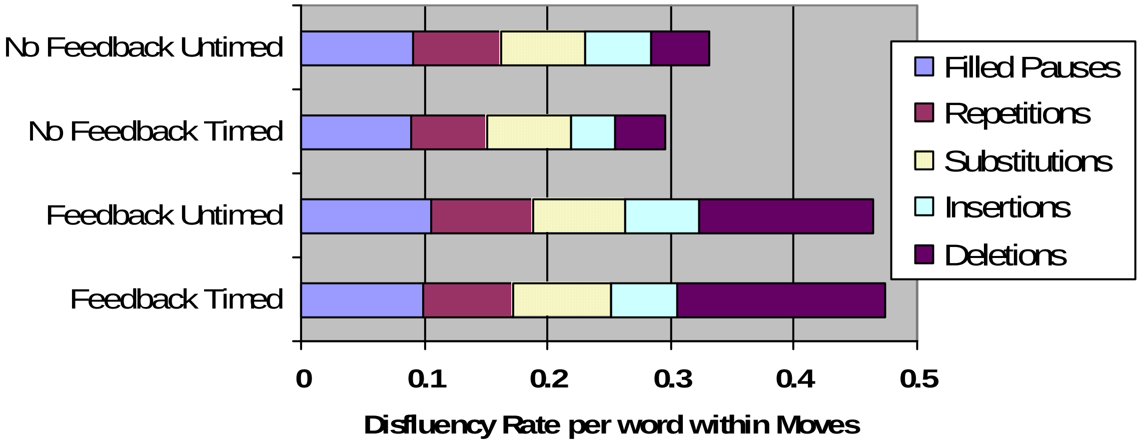


Figure 13. Disfluency rate by disfluency type per number of words within Instruct Moves

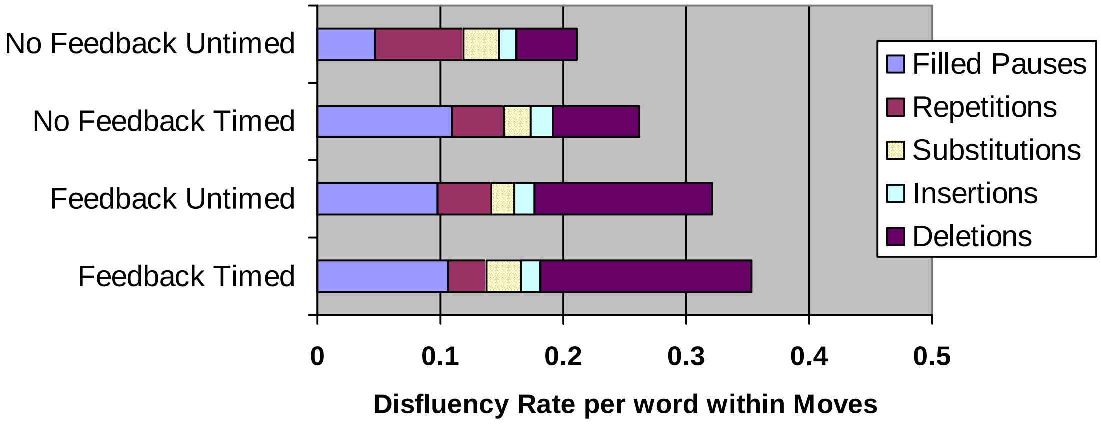


Figure 14. Disfluency rate by disfluency type per fluent words within Explain Moves

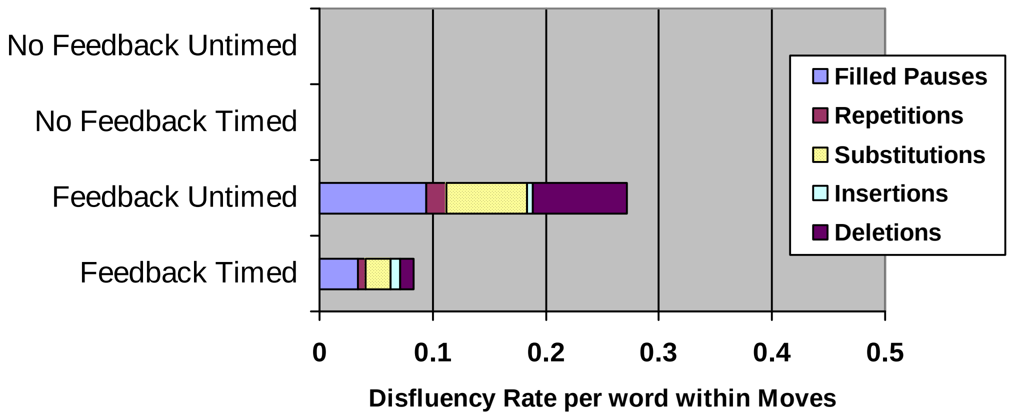


Figure 15. Disfluency rate by disfluency type per fluent words within Interactive Moves

Bard et al. (2003) report that Instruct Moves are the most common type of move in the MONITOR corpus. The disfluency rate results within Conversational Moves seem to reflect this aspect of the experimental design. As Figures 13, 14 and 15 illustrate, repetitions, substitutions, insertions and filled pauses occur more frequently within Instruct moves than they occur within Explain or Interactive moves. Deletions occur at about the same rate within Explain and Instruct moves. Surprisingly, although deletions are more frequent in the presence of feedback, they were not as common in Interactive moves as they were in Instruct or Explain moves. This result could indicate that when speakers make interactive moves to help the Follower in specific circumstances, they are not generally very disfluent. Rather, they are more disfluent when presented with the task of describing the route to the Follower. Typically, an Instruct move will be more syntactically complex than any other type of move because the speaker is engaged in trying to describe the route. Thus, we can conclude that speakers are more disfluent when describing the route than they are when they interact directly with the feedback square. We can now turn to the next section in order to understand the gaze behaviour of the speaker during disfluency.

3.11 Disfluency and Gaze within a Feedback Episode

In this section, we analyze a further indicator of speaker behaviour, eye-gaze, in order to complement the knowledge we have already about speaker behaviour during different Conversational Moves. Eye-gaze information is a valuable resource because it indicates what the

IG was attending to at the time s/he was disfluent. If the IG was truly interested in the well-being of the listener and in attending to the listener's feedback, then the Strategic-Modelling View of repairs would predict that disfluencies are concurrent with at the very least the most problematic spans of the discourse, i.e. when the speaker is dealing with an errant follower. Therefore, if we know what the speaker was attending to when s/he became disfluent, we can know a little more about the causes of disfluency.

The effects of feedback condition on disfluency rate in Section 3.9 suggest that deletions in particular might be related to visual information. In order to test this hypothesis, an analysis was undertaken of all feedback episodes in the feedback trials (N = 694) to see whether

a. the Giver attended to the Follower's location
b. the Follower was where she was meant to be or whether she had deviated off-course
c. the Giver was disfluent during the episode

A feedback episode begins when the feedback square moves to the next landmark on the route after the first mention of a landmark name by the IG. If the feedback square was scheduled to move to a correct landmark, then it will move to the landmark just mentioned by the IG (Figure 16). If, on the other hand, the feedback square was scheduled to go to a wrong landmark, it will deviate off course to a landmark that was not just mentioned by the IG (Figure 17).

The distribution of Feedback episodes is shown in Table 9. This Table breaks the Feedback episodes down into either 'Correct' (instances where the IF moved to the mentioned landmark) or 'Wrong' (instances where the IF diverged off-route) sequences. An example of Correct feedback is depicted in Figure 16 and an instance of Wrong feedback is depicted in Figure 17. Table 9 further shows whether the Giver attended to the feedback square or whether he or she was busy looking somewhere else on the route. An opportunity was considered 'Looked at' if the Giver's gaze hovered over the feedback square for even a short period during the episode, or 'Not Looked otherwise. Finally, the episode was labelled 'disfluent' if a disfluency of any type or filled pause occurred during that episode and fluent otherwise.

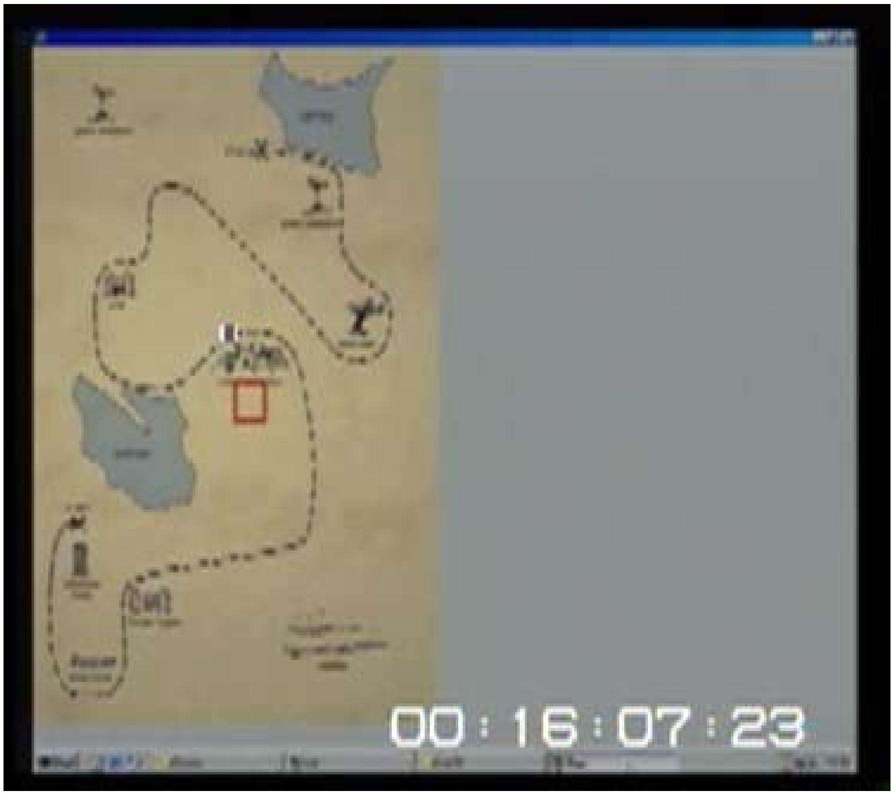


Figure 16. An instance of 'correct' feedback where the Follower's square hovers over the intended landmark. The black dot = IG gaze. The square = 'IF' gaze

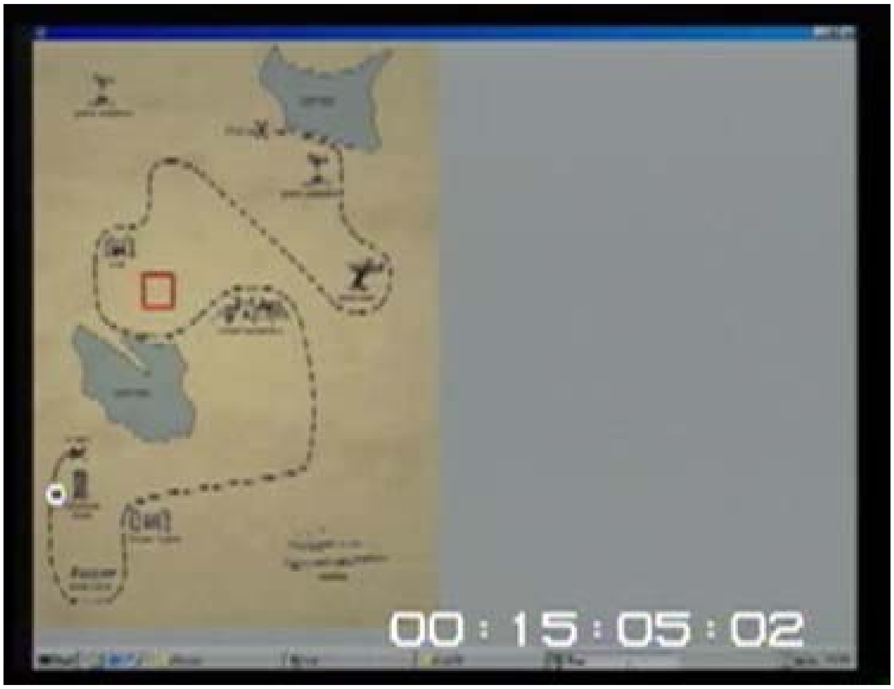


Figure 17. An instance of 'wrong' feedback where the Follower's gaze is diverted to another location other than the one intended. The dot = IG gaze. The square = 'IF' gaze

Table 9. Distribution of Correct and Wrong Feedback opportunities in the Feedback conditions. Opportunities are divided with respect to Giver Attention (Looked vs. Not Looked) and Fluency (disfluent vs. fluent). Overall Means for each cell over all 24 participants is given in parentheses.

	Correct N = 507				Wrong N = 186			
Gaze behaviour:	LOOKED N = 389		NOT LOOKED N = 118		LOOKED N = 113		NOT LOOKED N = 73	
Condition:	disfluent	fluent	disfluent	fluent	disfluent	fluent	disfluent	fluent
Feedback Timed	80 (.317)	93 (.369)	22 (.088)	52 (.212)	27 (.292)	23 (.253)	21 (.226)	21 (.219)
Feedback Untimed	99 (.409)	117 (.415)	19 (.079)	25 (.098)	40 (.389)	23 (.240)	21 (.222)	11 (.108)
TOTAL	**179**	**210**	**41**	**77**	**67**	**46**	**41**	**32**

To check for the effects of feedback type on disfluency, I ran an ANOVA with the proportion of disfluent feedback opportunities out of the total number of feedback opportunities as the dependent variable. The independent variables were Square (2: Correct vs. Wrong) and Time-pressure (2: Timed vs. Untimed). As Table 9 shows, there were more correct feedback episodes than wrong feedback episodes and for this reason the dependent variable must be the proportion of disfluent events. Givers were more disfluent when the feedback square was at a wrong landmark (.382) than they were when it visited a correct landmark (.292) (Square: $F_1(1,23) = 5.75$, $p < .05$). Givers were also more disfluent in untimed feedback episodes (.376) than in timed episodes (.297) (Time-pressure: $F_1(1,23) = 5.28$, $p < .05$). These results suggest an association between difficult feedback (i.e. wrong feedback) and disfluency but do not tell us whether the Giver gazed at the wrong feedback while being disfluent.

To check for the effects of attention and feedback type on disfluency, I ran an ANOVA for disfluent "looked at" episode rate per feedback episode as the dependent variable with Feedback Square (2: Correct vs. Wrong) and Time-pressure (2: Timed vs. Untimed) as independent repeated measures. The ANOVA revealed only an effect of time-pressure: Givers were more disfluent when they had unlimited time (.399) than when they had a time-limit (.340) (F_1(1,23) = 4.27, p < .05). An ANOVA for the "not looked at" disfluent episodes showed that Givers responding to a lost Follower but not looking at the Wrong square (.224) were more disfluent than Givers responding to a correct Follower but not looking at the feedback square (.083) (F_1(1,23) = 21.12, p < .05, α < *.001*). This result matches a general gaze and feedback interaction described by Bard et al. (2003) and Bard et al. (2004). According to Bard et al., Givers spent more time gazing at feedback that was easy to process (i.e. the correct feedback which they would look at because it is next on the route) compared to feedback that was hard to process, or wrong feedback in which they would have to find the lost Follower.

3.12 Function of Structural Disfluency Type

The results of the previous section on disfluency and gaze investigated disfluency in general and found no indication that complicated feedback, such as gazing at the Follower's feedback on a wrong, off-route landmark, induced disfluency. Though responding to a lost Follower did seem to make the Giver more disfluent. In this section, we expand upon this research by investigating deletion disfluencies. Deletion disfluencies were the only type of disfluency to show any sensitivity to feedback and so by conducting a deletion-specific analysis we can understand more about which circumstances seem to induce them (Section 3.9.2). There are two obvious functions for a deletion: one, the speaker could abandon an utterance because of something s/he saw on the screen which indicates that the speaker needs to re-plan the current utterance. I will call these 'planning deletions'. In the second type, the speaker abandons an utterance when no salient external event has occurred but when instead s/he decides either to restart the utterance anew or rephrase the utterance in a different manner. I will call this type of deletion a 'hesitation deletion'.

In order to conduct an analysis of planning versus hesitation deletions, I listened to 155 deletions and watched video MPEG recordings of the screen during the deletion. If the feedback square moved to a different landmark and the speaker's gaze track moved as a result, a deletion within this episode was considered a planning deletion (*If you go to the well, if you look…that's it yeah, that's that's the start)*. All remaining deletions which could not be pinned down as occurring for planning reasons were considered hesitation deletions (*If you can turn west you sh-…uh there's a swan pond)*. Raw numbers of planning and hesitation deletions are given in Table 10. More planning and hesitation examples from Experiment 1 are given in Appendix D.

A second reliability study was performed in order to ensure that a cognitive classification system devised by the author in order to test predictions made in this thesis is also replicable by future research. The system required a coder to judge whether a disfluency is due to planning or hesitation functions. The coder for this reliability test was an MSc. Student[15] in psycholinguistics. At the beginning of the training period, the coding system was presented by the author to the coder through the medium of MPEG video clips and detailed transcripts of the dialogue featured in the video. Following the training period, the coder was asked to code 66 disfluencies according to their function, i.e. either planning or hesitation. The coder and the author then met following each block of coding to discuss the judgments. In some cases (23 out of 112), the discussion led to a recoding of the disfluency in question; in other cases (89 out of 112), judgments were left as they originally were by both parties. Agreement for this reliability test was good for both Kappa (K = .73) and Krippendorff's alpha *(*α =.74; .58 < α < .90).

The results on disfluency and speaker attention suggest that Givers were more disfluent when the feedback square was on a wrong landmark. These results did not take into account whether the Giver had actually attended to the landmark while he was disfluent. A second ANOVA for the proportion of disfluent, "looked at" feedback episodes showed that only time pressure affected the results. It seems then that the Giver encounters difficulty when he or she has to reorient a lost Follower. It is immediately evident from Table 10 that no planning deletions occur in the No Feedback condition. Nor can they as they depend on feedback. I will omit these cells when doing statistical analyses.

Table 10. Distribution of planning and hesitation deletions in Experimental conditions Rates are given in parentheses.

	Experimental Condition			
Disfluency Function:	Feedback Timed	Feedback Untimed	No Feedback Timed	No Feedback Untimed
Planning	17 (.003)	32 (.004)	0	0
Hesitation	21 (.004)	38 (.005)	24 (.005)	23 (.004)
TOTAL	**38**	**70**	**24**	**23**

Independent ANOVAs for hesitation deletion rate per word with Feedback (2: Feedback vs.

8/5/078/5/07

[15] Thanks to Ryan Gramacy for assistance in this regard.

No Feedback) x Time-pressure (2: Timed vs. Untimed) revealed no significant effects (Feedback: $F_1(1,23) = 0.286$, $p = .598$; Feedback: .005 No Feedback: .004 ; Time-pressure: $F_1(1,23) = 0.043$, $p = .637$; Timed: .004; Untimed: .004). Likewise, an independent ANOVA for planning deletion rate per word for Time-pressure (2: Timed vs. Untimed) revealed no significant effects (Time-pressure: $F_1(1,23) = 0.607$, $p = .444$). Thus, there does not seem to be any clear association for the function of disfluency in the present experiment.

3.13 Discussion

As outlined in Chapter 2, there are two hypotheses which make predictions regarding the effects of time-pressure and feedback on disfluency. The first hypothesis, the Strategic-Modelling view, may predict high disfluency rates throughout a dialogue when listener feedback is involved but no effect of time-pressure. Next, the Cognitive Burden view may also predict high disfluency rates when a listener is involved but highest rates when both feedback and time-pressure are present. The purpose of the eye-gaze and disfluency analysis presented in this chapter is twofold: first we must ground the experimental paradigm to be certain that speakers respond to the novel visual stimulus. Once that is done, we can determine the chronology of interaction and whether a lost IF induces difficulty and therefore disfluency for the IG.

The results from this experiment show that time-pressure does make for shorter trials and less speech. A higher rate of Retrieval Transactions in the presence of listener feedback indicates that speakers do attend to the Follower's visual Feedback. Secondly, the fact that disfluency rate is highest in conditions in which visual feedback was present suggests a connection between the presence of feedback, Retrieval Transactions and disfluency. Perhaps, the Strategic-Modelling view does make the correct prediction and speakers use disfluency as a method for indicating their commitment to a listener.

A breakdown by disfluency type, however, reveals that only deletions are responsible for the higher disfluency rates with feedback. This finding suggests a possible link between deletions and the speaker's remedy for an errant IF. Clark & Wasow's (1998) commitment hypothesis relied upon repetitions as evidence for their theory that disfluencies can be used as signals. In contrast, the analysis presented in this chapter found a feedback effect with deletions, which are essentially a marker of the abandonment of a current utterance and therefore the exact opposite of a repair which involves repetition. A further detailed analysis of deletions in which deletions were grouped according to the likely function of their occurrence revealed that deletions can occur for planning (*Eh down the bottom bi-…You look like you're looking in the wrong spot)* or hesitative reasons (*Then right along to the right-hand corner of the page, there are….I think it said white mountain).* Although hesitative deletions were more common on the whole, there were no significant results for individual ANOVAs for either feedback or time-pressure. Overall, the analysis as whole shows that deletions (and for that matter any type of disfluency) that are classified according to a purely word surface structure coding system (e.g. Lickley, 1998) can in fact differ in origin and function. The planning deletions show that speakers were sensitive to the visual stimulus as they abandoned their utterances upon observation of it; for some, this might not be considered disfluent at all but just natural speaker behaviour given the experimental task. Hesitative deletions, on the other hand, can be conceived of as a genuine disfluency. Further investigation of the function of deletions in a more natural setting is required and will be discussed in Chapters 4 and 5.

At the outset of this chapter, I predicted according to the claims of Clark & Wasow's continuity hypothesis that one could determine why a speaker retraces an utterance rather than simply beginning where they left off. If speakers are repeating solely to facilitate a signalling function in production when they make a repetition, as predicted by Fox Tree and Clark (1997), it might be the case that the feedback manipulation would not affect the filled pause and repetition rates because they fulfil such specific functions. On the other hand, if speakers retrace utterances for the benefit of the listener or to maintain an acceptable social appearance, one would expect higher rates of repetition in interactive circumstances. Results in this chapter failed to find a significant difference between the no feedback and the interactive trials, although a near significant trend was observed. This result could suggest that perhaps the feedback manipulation used in this experiment did not permit speakers to fulfil the specific functions required to signal with repetitions and filled pauses because there was no difference in disfluency rate between feedback and no feedback trials. This does not however mean that repetitions and filled pauses are intentional signals in the manner suggested by Clark and Fox Tree (2002) and Fox Tree and Clark (1997). Closer examination of the patterns of individual subjects revealed that for 16 out of 24 speakers the mean repetition rate in the feedback condition was greater than the mean repetition rate in the no feedback condition. Thus, it would seem for the moment that individuals retrace for different reasons with some speakers ostensibly retracing for the benefit of the speaker. In the present experiment, speakers only had visual feedback from their listener. Perhaps repetition rates did not differ between conditions because speakers were not permitted any verbal feedback from participants. Further investigation which looks at the particular function of structural disfluency types is necessary.

An analysis of disfluency and feedback episodes revealed that disfluency was linked to situations during which the IF's feedback was in a wrong location. When speaker attention and disfluency were analysed together, however, there was only a significant effect of time-pressure. Thus, all we conclude is that Givers were more disfluent when faced with the task of having to re-align with a lost listener. In taxing situations, as predicted by Pickering and Garrod (2004) an IG encounters fluency problems when attempting to re-align with the IF. We must therefore conclude that we do not have any gaze-related evidence to support the Cognitive Burden view. Likewise, we cannot claim full support for the Strategic-Modelling view either because the current experiment did not find overwhelming support that Givers tracked their Followers assiduously, especially during periods when the Follower needed the most help. Furthermore, repetitions were not found to occur significantly more frequently in the feedback condition, as one would predict if they are truly being used as signals to the listener. For the moment, there is no strong pattern between disfluency and gaze.

Since this experiment used a surrogate for eye gaze rather than face-to-face interaction, there are a number of issues to consider regarding experimental control. For example, the experimenter was placed in charge of moving the red square in a timely and believable fashion. As previously mentioned, if the experimenter missed a single wrong landmark cue or more than 30% of the correct landmark cues, the trial was discarded. Six subjects were replaced because their data did not meet the 70% capture rate criterion. Furthermore, all subjects were questioned during debriefing whether they found anything 'odd' about the movements of the red square. The results of any subjects who disbelieved the visual feedback were also discarded. No subjects were removed in Experiment 1 for this reason.

Another issue that arises when considering the degree of experimental control is the fact that the red square provides both more (e.g. the precise location of the Follower) and less information (e.g. no facial cues or gestures) than is available during a face-to-face dialogue. This fact suggests that the results obtained in this experiment may pertain only to the specific paradigm. This does not mean that any results found in this experiment are invalid. I believe that the current paradigm is no less natural than phone conversations where interlocutors have only verbal feedback and yet still engage in collaborative dialogue or gaze experiments which require the participant to wear a head-mounted eye-tracker, a potentially unnatural situation for unpracticed participants. As in any experiment, however, the results reported in this chapter should be taken at face-value: when Givers are presented with a visual-only stimulus in a wrong location, they tend to incur more disfluencies when attempting to realign. As mentioned above, further experimentation on this issure is required to distinguish between the Strategic-Modelling and the Cognitive Burden views.

Since this experiment tested only the Giver's attention to visual feedback (i.e. a surrogate for real gaze), it shows only part of what happens in a real dialogue. In fact to face dialogues, interlocutors have access to both visual and verbal feedback. In such a scenario, the Strategic-Modelling view predicts that speakers will be capable of attending to both the visual and verbal feedback of the Follower without difficulty. The Cognitive Burden theory, on the other hand, predicts that disfluency will increase with task difficulty. In order to address these predictions, it would be worthwhile analyzing the relationships between disfluency, task difficulty and attention to the Follower's feedback. This analysis will be the subject of Chapter 4.

CHAPTER 4 – DISFLUENCY AND ATTENTION IN DIALOGUE

4.1 Introduction

Since the goal of this thesis is to address why disfluency occurs, an investigation that includes both visual and verbal feedback is necessary to approach what happens between people in real dialogue. In order to understand why disfluency occurs, one needs to know what sort of behaviours is associated with disfluency. Experiment 1 showed that Givers were more disfluent in the Feedback conditions when they could access visual feedback from the Follower. Such a finding is in line with the predictions of the Strategic-Modelling view, which predicts that speakers will signal commitment through disfluency only when a listener is present. When disfluencies were analyzed by type, however, only deletions were significantly more frequent in the presence of feedback. Fox Tree and Clark (1997) and Clark and Wasow (1998) predict that repetitions are signals of commitment made for the benefit of the listener. Experiment 1 showed that deletions, or abandoned moves, are actually more responsive which suggests that Clark and Wasow's predictions need to be revisited. In terms of Giver attention to gaze, the previous chapter found that Givers were more disfluent when the Follower feedback square hovered over a wrong landmark. This is evidence in support of the Cognitive Burden theory, which predicts that Givers' disfluency rate will increase with task difficulty. The current chapter will revisit these tests of disfluency rate by type to see whether disfluency rate is affected differently by visual or verbal feedback.

Another way to analyse the association between speaker behaviour and disfluency is to investigate what the speaker's dialogue goals were when he or she became disfluent. Givers who participated in Experiment 1 were found to be most disfluent during Instruct Moves when compared to Interactive Moves. What types of Transactions cause the speaker to be more disfluent? The current chapter will present an analysis of this sort.

Finally, as reported in Chapter 3, no reliable function of disfluency could be found from Experiment 1. The current chapter will revisit this issue by investigating the functions, i.e. planning or hesitation for repetitions and deletions. Recall that the Strategic Modelling View predicts that repetitions will be associated with a planning function, because they are made as signals of commitment for the listener.

Although Experiment 1 provided an indication of the distribution of disfluency relative to feedback from a listener, it said nothing about the distribution when Givers have both visual and verbal feedback. Thus, Experiment 2 uses the same screen-based task as Experiment 1, but gave Givers access to verbal feedback, to visual feedback, and in some cases both simultaneously. Disfluencies were classified both structurally and according to the ‘dialogue goal’ and for the function of the repair. Analyses of disfluency, dialogue goal and gaze was then conducted, the results of which are explained in this chapter.

4.2 Rationale and Predictions

As mentioned in Chapter 3, the Strategic-Modelling view predicts that speakers will signal their commitments to utterance and interlocutor through disfluency. In terms of attention, the Strategic-Modelling view predicts that speakers will attend to the listener’s feedback throughout a dialogue, especially in circumstances when the follower deviates from the planned route. According to one possible prediction of the Strategic-Modelling view, disfluency rate should be high once more in conditions where feedback is present and low in monologue conditions. Alternatively, as observed in Chapter 3, if the signalling function of repetitions and filled pauses is highly specialised, it might be the case that the feedback manipulation used in Experiment 1 did not allow speakers to make these specific signals. In order to further rule out this possibility in Experiment 2, we must make the feedback manipulation as close to dialogue as possible by adding verbal feedback. Furthermore, according to the predictions of the Strategic-Modelling view, the speaker should gaze most when the Follower is lost, that is at a wrong landmark or when the Follower indicates verbally with negative feedback that she needs help. According to the predictions of Clark and Wasow (1998), repetitions should occur for planning reasons because speakers use repetitions as signals of commitment to the utterance for the benefit of the listener.

In contrast, the Cognitive Burden view predicts that speakers will avoid attending to information when the cognitive cost of doing so is high. In terms of responsiveness, Givers will only respond to the Follower’s needs when the cognitive cost of doing so is low. This suggests that Givers do not monitor the listener. According to the Cognitive Burden view, disfluency rate is predicted to be high when task difficulty is also high. In response to this difficulty, Givers are predicted to avoid attending to the Follower’s feedback occasionally, even if she indicates that she is lost. If the Giver does attend to difficult feedback, for example a lost Follower, the Cognitive Burden view predicts that disfluency rates will increase in these situations because the speaker has to pay a cost for this difficulty. Similarly, a Giver who is in the process of Retrieving a lost Follower would be predicted to be more disfluent because retrieving a lost Follower is more difficult than simply describing the route from landmark to landmark.

Table 11. Table summarising the predictions for the Cogntive Burden and Strategic-Modelling Views with regards to the Independent variables Feedback, Time-pressure, and Group

	COGNITIVE BURDEN			**STRATEGIC-MODELLING**		
Dependent Variable	**Feedback**	**Time-Pressure**	**Group**	**Feedback**	**Time-Pressure**	**Group**
Disfluency Rate	High disfluency in the most difficult condition	High disfluency with time-pressure	No difference between groups	High disfluency in the most Interactive condition	No prediction	No difference between groups
Gaze Proportion	Avoid gazing during difficulty	Avoid gazing when difficult, i.e. with time-pressure	Group with most experience should gaze most, i.e. Visual group	Gaze at IF throughout	No prediction	Equal rates for both Visual and Verbal groups
Disfluency Types	No prediction	No prediction	No prediction	High repetition and filled pause rates	No prediction	No difference between groups
Disfluency Rate by Transaction Type	High disfluency in the most difficult Transaction Type (Retrievals)	Higher disfluency rate in time-pressure	No difference between groups	No change in disfluency rates across Transaction types	No prediction	No difference between groups
Disfluency and Gaze within a Feedback Episode	High disfluency when the Feedback is difficult to process	Higher disfluency rate with time-pressure	Verbal group should be most disfluent with visual feedback	High disfluency rates when the Follower is Lost	No prediction	No difference between groups
Function of Structural Disfluency Type	No prediction	No prediction	No prediction	Repetitions fulfil a signalling function Deletions show that the Giver is opportunistic.	No prediction	No difference between groups

Chapter 3, Section 3.2 summarised the predictions of the Cognitive Burden and Strategic-Modelling views with regards to the dependent and independent variables tested in Experiment 1. To these predictions Experiment 2 adds a between-subjects Group variable that tests whether a group that received Verbal feedback behaves differently in a Dual-feedback situation (i.e. both visual and verbal feedback are present simultaneously) from a group that received Visual feedback. As shown in Table 11, the Cognitive Burden view predicts that gazing should be easiest for those group participants who have had more exposure to the visual feedback, namely the Visual Group. The Strategic-Modelling view predicts that all speakers should attend to the listener's feedback with the same frequency, regardless of whether they've had visual or verbal feedback in earlier trials. For the analysis of the Function of disfluency, the Cognitive Burden view predicts that it should be difficult for Verbal group Givers to adjust to the addition of visual feedback and therefore they would be more disfluent than Visual group Givers. The Strategic-Modelling view predicts no difference in disfluency rates between groups in this case because both groups should signal equally.

With regards to the independent variables of Feedback and Time-pressure, the Cognitive Burden and Strategic-Modelling views make the same predictions as presented in Table 3 (page 98) in Chapter 3. If anything these predictions are enhanced by the addition of verbal feedback which makes the task more interactive on the one hand and therefore possibly harder for the speaker to manage on the other. If one modality of feedback conflicts with the other (e.g. the Follower says one thing but does another), then the speaker has the responsibility of clarifying the issue. In this respect, Experiment 2 is really a test of Cognitive Burden.

4.3 Experiment 2 Method

Experiment 1 showed evidence of poor uptake of visual wrong feedback, evidence which supports the Cognitive Burden view that speakers do not monitor their listeners during dialogue. Experiment 2[16] contrasted visual feedback with verbal feedback to assess whether the speaker responded differently to verbal feedback. In Experiment 2, the design crossed Time-pressure (2) with Feedback Modality (3). As for Experiment 1, Time-pressure in Experiment 2 could either be present or absent. In the timed condition of Experiment 2, Givers were limited to two minutes, a

8/5/078/5/07

[16] This experiment was run by Catriona Havard in the Department of Psychology at the University of Glasgow.

minute longer than in Experiment 1. The Feedback condition in Experiment 2 consisted of a no-feedback trial, a Single-Modality feedback trial and finally a Dual-Modality trial in which the speaker had access to both verbal and visual feedback. Experiment 2 actually consisted of two separate smaller experiments, Experiment 2A and Experiment 2B, which differed only in the Single-Modality. In Experiment 2A, the Single-Modality feedback consisted of only verbal feedback from a confederate participant. Experiment 2A Givers can be referred to as the 'Verbal Group'. In Experiment 2B, the Single-Modality feedback consisted of only visual feedback. Experiment 2B Givers will be referred to as the 'Visual Group'. Like the visual feedback, the verbal feedback was not derived from a naïve participant but instead a confederate who read from a prepared script. The confederate's comments were designed to reflect a lost follower with statements such as *I don't see it* as well as affirmations *Yeah, that's fine*. Confederates were requested to stay as close as possible to the provided script but could add backchannels like *yeah* or *Ok* when necessary.

In the Dual-Modality trials of both Experiment 2A and 2B, each trial was pre-programmed to contain both consonant and dissonant verbal visual feedback pairs so that a Follower's red square might be physically located on the correct landmark but the verbal feedback from the confederate reflects confusion (*I don't see it)*. Alternatively, the Follower's square could deviate off the route while the confederate responds as if she understands where she should be *(Yep, got that*). Finally, consonant feedback pairs (ie. visual-positive and verbal-positive; visual-negative and verbal-negative) also occurred. In this way, it was possible to test whether the speaker responded differently to the separate types of feedback. The visual feedback provides the speaker with an exact description of where the follower actually is at the moment where the verbal feedback conveys only a sense of where the speaker believes s/he should be. The theory of disfluency as a sign of cognitive burden predicts that speakers should be most disfluent in times of difficulty, hence when the follower has deviated off-course.

4.4 Experimental Procedure

The majority of the Experimental Procedure for Experiment 2 reduplicated the procedure used in Experiment 1. The same rooms, eye-tracking equipment, eye-tracking software, video recording software, and audio equipment were re-used. The role of confederate Information Follower was played by a different graduate student from the Psychology Department at the University of Glasgow.

As discussed in Section 3.5, there is a possibility that IGs could have used their gaze deicticly to point out the correct location because they believed that the IF could see their gaze. The experimenter once again explained that the IG would be able to see the IF's gaze but the IF wouldn not be able to see the IG's gaze. To be sure that no IGs used their gaze in a deictic fashion, the transcripts and videos from Experiment 2A and 2B were examined and no indication (i.e. explicitly deictic language indicating that gaze was being used to point as observed by Clark and Krych, 2004) was found. Two subjects were replaced because they were suspicious of the confederate participant.

Thirty-six participants from the community of The University of Glasgow partook in Experiment 2 in exchange for £5 per hour. The same subject criterion for normal uncorrected vision was upheld. A subject's data was discarded if the data did not meet the criteria for feedback or capture quality. The data from thirteen subjects were discarded because less their data fell below the 70% capture rate criterion. These subjects were replaced with an additional thirteen subjects so that a total of fifty-one participants were needed (i.e. and additional thirteen to uphold the 70% capture criterion and an additional two to uphold the "naïve" nature of the confederate) in order to obtain thirty-six usable trials. A copy of the instruction sheet and the consent form that the subjects were asked to sign is given in Appendix AA.

4.5 Experimental Design

Experiment 2 was run on 36 subjects according to 3 x 2 Repeated Measures design for Feedback Modality (3) x Time-pressure (2). Feedback-was within-subjects variable with 3 levels: no feedback, during which the speaker received no feedback whatsoever from the Follower, a Single-Modality feedback, during which the speaker received either visual or verbal feedback, and Dual-Modality feedback, during which the speaker received both visual and verbal feedback. Feedback type in the Single-Modality condition was a between-subjects variable. The 18 Subjects who participated in Experiment 2A received a verbal-only feedback condition and the 18 subjects in Experiment 2B received a visual-only feedback condition. Each subject participated in 6 trials using a new map each time. Time-pressure was a within-subjects variable. Time-limited trials had a 2 minute time limit; Untimed had no time-limitations.

4.6 Materials

Since there were six conditions, the same four maps from Experiment 1 were used again and an additional two maps from the HCRC Map Task corpus were re-used[17]. The addition of verbal feedback meant that more landmarks had to be added to all the maps so that each map contained 8 visual=correct verbal=positive landmarks, 3 visual= correct verbal= negative landmarks, 3 visual=wrong verbal=positive landmarks and 3 visual=wrong verbal=negative landmarks.

As explained in Chapter 3, Materials Section 3.4, the visual feedback used in the MONITOR project was designed to correspond to the mismatched landmarks that subjects encountered in the HCRC Map Task Corpus. Similarly, the verbal feedback was designed to correspond to the original mismatch. For example, if the IG had 2 Allotments landmarks, the first one on the route at the top of the page and the second one at the bottom of the page, the IF would indicate a mismatch by going to the second Allotments landmark and whilst providing negative verbal feedback *I don't see it*. The schedule of verbal and visual feedback for the Crane Bay map is shown below. All other schedules appear in Appendix F.

LM	**Verbal Response**	**Visual FB**
Start / Sandy Shore:	Ok got that.	Correct
Well:	*Ok, yes.*	Correct
Hills:	*Yep, fine*	Correct
Local Residents:	*Can't see it*	Correct
Iron Bridge:	I *don't see it*	Wrong
Wood:	*Okay, fine*	Correct
Forked Stream:	*Got it.*	Wrong
Farmed Land 1:	*Don't know where you mean.*	Wrong
Dead Tree:	*Okay, got it.*	Correct
Pine Grove:	*Ok, got that*	Wrong
Farmedland 2:	*Can't see it.*	Correct
Lagoon:	*Yep, got it.*	Wrong
Crab Island:	*Ok, I'm with you*	Correct
Rock Fall:	*No, not with you.*	Correct
CCSub[18]:	*Stop, where's that?*	Wrong
Pirate ship / Finish:	*Yes, ok.*	Correct

Figure 18. The schedule of verbal and visual feedback for the Crane Bay map

8/5/078/5/07———————————————

[17] See Appendix E for the full set of Experiment 2 maps.

[18] CCSub = Computer Controlled Submarine

Maps were paired and then run through a Latin Square design where order of presentation, experimental condition and subject were counter-balanced. This order can be seen in Appendix G. For example, the ‘Crane Bay’ map and the ‘Safari’ map were subjected to all of the experimental conditions. In order to ensure that there was no effect of map difficulty, the rate of words per number of landmarks was submitted to an ANOVA with a 6-valued independent variable for Maps. This ANOVA failed to retrieve a significant effect for Map (*F1*(5,170) = 1.44, *p* = .214). This suggests that although one of the maps had 18 landmarks (Pyramid), four maps had 17 landmarks (Diamond Mine, Mountain, Safari, Telephone Kiosk) and one map had 16 landmarks (Crane Bay), these maps took the same amount of work to complete.

4.7 Data Coding

The dialogues were transcribed and coded for disfluencies, Conversational Moves and Transactions in the manner explained in Chapter 3, Section 3.7.1. The eye-gaze data from the videos was coded frame by frame in Observer Pro software at The University of Glasgow[19], as explained in Chapter 3.

The procedure of Experiment 2 introduced a new form of data: verbal feedback from the Follower. Verbal feedback was first transcribed verbatim in a separate file from the transcript of the Giver. Later, this was coded at the University of Edinburgh[20] in a similar fashion to visual feedback as either ‘positive’ or ‘negative’. Positive verbal feedback *Yeah got it* indicates that the confederate Follower understood which landmark she was meant to find. Negative *No, not with you* feedback suggests that the confederate was confused and unaware. The verbal feedback was transcribed, time-stamped and then output into a separate XML file for analysis.

4.7.1 Coder Reliability

The Coder Reliability tests used for Experiment 2 were the same as those used for Experiment 1 (Chapter 3, Section 3.7.4).

8/5/078/5/07

[19] Thanks to Catriona Havard for assistance in this manner

[20] Thanks to Yiya Chen for overseeing the coding and transcription of all Experiment 2 dialogues.

4.7.2 Data Analysis

The data were analyzed in a similar fashion to the analysis method described in Chapter 3, Section 3.7.5. For analyses which involved the fine details of timing of disfluency with regard to gaze (see Section 4.9), I accessed the data by referring to the time-stamped XML files and by watching the MPEG video files. The results of such an analysis are described in Section 4.12.

4.8 Words and Speech Overall

Tables 12 and 13 show the overall distribution of transactions, words, disfluencies, filled pauses and average time a trial took for the Verbal and Visual groups, respectively. Transaction and word counts are broken down into Normal, Retrieval and Other (e.g. Irrelevant, Review and Overview) Transactions to show where the most speech occurs. Appendix H shows the Total distribution for both Experiment 2A and 2B combined as well as by subject.

4.8.1 Word Count

Word counts for whole and part-words show less speech with time-pressure (418 words/trial on average) than without (580): (F_1 (1,34) = 25.34, p < .001). Visual Group Single-Modality trials (461 words) were shorter than the corresponding Dual-Modality trials (585 words) (Feedback Modality x Group: (F_1(2,68) = 8.87, p < .001; Bonferroni: t = -6.6, p < .003, α < .003). For the Verbal Group, No Feedback trials (355 words) were shorter than both Single-Modality trials (545) and Dual-Modality trials (611) (Bonferroni: t = -5.87, p < .003, α < .003; Bonferroni: t = -5.22, p < .003, α < .003), which did not differ. The interaction between Feedback-Modality and Group is depicted in Figure 19.

Table 12 Total speech, total disfluencies, and average time spent for the Verbal Group (Experiment 2A)

	Timed			Untimed		
MEASURE	None	One	Dual	None	One	Dual
Transactions	**256**	**384**	**398**	**282**	**422**	**454**
Normal	252	263	279	276	304	324
Retrieval	0	106	116	0	106	122
Others	4	15	3	6	12	8
Words	**5235**	**8180**	**9134**	**7502**	**11417**	**12810**
Normal	5179	5261	5790	7386	7338	8769
Retrieval	0	2798	3305	0	3880	3967
Others	56	121	39	116	199	74
Time in Seconds	**121.81**	**189.33**	**214.94**	**186.66**	**277.73**	**311.66**
Disfluencies	**152**	**249**	**265**	**203**	**446**	**530**
Repetitions	59	84	87	95	205	251
Substitutions	54	91	87	58	120	122
Insertions	27	38	34	27	62	63
Deletions	12	36	57	23	59	94
Filled Pauses	**134**	**205**	**234**	**205**	**340**	**346**

Table 13 Total speech, total disfluencies, and average time spent for the Visual Group (Experiment 2B)

	Timed			Untimed		
MEASURE	None	One	Dual	None	One	Dual
Transactions	**239**	**259**	**342**	**303**	**334**	**427**
Normal	232	240	243	298	296	295
Retrieval	2	18	93	0	34	123
Others	5	1	6	5	4	9
Words	**6596**	**7443**	**8553**	**9133**	**9121**	**12443**
Normal	6403	7022	5819	9038	8257	8600
Retrieval	175	411	2711	0	819	3804
Others	18	10	23	95	45	39
Time in Seconds	**149.00**	**158.93**	**188.13**	**208.63**	**219.30**	**285.98**
Disfluencies	**150**	**180**	**240**	**183**	**219**	**366**
Repetitions	60	64	91	61	57	135
Substitutions	57	75	68	73	89	118
Insertions	22	18	23	29	33	42
Deletions	11	23	58	20	40	71
Filled Pauses	**157**	**181**	**221**	**280**	**259**	**369**

4.8.2 Word Count

Word counts for whole and part-words show less speech with time-pressure (418 words/trial on average) than without (580): (F_1 (1,34) = 25.34, p < .001). Visual Group Single-Modality trials (461 words) were shorter than the corresponding Dual-Modality trials (585 words) (Feedback Modality x Group: (F_1(2,68) = 8.87, p < .001; Bonferroni: t = -6.6, p < .003, α < .003). For the Verbal Group, No Feedback trials (355 words) were shorter than both Single-Modality trials (545) and Dual-Modality trials (611) (Bonferroni: t = -5.87, p < .003, α < .003; Bonferroni: t = -5.22, p < .003, α < .003), which did not differ. The interaction between Feedback-Modality and Group is depicted in Figure 19.

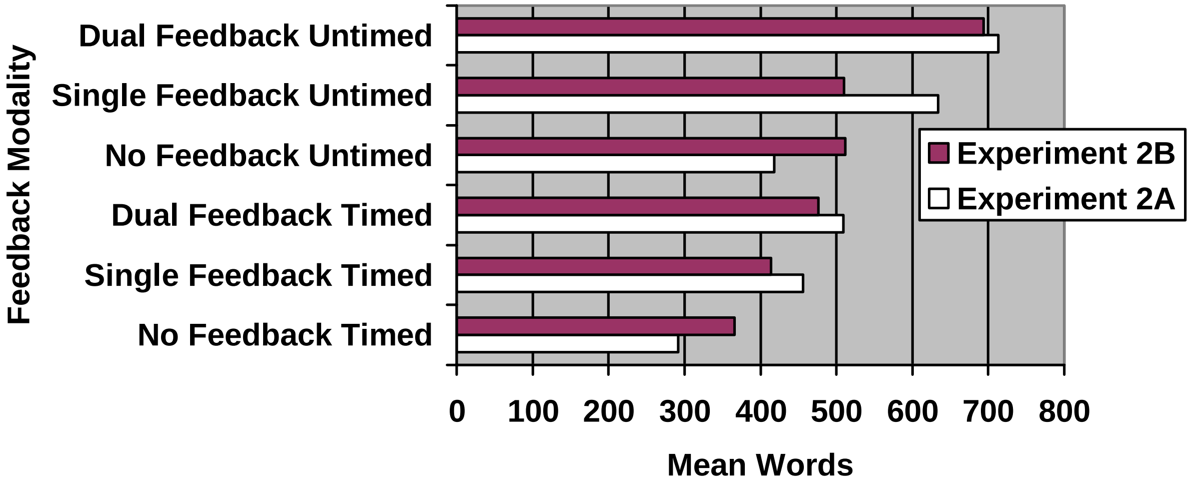


Figure 19. Raw word counts for both the Verbal and Visual Groups in Experiment 2

Since Dual-Modality conditions do not differ between groups (Verbal: 610, Visual: 584), we can use this condition to examine the relationships between disfluency and gaze or dialogue events.

4.8.3 Speech Rate

I also examined speech rate in order to be certain that experimental conditions are comparable for the disfluency analyses. To calculate speech rate, I divided the number of Giver words per map by the total Giver speaking time for the map (the summed durations of all conversational moves less the summed durations of both filled and simple pauses). Time-pressure had no significant effect on speech rate. The interaction between Feedback Modality and Group (F_1(2,68) = 4.87, p < .02) presented in Figure 20, is due only to a difference between the No-Feedback (.34) and Dual-Modality (.30) conditions for the Verbal Group (Bonferroni, p = .006,

α < .003). Again Dual-Modality conditions are alike.

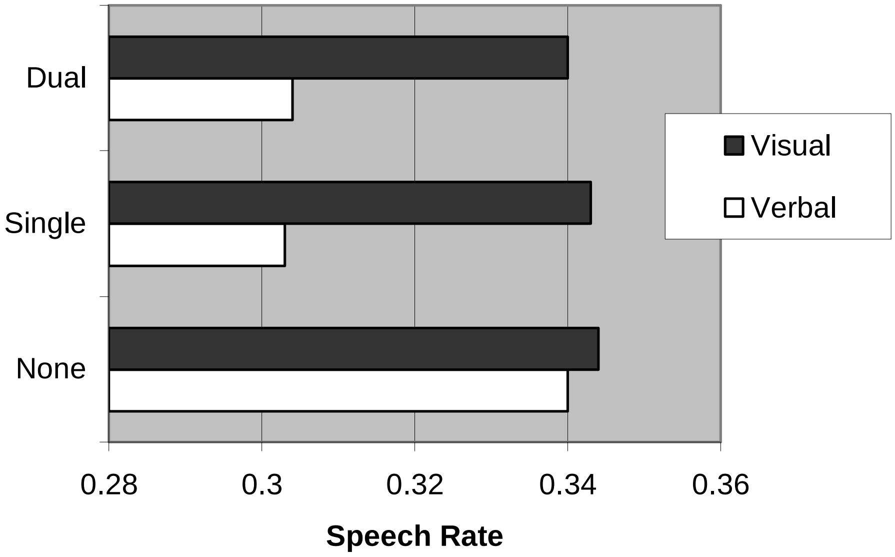


Figure 20. Mean Speech rate (word / Total speaking time) in seconds from Feedback Modality for the Visual and Verbal Groups

4.8.4 Transaction Rate

To understand how Givers break the route description down into sub-goals, I analyzed the number of Normal, Retrieval and Other transactions per trial. A Normal transaction occurs when the speaker provides instructions to get from one landmark to the next. The raw number of Normal transactions was submitted to an ANOVA for Feedback-Modality (3) x Time-pressure (2) x Group (2). The untimed conditions (16.6 per trial) contained more Normal transactions than the timed condition (13.97) (Time-pressure: $F_1(1,34) = 26.66$, $p < .001$). There were no other significant results or interactions for Normal transactions. The result for time-pressure reflects the fact that when speakers have more time, they will say more.

Recall that each map had 9 scheduled landmarks with either wrong visual feedback or negative verbal feedback (i.e. 3 visual=wrong verbal=positive landmarks, 3 visual=wrong verbal=negative landmarks and 3 visual=correct verbal=negative landmarks). The Giver could be expected to retrieve a lost Follower in any of these situations.

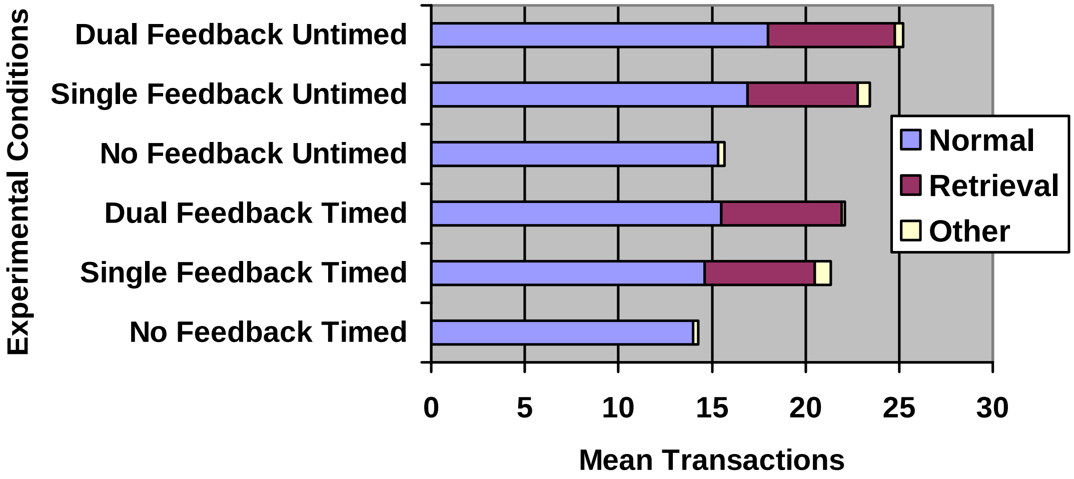


Figure 21. Transaction rate per trial by transaction type and condition in Experiment 2A: The Verbal Group

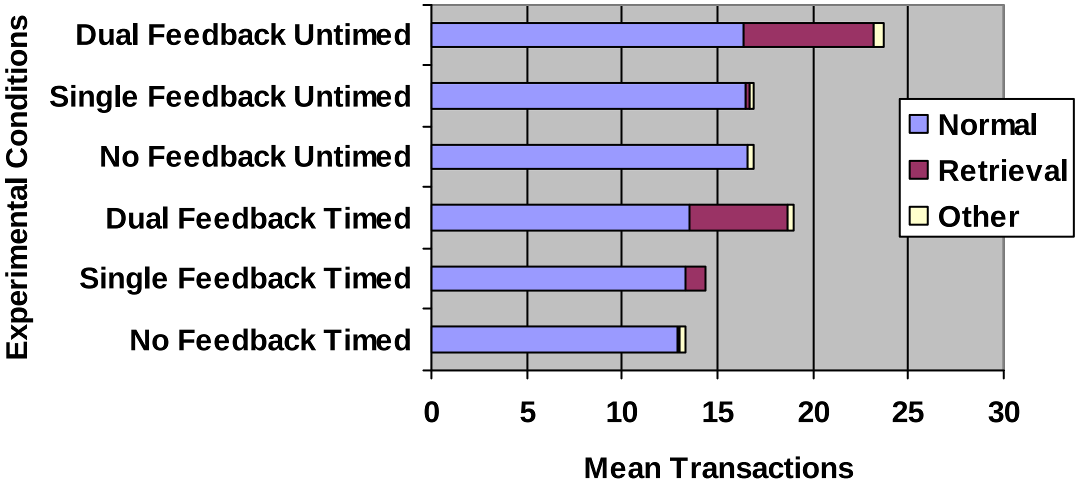


Figure 22. Transaction rate per trial by transaction type and condition in Experiment 2B: The Visual Group

Instead, a different pattern emerged when the rate of Retrieval transactions per trial was submitted to a Mixed Between and Within by-subjects ANOVA for Feedback Modality (3) x Time-pressure (2) x Group (2). In terms of Feedback Modality, Verbal Group Givers made more retrievals in their Single-Feedback modality (5.89) than Visual Group Givers made in their Single-Feedback modality (1.44) (Feedback Modality x Group: $F_1(2,68) = 55.44$, $p < .001$; Bonferroni $t = 10.04$, $p < .001$, $\alpha < .003$). This result suggests that Givers respond more often to verbal feedback than to visual feedback. The same interaction also revealed that Verbal Group Givers made more Retrieval transactions in the Single-Feedback (5.89) and the Dual-Feedback

(6.61) modality than in the No Feedback modality (.000) (Bonferroni *t*-tests, $p < .001$, $\alpha < .003$). Visual Group Givers made more Retrievals in the Dual-Feedback modality (6.00) than in either the Single-Feedback (1.44) or No Feedback (.056) modality (Bonferroni *t*-tests, $p < .001$, $\alpha <$ .003).

In terms of Time-pressure, Givers in the Verbal Group retrieved more often in both the timed (4.11) and untimed (4.22) conditions than Visual Group Givers did in their timed (2.09) or untimed (2.91) conditions (Time-pressure x Group: $F_1(1,34) = 6.36$, $p < .02$; Bonferroni *t*-tests, $p < .001$, $\alpha < .008$). An interaction between Time-pressure and Feedback showed that Givers retrieved more often in the Timed Dual-Feedback modality (5.81) than in either the Timed Single-Feedback (3.44) or Timed No Feedback (.056) modality (Time-pressure x Feedback Modality: $F_1(2,68) = 4.94$, $p = .01$; Bonferroni *t*-tests, $p < .001$, $\alpha < .003$). The difference between Single-Feedback (3.44) and No Feedback (.056) modalities was significant (Bonferroni $t = -12.16$, $p < .001$, $\alpha < .003$). In the Untimed condition, the same pattern emerged: Givers retrieved more in the Dual-Feedback modality (6.81) than in the Single-Feedback (3.89) or No Feedback modality (.000) (Bonferroni *t*-tests, $p < .001$, $\alpha < .003$). Once again, the difference between Single and No Feedback modalities was significant (Bonferroni $t = -12.47$, $p < .001$, $\alpha < .003$).

The rate of Other (Review, Overview and Irrelevant) transactions per trial was also submitted to an ANOVA for Feedback Modality (2) x Time-pressure (2) x Group (2). There was a significant interaction between Feedback Modality x Group ($F_1(2,68) = 4.83$, $p < .02$) which showed that Verbal Group Givers had the highest rate of Other transactions in the Single-Feedback modality (.75) compared to any of the other cells (Verbal Group: No Feedback: .306; Dual: .306; Visual Group No Feedback: .278; Single: .139; Dual: .417). Post-hoc comparisons for this interaction were not significant, however.

4.9 Gaze

Recall from Chapter 2 that Anderson et al. (submitted) report raw gaze patterns (i.e. average time spent fixating on feedback) which show that speakers in the MONITOR Project avoided gazing at the Follower feedback when it hovered over a wrong landmark. For Experiment 2, Anderson et al. report that when speakers were presented with verbal feedback, their tendency to gaze at the Follower increased. These results show that verbal feedback affects Giver gaze in terms of total fixation. Is the same result true for an analysis of Giver gaze per feedback episode? This measure is important because a similar measure, disfluency per gazed at episodes, will be used later in order to test the relationship between disfluency and Giver attention. Furthermore, it is necessary to determine whether all conditions (e.g. verbal = negative visual = correct versus verbal = wrong visual = wrong) in which a Giver might gaze at a feedback square actually succeeded in directing the Giver's attention to the square.

To check for overlap of gaze between Giver and purported Follower feedback, the video record of feedback and Giver Gaze were analyzed frame by frame for the landmark at which each was directed. When Follower Gaze and Giver Gaze were on the same landmark, the Giver was considered to be looking at the feedback square. The No Feedback condition has shown us what the baseline gaze time if for landmarks when there is no feedback from the Follower. A feedback episode, or task sub-portion containing feedback, starts with the departure of the feedback square for a landmark and continues until the feedback square departs for the next landmark. A by-subjects ANOVA with the number of feedback episodes as the dependent variable was run with Group (2: Experiment 2A vs. Experiment 2B), Verbal Feedback (2: positive vs. negative) and Visual Feedback (2: correct vs. wrong) as independent factors.

Givers did not make use of all their opportunities by any means (Figure 23). Nor did they use their opportunities equally in all conditions (Visual Feedback x Verbal Feedback: $F_1(1,34) = 7.70$, $p < .01$). Strangely enough, Givers used fewest opportunities in an important concordant condition, where the Follower was clearly lost: the Follower square was hovering over a wrong landmark while the Follower simultaneously provided negative verbal feedback (verbal=positive visual=correct: .366).

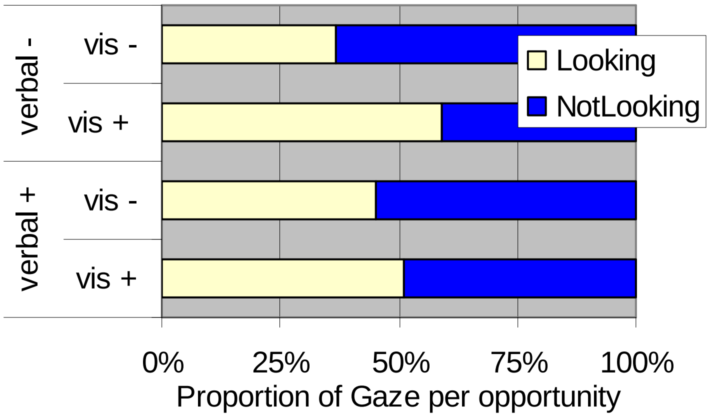


Figure 23. Proportion of feedback episodes attracting speaker gaze to feedback square in Experiment 2A and 2B: Effects of combinations of visual and verbal feedback in dual channel conditions

These attracted fewer looks than another concordant condition – when the Follower needed no help because she was in the right place and said so (verbal=positive, visual=correct: .511).

Similarly, Givers looked less when the Follower was lost but claimed not to be (verbal=positive visual=negative: .448) than when she was correct but claimed to be lost (verbal=negative visual=correct: .591) (Bonferroni *t*-tests, $p = .005$, $\alpha < .008$). Put simply speakers are most likely to track listeners, when the listener's location falls under their own gaze, which is occupied by the things they are describing. Apparently, speakers prefer not to go off-route to learn the whereabouts of an errant Follower.

4.10 Disfluencies

Experiment 1 found that speakers were more disfluent in interactive circumstances. How does the addition of verbal feedback in Experiment 2 affect this tendency? In order to answer this question, I plotted disfluency rate under each experimental condition. Once again, because disfluencies are more common in longer utterances (Bard et al, 2001; Oviatt, 1995; Plauché & Shriberg, 1999) the disfluency rate per total words was plotted. Rates were calculated for entire trials speaker by speaker. Overall, there were 3183 speech repair disfluencies (Verbal Group: 1845; Visual Group: 1338) and 2931 filled pauses (Verbal Group: 1464; Visual Group: 1467).

Disfluency rates from both experiments were submitted to a Mixed by-subjects ANOVA for Group (2: Verbal vs. Visual) x Time-Pressure (2: timed vs. untimed) x Feedback Modality (3: none, Single or Dual-Modality). Interactive conditions were more prone to disfluency: the Dual-Modality condition (.030) and the Single-Modality (.028) condition were both more disfluent than the No Feedback (.024) condition (Feedback Modality: $F_1(2,68) = 8.04$, $p = .001$). There was no significant difference between the Dual-Modality and Single-Modality conditions for disfluency rates alone. The results for the Verbal Group are shown in Figure 24 and the results for the Visual Group are depicted in Figure 25. Since Single and Dual-Modality conditions do not differ, we can proceed to examine only the Dual-Modality conditions in the expectation that conflicting feedback (only found in the Dual Modality) *per se* is not an overall cause of disfluency.

A by subjects ANOVA of disfluency rate per words which included filled pauses revealed a similar finding to the one just reported about disfluency rates alone. Givers were more disfluent in the Dual-Modality condition (.058) than in the No Feedback condition (.05) (Feedback Modality: $F_1(2,68) = 4.98$, $p = .01$). There was no significant difference between the Single-Modality and Dual-Modality condition. As in the ANOVA of disfluency rates alone, the ANOVA including filled pauses revealed no significant interactions at all.

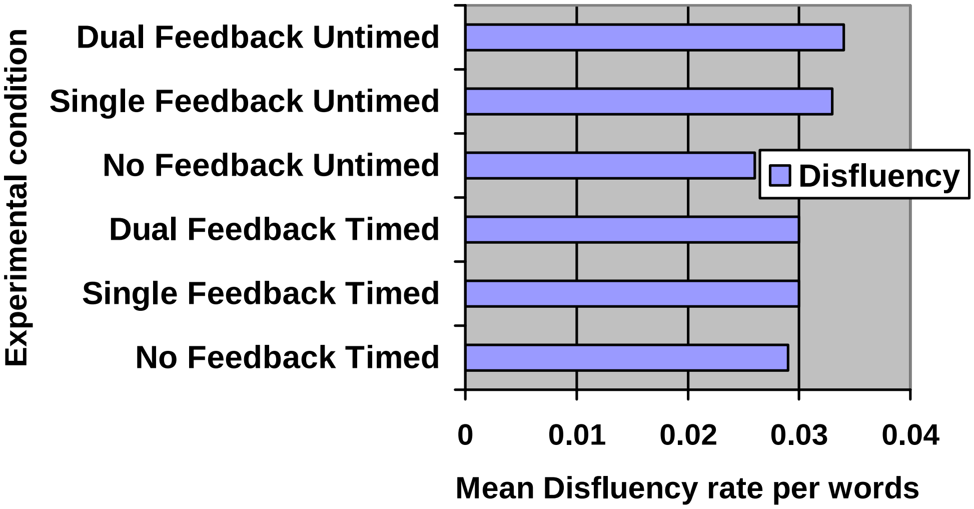


Figure 24. Rates of Disfluencies per fluent words for Experiment 2A, Verbal group**.**

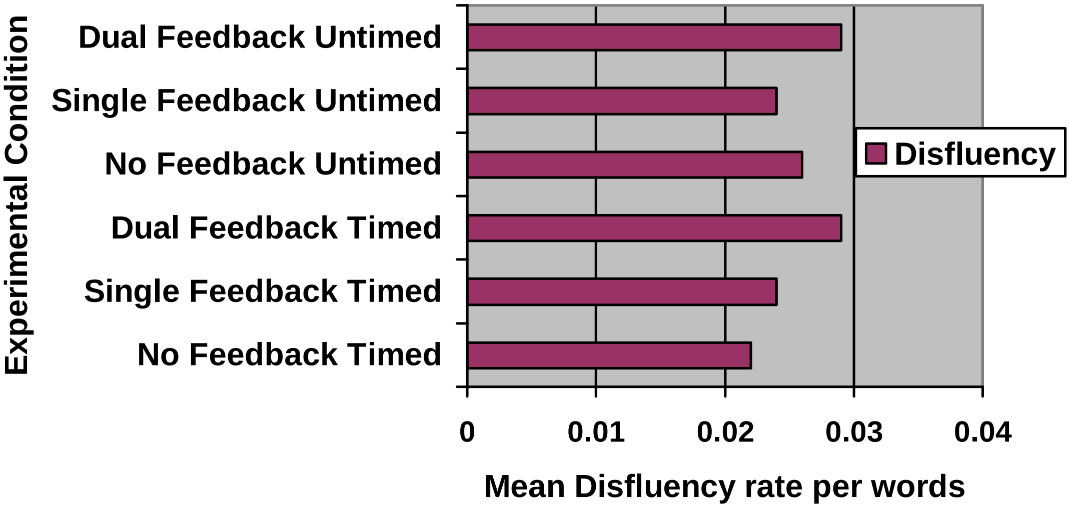


Figure 25. Rates of Disfluencies per fluent words for Experiment 2B, Visual group

4.10.1 Disfluency Types

Experiment 1 found a result contrary to the predictions of the Strategic-Modelling view that repetitions might occur more frequently in interactive circumstances if they are true strategic signals. Could it be the case that the paradigm in Experiment 1 wasn't interactive enough because it didn't have verbal feedback? To test this hypothesis and to determine the frequency of different disfluency types according the experimental conditions, I conducted independent analyses for each type of disfluency. The total counts of disfluencies in Experiment 2A and 2B are shown in Table 14. An initial investigation of disfluency types overall suggests differences between deletions and repetitions. Figure 26 displays the distributions of all disfluency types across experimental conditions for the Verbal Group. Figure 27 displays the distributions of disfluency types across experimental conditions for the Visual Group. Disfluency rate in this instance was calculated by dividing the total number of disfluency of a given type by the number of words for that trial.

Table 14. Distribution of disfluencies according to type in the Verbal and Visual Groups

	Repetitions	Substitutions	Insertions	Deletions	Filled Pauses
Verbal Group	781	532	251	281	1464
Visual Group	468	480	167	223	1467

As was the case for Experiment 1, deletion rate showed a significant effect of feedback: Deletion rate rose significantly with each additional Feedback Modality (No Feedback .002, Single .004, Dual .007; $F_1(2,68) = 21.02$, $p < .001$). There were no effects of time-pressure on deletion rate and no significant interactions. A by-subject ANOVA of only Verbal Group subjects revealed that speakers made more deletions in the Dual-Modality condition (.007) than in the No Feedback modality (.003) (Experiment 2A: Feedback Modality: $F_1(2,34) = 7.65$, $p < .01$). Similarly, an by-subject ANOVA for only the Visual Group revealed that speakers deleted more often in the Dual-Modality condition (.006) than they did in either the Single-Modality condition (.004) or the No Feedback condition (.002) (Experiment 2B: Feedback Modality: $F_1(2,34) = 15.28$, $p < .001$).

Repetition rate was submitted to an ANOVA for Feedback-Modality (3) x Time-pressure (2) x Group (2). Although there was a significant interaction between Time-pressure x Group, internal comparisons were not significant ($F_1(1,34) = 6.16$, $p < .02$; Experiment 2A Timed: .009; Untimed: .012; Experiment 2B Timed: .009; Untimed: .008). Though Verbal Group speakers considered alone showed no effect of conditions the Visual Group subjects had a higher repetition rate in the Dual-Modality condition (.011) than in either the Single-Modality (.007) or the No Feedback Modality condition (.007) (Feedback Modality: $F_1(2,34) = 6.66$, $p < .01$).

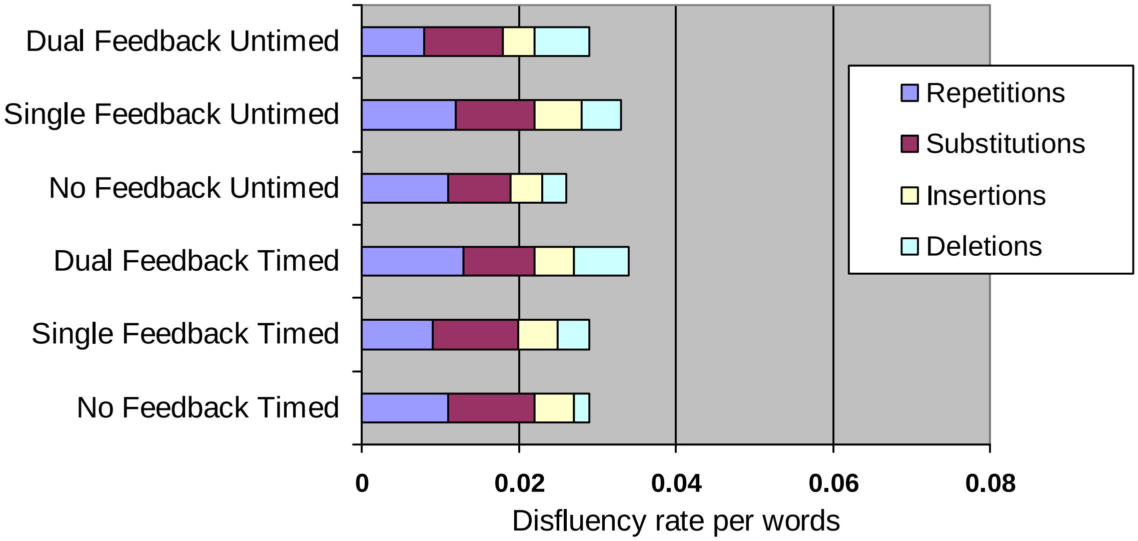


Figure 26. Rates of disfluency by type and experimental condition for Experiment 2A, the Verbal Group

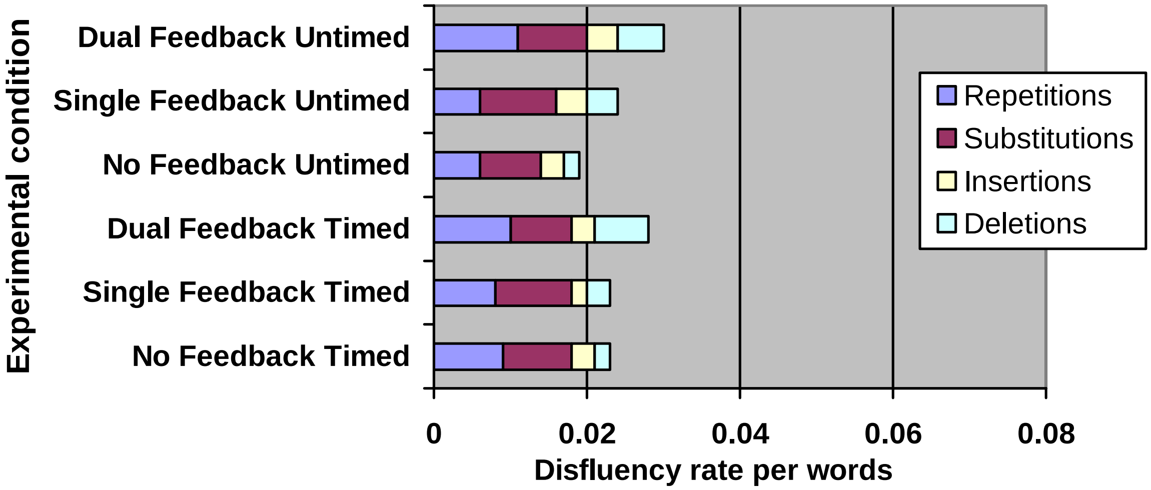


Figure 27. Rates of disfluency by type and experimental condition for Experiment 2B, the Visual Group.

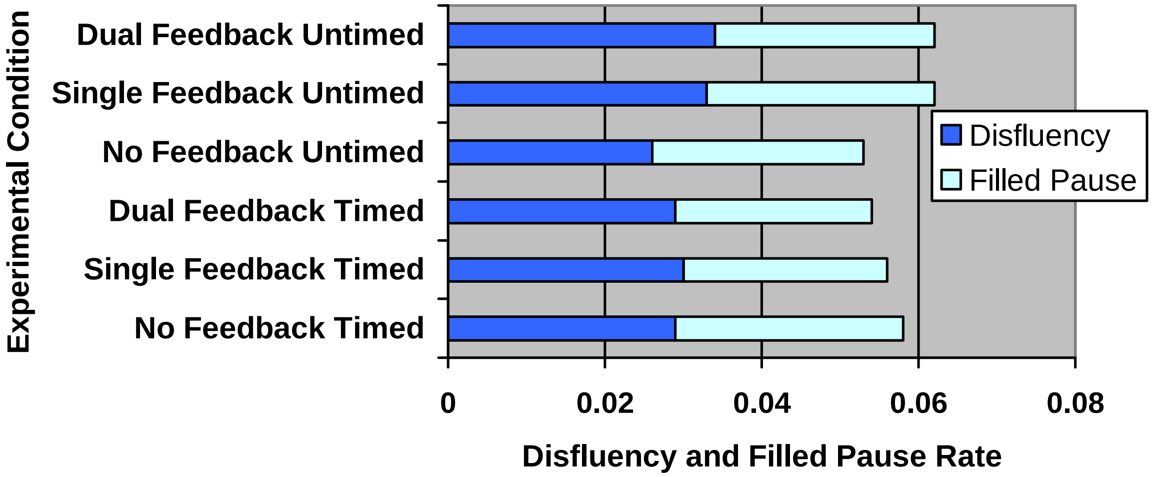


Figure 28. Rates of Disfluencies and Filled Pauses for Experiment 2A Verbal group

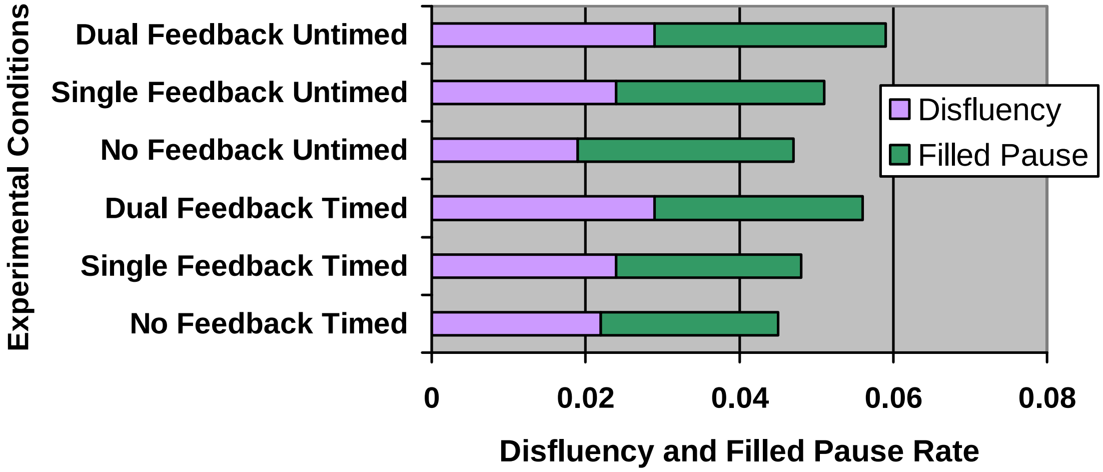


Figure 29. Rates of Disfluencies and Filled Pauses for Experiment 2B, Visual group

ANOVA for Feedback Modality (3) x Time-pressure (2) x Group (2) showed no effects on substitutions rate. Similarly, Insertion rate was also submitted to an ANOVA of Feedback Modality (3) x Time-pressure (2) x Group (2). Speakers from the Verbal Group (.005) made more insertions than speakers from the Visual Group (.003) (Group: $F_1(1,34) = 6.02, p < .02$).

A disfluency type analysis revealed an expected result. Deletions were once again associated with highly-interactive environments. As shown in Chapter 3, it could be the case that speakers make deletions for planning purposes. It makes sense that deletions should occur most in the Dual-Modality conditions because in these conditions the speaker has access to both visual and verbal feedback and can therefore see when the Follower's square diverts off-course or hear signs of hesitation or uncertainty. In these cases, the speaker abandons an utterance they were currently producing in favour of assisting the Follower or to provide confirmation that the Follower has reached the targeted landmark. As shown in Chapter 3, it could be the case that speakers make deletions for planning reasons. Repetitions also showed an effect of Feedback Modality for the Visual Group so the Strategic-Modelling prediction that the feedback manipulation used in this experiment would not affect repetition rate is disconfirmed, at least for the Verbal Group. This suggests that repetitions might also occur for planning reasons. An investigation of these two types of disfluencies can show not only the differences between them but can also be used to distinguish between the Cognitive Burden and Strategic-Modelling views. I will investigate this issue in Section 4.13.

4.10.2 Filled Pause Rate

Figures 28 and 29 (page 153) show the rates of filled pauses per words for Experiment 2A and Experiment 2B, respectively. Independent by-subjects ANOVAs of filled pause rate per fluent word failed to reveal any significant effects for Feedback ($F_1(2,68) = .303$, $p = .740$). There was a near significant trend for Time-pressure: Givers made more filled pauses in Untimed conditions (.028) compared to Timed (.025) ($F_1(1,34) = 3.31$, $p = .078$). As shown in Section 4.8.1, untimed trials were lengthier in terms of words. Bard et al. (2001) and Oviatt (1995) have shown that disfluency rate increases in longer trials. It could be the case that filled pause rate increases as a function of the number of words in a trial. This fact is not investigated further here since the effect did not reach significance.

4.11 Disfluency rate in Transactions

Section 4.8.4 showed that Normal transactions were as expected the most common type of Transaction and that Retrieval Transactions were more common in the presence of feedback than in the No Feedback condition. Chapter 3 investigated the rate of disfluencies within particular Conversational Move types to find evidence of the speaker's goals with regard to speech when they became disfluent. In the current section, I extend this analysis to Transaction types to determine whether disfluency was associated with any transaction-level speech goals. Results are reported as disfluency rate by transaction type: the number of disfluencies of a given type for any particular speaker within that speaker's transactions of the given type divided by the speaker's total words uttered within transactions of that given type. The dependent variable was then submitted to a by-subject ANOVA where Transaction Type (2: Normal vs. Retrieval), Time-pressure (2), Feedback Modality (2) and Group (2) were independent factors. The No Feedback conditions were omitted in this ANOVA because as Section 4.8.4 explained and Figures 21 and 22 (page 146) illustrate, there are few retrieval transactions in these conditions.

4.11.1 Overall Disfluency Rate

Overall, disfluency rates were higher in Retrieval transactions (.037) than in Normal transactions (.030) (Transaction: $F_1(1,34) = 14.50$, $p = .001$). A significant interaction between Transaction x Group, however, revealed that the disfluency rate of Retrieval transactions for the Verbal Group (.046) was higher than the rate of disfluency in Normal transactions in the same group (.031) (Transaction x Group: $F_1(1,34) = 14.94$, $p < .001$; Bonferroni $t = -5.17$, $p < .001$, $\alpha <$ .003).

Visual Group Givers were more disfluent in the Dual-Feedback modality (.033) than they were in the Single-Feedback modality (.023) (Feedback Modality x Group: $F_1(1,34) = 6.24$, $p =$ .02; Bonferroni $t = -3.25$, p = .005, α <.008). In terms of Time-pressure, Givers were more disfluent overall in Untimed Retrieval transactions (.042) than they were in Untimed Normal Transactions (.031) (Transaction x Time-pressure: $F_1(1,34) = 4.51$, $p < .05$; Bonferroni $t = -3.69$, $p = .002$, α <.008).

The Transaction x Group interaction which showed a significant difference for the Verbal Group might be due to a particular subject in the Verbal Group who made many more disfluencies than most subjects. An ANOVA for disfluency rate per words in transaction was rerun without the outlying subject. There were no differences between these results and the previous ANOVA for disfluency overall.

4.11.2 Repetitions

An ANOVA for repetition rate within transactions for Transaction Type (2) x Feedback Modality (2) x Time-pressure (2) x Group (2) including the outlier revealed that Visual Group Givers made more repetitions in the Dual-Feedback modality (.012) than they did in the Single-

Feedback visual modality (.006) (Feedback Modality x Group: $F_1(1,34) = 15.41$, $p <.05$; Bonferroni $t = -5.62$, $p < .001$, $\alpha < .008$). There was also a significant interaction between Transaction x Group ($F_1(1,34) = 8.48$, $p <.05$; Normal Transaction: Verbal: .011; Visual: .01; Retrieval Transactions Verbal: .015; Visual: .008) but post-hoc tests were not significant.

Once again, the ANOVA for repetition rate for Transaction Type (2) x Feedback Modality (2) x Time-pressure (2) x Group (2) was rerun without the outlying subject from the Verbal Group. Without the outlying subject, Givers made more repetitions in Retrieval Transactions in the Dual-Feedback condition (.014) than in Retrieval Transactions in the Single-Feedback condition (.007) (Transaction Type x Feedback Modality: $F_1(1,33) = 5.50$, $p <.05$; Bonferroni, $t = -3.764$, $p = .002$, $\alpha < .008$). Givers also made more repetitions in the Dual-Feedback in Retrieval Transactions (.014) than in Normal Transactions (.008) in the Single-Feedback modality (.008) (Bonferroni, $t = -3.364$, $p = .004$, $\alpha < .008$).

The outlying subject, subject 15, appears to have made enough repetitions in both Normal and Retrieval Transactions to cancel out these effects for the whole Group when his rates were included. In fact, most of the disfluencies made by this subject were repetitions at the beginning of a clause, for example where the repetition reparandum is highlighted in bold text *and* ***then we are going to****...we are going to travel along eh just the top of that missionary camp* or this example from the same dialogue *em and* ***then you are going to bear****...you are going to bear east* ***until you come to the****...until you come to the extreme right of the stones.* The means of subject 15 are compared to the other 17 subjects' means in Table 15. The repetitions for the other 17 Verbal Group Givers tended to have shorter reparanda than reparanda in Subject 15's repetitions, for example where the reparandum is indicated in bold: *If you just* ***go ehm…****go up the page about half an inch.* A further audio example of the outlying subject's disfluency is provided in Appendix I.

4.11.3 Substitutions

An ANOVA for substitution rate per words in transactions for Transaction Type (2) x Feedback Modality (2) x Time-pressure (2) x Group (2) including the outlying subject revealed a significant main effect of Time-pressure ($F_1(1,34) = 10.40$, $p <.01$; Timed: .006; Untimed: .009) as well as two significant three-way interactions. An interaction between Transaction Type x Time-pressure x Group showed that Visual Group Givers made more substitutions in Untimed Retrieval Transactions (.014) than Givers from the same group made substitutions in Timed

Retrievals (.005) (Time-pressure x Transaction Type x Group: $F_1(1,34) = 10.68$, $p = .002$; Bonferroni, $t = -4.229$, $p = .001$, $\alpha < .002$).

Table 15. Mean raw disfluencies by type across all conditions for the Verbal Group

	Repetitions	**Substitutions**	**Insertions**	**Deletions**
Subject 15	62.83	10	6.83	8.67
Verbal Group without subject 15	3.96	4.62	2.06	2.24
Verbal Group with subject 15	7.23	4.92	2.32	2.60

Verbal Group Givers made were more prone to substitutions in Untimed Retrieval transactions (.015) than they were in Untimed Normal transactions (.01) (Bonferroni $t = -3.89$, $p = .001$, $\alpha < .002$). A Between Group effect showed that Verbal Group Givers made more substitutions in Timed Retrieval Transactions (.016) than Visual Group Givers made substitutions in Timed Retrieval Transactions (.005) (Bonferroni, $t = 4.84$, $p < .001$, $\alpha < .002$). Although there was a significant interaction between Transaction Type x Time-Pressure x Feedback, internal comparisons were not significant ($F_1(1,34) = 6.03$, $p < .02$).

An ANOVA for substitution rate for Transaction Type (2) x Feedback Modality (2) x Time-pressure (2) x Group (2) without the outlying disfluent subject revealed the same results for the Transaction Type x Time-pressure x Group interaction that were observed when the outlying subject was included. One three-way interaction showed significant results without the outlying subject: the Verbal Group were more prone to Retrieval Transactions in the Single-Modality (.017) than speakers from the same group were prone to Normal Transactions in Single-Feedback modality (.011) (Transaction Type x Feedback Modality x Group: $F_1(1,33) = 4.27$, $p < .05$; Bonferroni, $t = -4.62$, $p < .001$, $\alpha < .002$). Thus, it seems that the only difference the outlying subject was responsible for in substitutions was the difference between Normal and Retrievals in the Single-Feedback Modality. The outlying subject made enough substitutions in both transaction types to cancel out the overall Group effect.

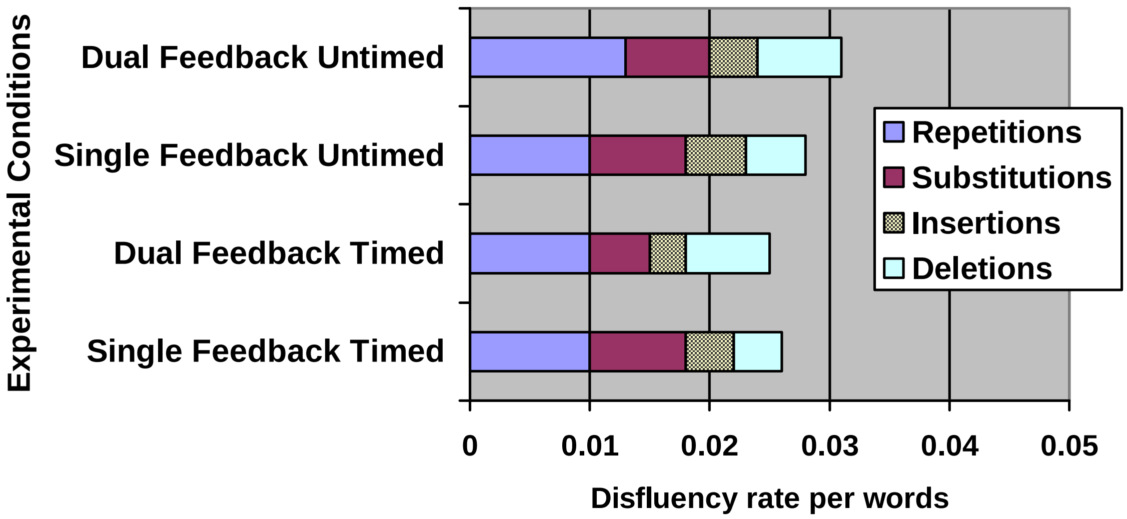


Figure 30. Disfluency rate per words in Normal Transactions for Verbal and Visual Groups of Experiment 2

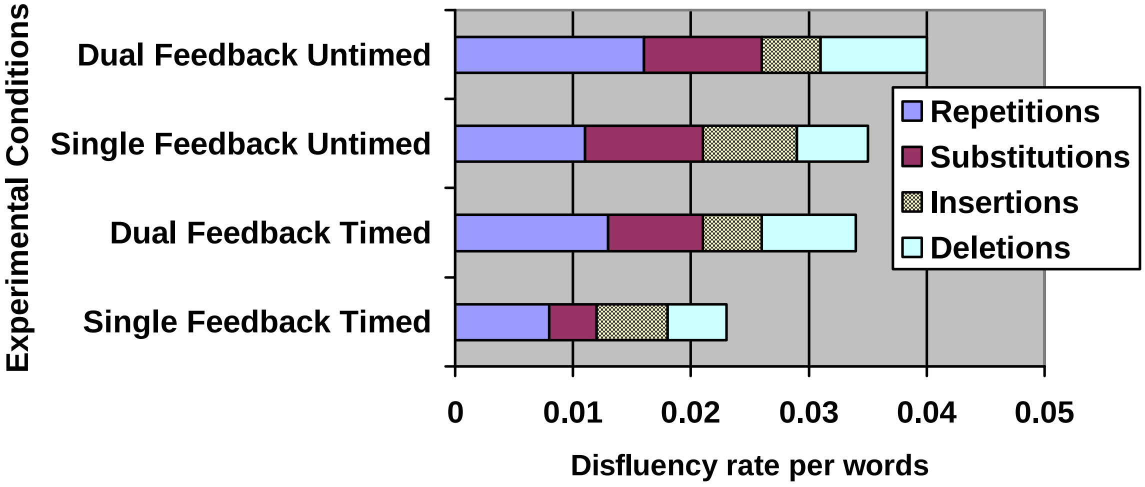


Figure 31. Disfluency Rate per words in Retrieval Transactions for Verbal and Visual Groups in Experiment 2

4.11.4 Insertions

An ANOVA for insertion rate for Transaction Type (2) x Feedback Modality (2) x Time-pressure (2) x Group (2) per words revealed only that Retrieval Transactions (.006) were more prone to disfluency than Normal Transactions (.004) ($F_1(1,34) = 5.33$, $p <.05$). There was no change to this effect when the outlying subject was removed from the ANOVA.

4.11.5 Deletions

Finally, an ANOVA for deletion rate revealed that deletion rate was higher in the Dual-Feedback modality (.008) than in the Single-Feedback modality (.005) (Feedback Modality: $F_1(1,34) = 7.57$, $p < .01$). Givers from the Verbal Group made more deletions per word in Retrieval Transactions (.009) than in Normal Transactions (.005) (Transaction Type x Group: $F_1(1,34) = 7.26$, $p <.02$; Bonferroni, $t = -3.70$, $p = .002$, $\alpha = .008$). There were no changes to the effects when the outlying subject was removed from the ANOVA.

4.11.6 Summary

This section has investigated speaker disfluency behaviour by looking at which Transaction types were more prone to disfluency. Overall, Retrievals seem to be more prone to higher rates of disfluency, although there were reasons to expect that an outlying subject influenced some of the results. For the analysis of all disfluency types combined, an ANOVA including the outlying subject revealed that Retrievals were more prone to disfluency than Normal Transactions, but only for the Visual Group and not for the outlying subject's group, the Verbal Group. When the outlying subject was removed from the Verbal Group, the effect remained non-significant for the Verbal Group and significant for the Visual Group. From this, we must conclude that the Visual Group was simply more disfluent overall in Retrieval transactions than speakers in the Verbal Group, even without the outlier.

ANOVAs of types of disfluency showed that Repetitions were more common in the Dual-Feedback Modality than in the Single-Feedback Modality; once again this was only true for the Visual Group. An ANOVA without the outlying subject, however, showed that Dual-Feedback Modality was more disfluent than the Single-Feedback Modality, suggesting that the outlying subject did play a role when it came to repetition rate.

To summarise for each of the Experimental conditions, Time-pressure affected only the Visual Group's substitution rate in Retrieval transactions. Visual Group Givers made more substitutions in Untimed Retrievals than in Timed Retrievals. The only important effect of

Feedback Modality is an interaction between Feedback Modality and Group that involves the Single Feedback Modality because this is the only condition where the groups differed. There were no differences of this sort for any individual disfluency type or all disfluencies considered together. The only differences of Feedback Modality found for Deletion rate within Transactions and Repetition rate within Transactions further confirmed the results found in Experiment 1 and Section 4.10 that interactive circumstances like the Dual-Feedback Modality are more prone to disfluency than the No Feedback Modality. Finally, Retrievals seem to be more prone to disfluency than Normal Transactions. Could this be indicative of a cognitive burden that subjects experience? Possibly, but one could also argue that subjects were attempting to send their listeners signals by being disfluent at critical moments. We shall turn to an analysis of Disfluency and Eye Gaze in the next section to understand speaker gaze behaviour during disfluent periods.

4.12 Disfluency Gaze within a Feedback Episode

Within the Dual-Modality condition, the experimental design contrasted positive and negative feedback in the two modalities. However, the modalities can be concordant or discordant only if the Giver actually takes up both visual and verbal feedback. The tendency for more speech in conditions with verbal feedback suggests that subjects were attending to what the confederate Follower said. Eye-tracking enabled us to tell when the Giver had actually looked at the Follower's visual feedback. Time-pressure has tended to reveal significant effects in the untimed conditions where Givers have more time to say more. As depicted in Figure 23 (page 148) and as explained in Section 4.9, Givers do not take up the same proportion of concordant and discordant feedback. They gazed most at one kind of discordant feedback (negative verbal + correct visual) and least at a concordant condition (negative + wrong feedback) when the Follower is in trouble and acknowledges that fact.

To look for disfluency in truly versus potentially concordant and discordant feedback situations, we examined disfluency per feedback opportunities in concordant and discordant situations contrasting those in which Givers did or did not look at Follower feedback. The dependent variable, disfluency per 'looked at' feedback episode, was submitted to a Mixed Within and Between by-subjects ANOVA with Group (2) x Concordance (2: concordant feedback vs. discordant) x Time-pressure (2) as independent variables.

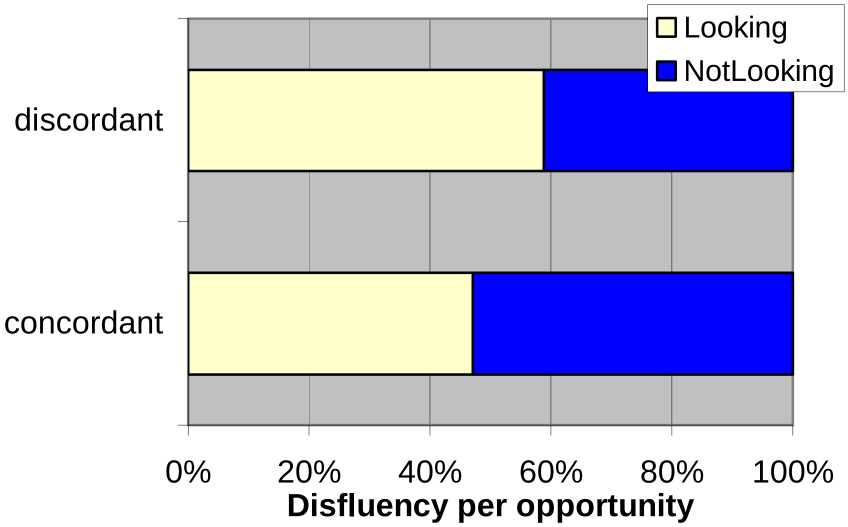


Figure 32. Proportion of disfluent concordant or discordant feedback opportunities with respect to whether the Giver was either looking or not looking at the Follower. The agreement difference is significant when the Giver looked at the Follower.

In fact, Givers who attended to discordant feedback from the Follower subsequently became disfluent. The number of disfluencies per feedback opportunity was greatest following a discordant feedback episode in which the Giver had actually gazed at the Follower feedback square (.325), a significantly higher rate than following a concordant feedback episode which had drawn the Giver's attention (.206) (Concordance: by subject: $F_1(1,34) = 9.60$, $p = .004$). One type of discordant landmark (verbal = positive, visual = wrong) (.624) attracted more disfluency when gazed at by the Giver than a concordant (verbal = positive, visual = correct landmark (.264) suggesting that Givers pay a cost for processing difficulty feedback (Visual Feedback x Verbal Feedback: by materials: $F_2(1,4) = 14.58$, $p < .02$) . Givers were more disfluent in Untimed episodes (.328) than in Timed episodes (.203) (Time-pressure: by subject: $F_1(1,34) = 7.28$, $p = .011$). This difference could be due to the fact that Givers say more in Untimed episodes, as shown in Section 4.8.1.

Recall from Section 4.9 that subjects avoided gazing at lost Followers who indicated both verbally and visually that they were having difficulties. The fact that Givers did not look at lost Followers supports the Cognitive Burden theory by suggesting that subjects found full uptake of Follower knowledge to be a difficult task. Overall, this section has shown that Givers tended to be more disfluent when presented with discordant feedback, which in conjunction with the General Gaze results from Section 4.9 point towards the Cognitive Burden hypothesis. The Strategic-Modelling View predicts that Givers will gaze at their Followers, especially in times of need.

The results presented here suggest that Givers do not always find it easy to gaze at their Followers and they make more disfluencies in the discordant condition as evidence of that fact.

4.13 The Function of Structural Disfluency Types

As shown in Chapter 3, deletions may occur for more than one reason. In some cases, it became clear that what the speaker said was classified as a deletion simply when the speaker changed their speech plan because of some external change in the Follower feedback, e.g. the Follower interrupted the speaker with verbal feedback or the speaker was interrupted by the moving feedback square. These instances show that the speaker needed to alter his or her plan of the discourse when, for example, she sees that the current utterance is redundant because the visual feedback shows that the goal has been achieved or that realignment is urgently required. Since these planning deletions are more interactive by nature, it is quite likely that the Feedback effect whereby deletion rate increases in more interactive circumstances is due to these planning deletions and not to the hesitation fresh starts. For this reason, I conducted separate ANOVAs of 'planning' and 'hesitation' deletions.

Section 4.10 and 4.11.2 showed that repetitions can be sensitive to the manipulations of Feedback Modality as well. Repetitions were found to occur more frequently in the Dual-Modality condition of the Visual Group of Experiment 2B than in the Single-Feedback Modality. This suggests that repetitions could stem from at least two different functions, planning and hesitation. For this reason, repetitions will be included in an analysis of the function of disfluency. Table 16 shows examples of planning and hesitation repetitions and deletions while Tables 17 and 18 show the distribution in raw numbers of planning and hesitation deletions and repetitions, respectively for the Verbal and Visual Groups.

This section presents the results of an analysis of the functions of two types of disfluency to determine whether any one type of disfluency is more associated with a particular function. Data were collected by watching MPEG videos for 1204 repetitions and 482 deletions in all conditions of Experiments 2A and 2B. Video examples and transcripts are included in Appendix J. A disfluency was considered to be from a 'planning' function if the movement of the visual feedback square or the verbal response from the confederate interrupted the speaker whilst speaking. All other cases were marked as occurring for a hesitation reason. Textual examples are provided in Table 16. Since there is no external function of feedback in the No Feedback Modality, this condition was omitted from the analysis explained below. The dependent variable for the analysis was calculated by dividing the number of planning disfluencies or the number of hesitation disfluencies by the number of words in the trial.

Table 16. Examples of disfluencies by goal and type. For repetitions, both reparandum and repair appear in bold text. For deletions, just the reparandum appears in bold text since the repair is effectively non-existent.

	Dialogue Goal	
Disfluency Type	Planning	Hesitation
Repetition	**'No ga- No gazelles?'**	'Eh you travel directly ehm sort of **north … north** and east'
Deletion	'So loop around the waterfall **over**…Yeah, there'	'Um **can you si-** … it's to the left of the pine grove'

Table 17. Distribution of planning and hesitation deletions across experimental conditions and within the Verbal and Visual Groups. Totals are given in bold text.

	Timed			Untimed		
FEEDBACK	None	One	Dual	None	One	Dual
Verbal Group	**11**	**35**	**56**	**19**	**56**	**93**
Planning	0	6	28	0	8	41
Hesitation	11	29	28	19	48	52
Visual Group	**11**	**23**	**57**	**18**	**34**	**69**
Planning	0	13	48	0	15	44
Hesitation	11	10	9	18	19	25
TOTAL	**22**	**58**	**113**	**37**	**90**	**162**

Table 18. Distribution of planning and hesitation Repetitions across experimental conditions and within the Verbal and Visual Groups. Totals are given in bold text.

	Timed			Untimed		
FEEDBACK	None	One	Dual	None	One	Dual
Verbal Group	**60**	**81**	**84**	**101**	**194**	**234**
Planning	0	9	17	0	9	22
Hesitation	60	72	67	101	185	212
Visual Group	**60**	**64**	**86**	**61**	**57**	**123**
Planning	0	6	20	0	9	22
Hesitation	60	58	66	61	48	101
TOTAL	**120**	**145**	**170**	**162**	**251**	**356**

The dependent variable, rate of planning disfluencies per words in the trial, was submitted to an ANOVA for Feedback Modality (2) x Time-pressure (2) x Group (2). Givers made more planning deletions in the Dual-Feedback Modality (.004) than they made planning repetitions in the Dual-Feedback Modality (.002) or planning deletions in the Single-Feedback Modality (.001) (Feedback Modality x Disfluency Type: $F_1(1,34) = 8.90$, $p < .01$). There were no significant between-subjects effects or any other significant interactions.

The rate of hesitation disfluencies per word was also submitted to an ANOVA for Disfluency Type (2) x Feedback Modality (2) x Time-pressure (2) x Group (2). A three-way interaction revealed that Visual Group Givers made more hesitation repetitions in the Single-Feedback modality (.006) and the Dual Feedback modality (.008) than speakers from the same group made hesitation deletions in either the Single-Feedback (.002) or the Dual-Feedback modality (.002) (Disfluency Type x Feedback Modality x Group: $F_1(1,34) = 4.28$, $p < .05$; Bonferroni, *t*-tests, $p < .002$, $\alpha. < .002$). A further interaction showed that Visual Group Givers made more hesitation repetitions in both Timed (.008) and Untimed (.007) conditions than they made hesitation deletions in either Timed (.001) or Untimed (.002) conditions (Disfluency Type x Time-pressure x Group: $F_1(1,34) = 6.70$, $p < .02$; Bonferroni, *t*-tests, $p < .002$, $\alpha. < .002$).

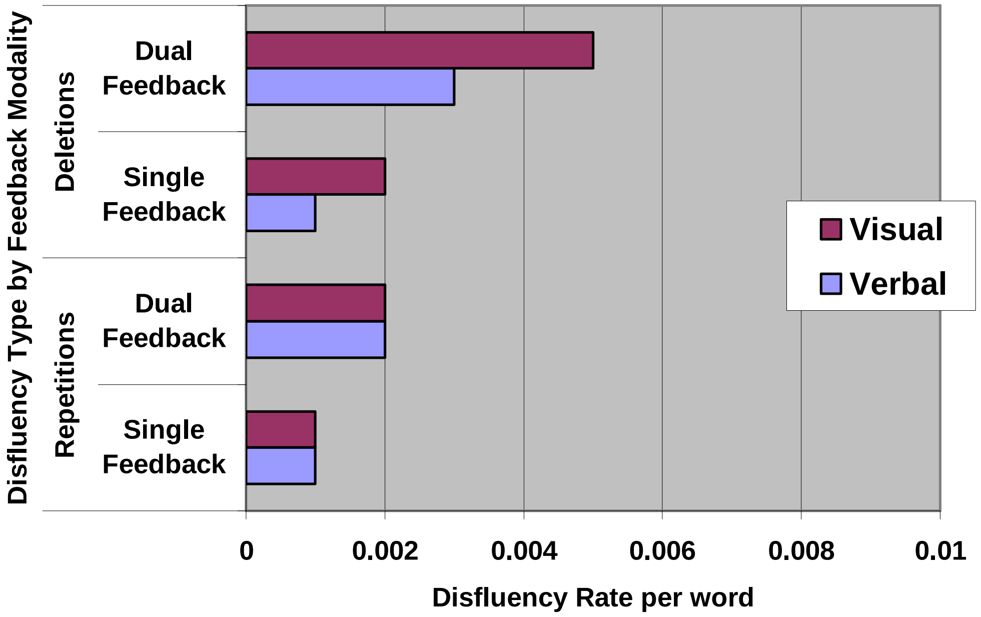


Figure 33. Planning Disfluency Rate per words by Feedback Modality and Group (Verbal vs. Visual)

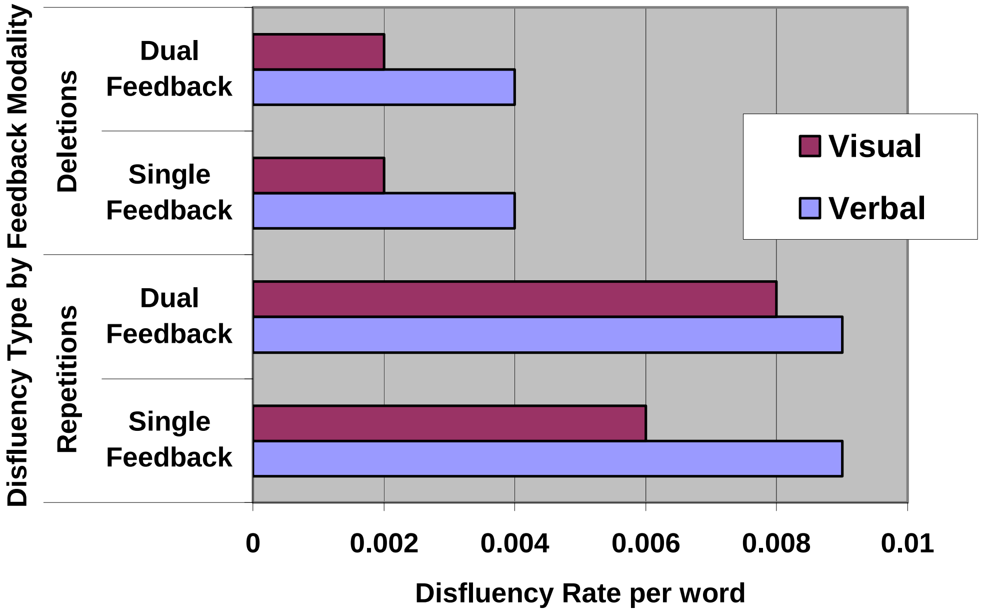


Figure 34. Hesitation Disfluency Rate per words by Feedback Modality and Group (Verbal vs. Visual)

Recall from Section 4.11 that the Verbal Group contained the outlying subject. The results just reported show significant effects for the Visual Group, the group without any outliers, but no significant results for the Verbal Group. The outlying subject could be the cause of this difference and so an ANOVA with his data removed is warranted. The rate of planning disfluencies was submitted to an ANOVA for Disfluency Type (2) x Feedback Modality (2) x Time-pressure (2) x Group (2). With the outlying subject excluded, a previously non-significant three-way interaction became significant: Visual Group Givers made more planning deletions in the Dual Feedback Modality (.002) than Verbal Group Givers made planning repetitions in either the Single-Feedback Modality (.001) or the Dual-Feedback modality (.001) (Disfluency Type x Feedback Modality x Group: $F_1(1,34) = 8.45$, $p < .01$; Bonferroni, *t*-tests, $p < .002$, $\alpha. < .002$). Visual Group Givers made more planning deletions in the Dual-Feedback Modality (.002) than speakers from the same group made planning repetitions in either the Single-Feedback (.001) or the Dual-Feedback modality (.002) (Bonferroni, *t*-tests, $p \leq .002$, $\alpha. < .002$). Even when the outlying subject was excluded from the analysis, there were no significant results for the Verbal Group. One possible explanation is that Visual Group Givers, who had more exposure to visual feedback, might be more sensitive to this modality of feedback and for that reason make more planning deletions following its movement. Alternatively, in order for the verbal feedback to interrupt the speaker, the confederate must provide her verbal response exactly when the Giver was about to speak or in some cases must interrupt the speaker. The visual feedback, which was controlled by an experimenter running a computer simulation, may not be as polite as the confederate human and for this reason might be more prone to interrupt speakers than a confederate.

An ANOVA of hesitation disfluency rate for Disfluency Type (2) x Feedback Modality (2) x Time-pressure (2) x Group (2) without the outlying subject revealed significant differences for the Verbal Group: Verbal Group Givers made more hesitation repetitions in both Timed (.005) and Untimed (.009) trials than Givers from the same group made hesitation deletions in Timed (.003) or Untimed (.004) trials or than Visual Group Givers made hesitation deletions in Timed (.001) or Untimed (.002) trials (Disfluency Type x Time-pressure x Group: Group $F_1(1,33) = 5.41$, $p < .05$; Bonferroni, *t*-tests, $p \leq .002$, $\alpha. < .002$). Likewise, as previously observed with the outlying subject, Visual Group Givers made more hesitation repetitions in both Timed (.008) and Untimed (.007) conditions than they made hesitation deletion in Timed (.001) or Untimed (.002) conditions (Bonferroni, *t*-tests, $p \leq .002$, $\alpha. < .002$). The fact that these results were significant when the outlying subject was removed suggests that his higher than average repetition rates in all conditions were responsible for the non-significance of the results when his data was included.

Thus, it seems that overall there is a tendency for repetitions to occur for reasons of hesitation. Overall, repetitions tended to be associated with hesitation functions.

For deletions, the rates of planning disfluencies were higher in the Dual-Modality than in the Single-Feedback Modality. This was also true, however, for hesitation deletions in the Dual-Feedback Modality so it is hard to determine whether deletions were more associated with one function over another. As shown in Figures 33 and 34 (page 165), the difference in planning and hesitation functions to deletion seemed to depend upon the group. The Verbal Group seemed to have a higher hesitation deletion rate while the Visual Group had a higher planning deletion rate. Recall that planning functions are the movement of the visual feedback square or the verbal reply of the confederate Follower while the Giver was speaking. It could be the case that the saccadic movements of the visual feedback square interrupt the speaker much more often than speech from the scripted confederate Follower ever could. Since the Visual Group encountered many more trials with visual feedback, one could suggest that possibly the Givers in these trials were more prone to interruption, thus explaining the higher occurrence of planning deletions.

In terms of the predictions of the Strategic-Modelling hypothesis and the Cognitive Burden hypothesis, this analysis has confirmed once again that two types of disfluencies, repetitions and deletions, stem from two possible functions. Participants in the experiment abandoned utterances in both the Single-Feedback and the Dual-Feedback modalities after attending to the movement of the visual feedback or after being interrupted by the verbal feedback of the Follower. Participants repeated more often for a hesitation reason, and not as the Strategic-Modelling hypothesis might predict, for the benefit of the hearer. Whether the hesitation repetitions are genuine intended signals of commitment as the Clark view would predict is known only to the speaker and is not immediately apparent in the current experimental paradigm.

4.14 Discussion

At the outset of this thesis, the predictions of the Strategic-Modelling view and Cognitive Burden hypothesis were contrasted. One version of the Strategic-Modelling view predicts that the Giver will attend to the Follower's feedback throughout the dialogue and that disfluency rate could increase when the Giver has access to Follower feedback. In contrast, one version of the Cognitive Burden view predicts that the Giver will avoid responding to the Follower if the cost of doing so is high. Therefore, disfluency rate is predicted to rise in times of difficulty, for example when the Giver's language production system is over-burdened.

Givers were more disfluent in the Dual-Feedback condition than they were in the No

Feedback condition, thus meeting the predictions of both the Strategic-Modelling and Cognitive Burden view that disfluency occurs most when interlocutors are co-present. Recall that the Cognitive Burden view predicts that feedback itself increases difficulty, thereby increasing disfluency rate. Another explanation for why disfluency rate may increase in the feedback condition is that perhaps the speaker tries harder to be understood when they have both visual and verbal feedback. In a sense, by trying harder to be understood when feedback is present, the speaker is changing his or her own level of difficulty but is doing so only when feedback is present.

As far as disfluency types are concerned, the distribution of disfluency types by Feedback Modality revealed that deletions increased significantly in the Dual-Feedback condition compared to the No Feedback condition. This finding partially supports the results found in Chapter 3: deletions occur more frequently when a speaker has access to Follower feedback. There was also a Feedback effect for repetition rate in the Visual Group where repetition rate was higher in both the Dual-Feedback and Single-Feedback Modality than in the No Feedback Modality. The results also suggest that repetition rate is also sensitive to the manipulations of the Feedback Modality. Since an effect of feedback was found, we have some evidence to rule out the possibility that the signalling function of repetitions is so highly specialised that the feedback manipulations used in Experiment 2 might not have created the necessary situations for an effect to occur. Note, however, that such a result did not occur until verbal feedback was added to the Visual Group; perhaps the verbal feedback enhanced the reality of the visual feedback.

Clark and colleagues suggest that repetitions are signals of commitment (Clark & Wasow, 1998; Fox Tree & Clark, 1997). However, results presented in both the current and previous chapters suggest that deletions were also used as indications of planning. Therefore, I conducted an analysis of the potential functions of disfluency for a speaker in dialogue. Two functions were identified: during a planning disfluency the speaker is interrupted by the movement of the visual feedback or by the verbal feedback of the confederate. During a hesitation disfluency¸ the Giver elaborates on something already uttered by adding, correcting or deleting spoken material. Initial results showed that repetitions were associated with hesitation functions, whilst deletions tended to be labelled as occurring for a planning reason. This suggests that deletions are used as signals in critical points of the interaction and not repetitions as predicted by the Strategic Modelling view. The speaker seems to have behaved according to the Joint Responsibility and abandoned an unnecessary utterance as soon as he or she learned that it was unnecessary so that s/he could help re-route the Follower instead. In order to confirm this suspicion, an analysis of Giver attention was necessary.

The Strategic-Modelling view predicts Givers will gaze most when the Follower clearly indicates that she is lost, that is, in the concordant wrong condition where both verbal and visual feedback indicate that the Follower is lost. An analysis of Giver attention to the Follower's feedback showed that Givers used most opportunities to look at the Follower's square when the Giver was presented with one kind of discordant feedback (visual=correct verbal=negative). Since the Giver would look at such on-route landmarks just to continue describing the route, the Giver has not gone out of her way to track the Follower in this instance. In contrast, when the Follower was lost in the concordant wrong condition Givers took less than fifty percent of the available opportunities to locate them. Therefore, Givers appear to prefer gazing at what is easiest for them rather than tracking a lost listener. This result supports the Cognitive Burden view.

A further prediction made by the Cognitive Burden theory is that Giver disfluency will increase in periods of difficulty. In order to answer this question, an analysis of disfluency and gaze was conducted. Givers were more disfluent when they had gazed at discordant feedback than at concordant feedback. Proponents of the Cognitive Burden view such as Barr and Keysar (2002) and Horton and Keysar (1996) suggest that merely having an interlocutor will increase difficulty and therefore disfluency should increase. Pickering and Garrod (2004) suggest that disfluency will arise out of misalignment in dialogue. Since discordant feedback is essentially misaligned feedback, both with the interlocutor and with itself, this prediction seems to be met. Furthermore, because the discordant feedback is difficult to process due to is conflicting nature, the predictions of the Cognitive Burden theory that disfluency increases with task difficulty are also met.

This chapter has investigated the relationship between disfluency, functions of disfluency and gaze in order to learn more about why disfluency occurs. In Chapter 2, I suggested that one way of answering this question is to determine whether structurally-classified disfluencies can be linked to a cognitive motivation or 'dialogue goal'. As this chapter has shown, structurally-classified disfluencies were indeed associated with certain goals: deletions tended to occur for planning reasons, although there were differences between the Visual and Verbal Groups. On the other hand repetitions were strongly associated with hesitation functions. Moreover, by classifying disfluencies according to their functions, we have observed that ideal versions of two prominent psycholinguistic theories are certainly just that, ideal. In real dialogue situations, speakers seem to be capable of attending to the listener's feedback at certain points while still avoiding attending to the listener's feedback at other points. When the speaker makes a planning deletion, the only thing the speaker is 'signalling' is a necessary change in direction for the Follower's feedback. At other times, the same speakers looked elsewhere on the map and made a repetition which stemmed from a hesitation function.

Finally, while the present chapter has revealed a great deal about the nature of disfluency in dialogue, it has still left a few questions remaining. One question is whether the speakers were perhaps just not motivated enough to perform the task because it was held in an experimental setting. In order to address this question, a further experiment, reported in Chapter 5, was designed and carried out. This experiment re-uses the basic design of Experiment 2, but adds a ‘Motivation’ condition in which speakers were offered additional incentive for highly successful trials.

CHAPTER 5 – MOTIVATION, DISFLUENCY, AND GAZE IN DIALOGUE

5.1 Introduction

Chapter 4 investigated the role of two types of feedback, verbal and visual, on disfluency. In conjunction with the predictions of the Cognitive Burden theory, Givers made more disfluencies during periods of difficulty and avoided looking at the Follower when it was difficult to do so. Proponents of the Strategic-Modelling View might suggest that the participants who performed as Information Givers were simply not committed enough to the task. A highly committed individual would be more likely to make collateral signals to their listener, according to Clark (2002). Fox Tree and Clark (1997) suggest that repetitions are one example of such a signal.

Proponents of the Cognitive Burden view would argue that the speakers' altruism competes with the demands of language production and is limited by the available cognitive resources (Bard et al., 2004; Horton & Keysar, 2004). Participants offered additional incentive would, according to the Cognitive Burden view, be expected to attend to the listener's feedback only when it was easy for them to do so. Disfluency rate is predicted to increase in difficult circumstances (Bard et al., 2001). In this Chapter, I report the results of an experiment which tested the effect that additional incentive had on the Giver during the dialogue to further tease apart these two hypotheses and furthermore how the function of disfluency maps onto the structure.

5.2 Rationale and Predictions

The current chapter sets out to test whether speakers who are offered extra compensation for optimal performance behave differently from a control group who were not offered additional compensation. The predictions of Experiments 1 and 2 were summarised in Table 3 (page 98) and Table 11 (page 135), respectively. The predictions with regards to Feedback, Time-pressure and Motivation, a between-subjects variable, are summarised in Table 19 below.

Table 19. Table summarising the predictions for the Cogntive Burden and Strategic-Modelling Views with regards to the Independent variables Feedback, Time-pressure and Motivation

	COGNITIVE BURDEN			**STRATEGIC-MODELLING**		
Dependent Variable	**Feedback**	**Time-Pressure**	**Motivation**	**Feedback**	**Time-Pressure**	**Motivation**
Disfluency Rate	High disfluency in the most difficult condition	High disfluency with time-pressure	More disfluencies for Group members with most difficulty	High disfluency in the most Interactive condition	No prediction	High disfluency if Givers are truly committed
Gaze Proportion	Avoid gazing during difficulty	Avoid gazing when difficult, i.e. with time-pressure	Motivated Givers may be willing to do difficult things, i.e. gaze at IF	Gaze at IF throughout	No prediction	Motivated participants should gaze more often than Controls
Disfluency Types	No prediction	No prediction	Deletion rate is high if Givers perform their responsibiliy	High repetition and filled pause rates	No prediction	Motivated Givers should make more repetitions
Disfluency Rate by Transaction Type	High disfluency in the most difficult Transaction Type (Retrievals)	Higher disfluency rate in time-pressure	No difference between groups	No change in disfluency rates across Transaction types	No prediction	No difference between groups
Disfluency and Gaze within a Feedback Episode	High disfluency when the Feedback is difficult to process	Higher disfluency rate with time-pressure	Motivated Givers should gaze more often and may be more disfluent	High disfluency rates when the Follower is Lost	No prediction	Motivated Givers and Controls will gaze and signal through disfluency
Function of Structural Disfluency Type	No prediction	No prediction	Deletions may fulfil a planning function	Repetitions fulfil a signalling function Deletions show that the Giver is opportunistic.	No prediction	Repetitions may fulfil a planning function

According to the arguments of Fox Tree and Clark (1997) mentioned above, speakers repeat themselves more often prior to complex utterances; these repetition disfluencies are a signal of commitment to the listener. Speakers who are more motivated should then produce more disfluencies, particularly repetitions, than speakers in the control group. Motivated speakers should also look more often at their task partners. If repetitions are signals of commitment to both the listener and the utterance, then one would expect repetitions to be associated with dialogue behaviour in which the goal of the utterance is to be attentive to the listener's feedback.

Speakers who are participating according to the Cognitive Burden view, would be expected to perform in an economical and cooperative manner with respect to the joint effort required from both speakers to accomplish the task. Rather than making repetition disfluencies to indicate commitment to the listener, a conscientious interlocutor who is behaving in a cooperative manner might be expected to abandon utterances in order to provide pertinent information. Motivated speakers are therefore predicted to exhibit a higher deletion rate in interactive circumstances than elsewhere. In terms of cost-sharing, the Cognitive Burden view would predict that speakers incur a cost for careful attention to their listeners' feedback. Therefore, one would predict that although motivated speakers may be more attentive, there is a cognitive cost for this effort that may be paid in terms of fluency.

5.3 Experiment 3 Method

With two exceptions, the same experimental method used in Experiment 2 was re-used in Experiment 3. Firstly, the Time-Pressure condition was eliminated since it only affected the amount of speech in Experiment 2. Secondly, a Motivation condition was added to test the effects of Motivation on the outcome of the task. Speakers were allocated to one of two groups, either the Control group or the Motivated Group. The Control Group consisted of nine participants from Experiment 2B who were told that they would receive £5 for their time. Only the untimed trials of the Control Group were used. Participants in the Motivated group were told that they would be offered £5 for their time regardless of how they performed, but if their description of the route on the map matched a certain criterion, they would be offered double their money, or £10 for the hour. The naïve participant and the confederate were then asked to decide amongst themselves who would be the Information Giver and who would be the Follower. The confederate always urged the naïve participant to perform the role of Giver and the naïve participant always agreed.

5.4 Experimental Procedure

The majority of the Experimental Procedure for Experiment 3 replicated the procedure used in Experiments 1 and 2. The same rooms, eye-tracking equipment, eye-tracking software, video recording software, and audio equipment were re-used. The role of confederate Information Follower was played by a different graduate student from the Psychology Department at the University of Glasgow.

Eighteen participants from the community of the University of Glasgow took part in Experiment 3. Nine participants, who had been paid £5 per hour, were taken from Experiment 2B; they constituted the Control Group. Nine Motivated participants were offered £10 per hour if they performed well or for £5 for their time. After each trial, each Motivated subject was paid £10 per hour. The same subject criterion for normal uncorrected vision was upheld. A subject's data was discarded if the data did not meet the 70% capture rate criterion for feedback or capture quality. In total, ten subjects were discarded because two were suspicious of the confederate's role in the experiment and eight did not meet the 70% capture criterion. All subjects were native English speakers. A copy of the instruction sheet and the consent form that the subjects were asked to sign is given in Appendix AA.

Motivated participants were paired with Control group subjects from Experiment 2B. Each Motivated subject saw the same maps in the same order and the same experimental condition as a Control group subject (Appendix E). 6 subjects (3 Motivated and 3 Control) had the 'Crane Bay' map, followed by the 'Diamond Mine' map, followed by the 'Pyramid' map. Another 6 subjects saw the 'Pyramid' map first, the 'Safari' map second and the 'Telephone Kiosk' map last. Finally, the last 6 subjects saw the 'Mountain' map first, the 'Telephone Kiosk' map second and the 'Crane Bay' map last. This ordering and pairing of subjects and maps means that the Crane Bay, Pyramid and Telephone Kiosk maps were each seen 6 times. The Pyramid and Crane Bay maps were each seen 3 times in the No Feedback condition and 3 times in the Dual-Feedback modality condition. The Telephone Kiosk map appeared 3 times in the Single-Feedback modality and 3 times in the Dual-Feedback modality. The other maps, Safari, Mountain and Diamond Mine, were only used a total of 3 times each in Experiment 3. The Safari and Diamond Mine maps appeared only in the Single-Feedback condition while the Mountain map appeared only in the No Feedback modality.

5.5 Experimental Design

The experiment crossed Feedback Modality (3) and Motivation (2: Control vs. Motivated). In the Dual-Feedback Modality condition, subjects received both visual and verbal feedback. Subjects in the Motivated group were offered double their money for excellent performance whilst subjects in the Control group were merely requested to do the task.

5.6 Materials

The six maps from Experiment 2 were reused in Experiment 3. Once again, maps were paired to make a balanced design. Maps can be found in Appendix E.

5.7 Data Coding

Following the completion of Experiment 3, the dialogues were transcribed and coded for disfluencies, Conversational Moves and Transactions in the same manner explained in Chapter 3, Section 3.7.

5.7.1 Coder Reliability

The Coder Reliability tests used for Experiment 3 were the same as those used for Experiment 1 (Chapter 3, Section 3.7.4) and Experiment 2 (Chapter 4, Section 4.13).

5.7.2 Data Analysis

The data were analysed using the method used for Experiment 1 (Chapter 3, Section 3.7.5) and Experiment 2 (Chapter 4, Section 4.7.2).

5.8 Words and Speech Overall

Table 20 shows the distribution of Transactions, Words, Disfluencies, and Filled Pauses for the Motivated and Control Givers in Experiment 3.

Table 20. Distribution of Transactions, Words, Disfluencies, and Filled Pauses for 9 Motivated and 9 Control Group Givers

	GROUP GIVERS					
	MOTIVATED			CONTROLS		
FEEDBACK MODALITIES	None	Single	Dual	None	Single	Dual
Transactions	**159**	**240**	**259**	**152**	**174**	**221**
Normal	144	168	172	151	150	156
Retrieval	0	52	75	0	21	61
Others	15	20	12	1	3	4
Words	**3070**	**6037**	**6446**	**4748**	**4737**	**6400**
Normal	3039	4246	4449	4702	4329	4608
Retrieval	0	1619	1921	0	365	1785
Others	31	172	76	46	43	7
Disfluencies	**52**	**119**	**171**	**183**	**221**	**383**
Repetitions	16	41	49	66	59	139
Substitutions	16	24	40	70	84	126
Insertions	8	15	16	28	35	44
Deletions	12	39	66	19	43	74
Filled Pauses	**67**	**77**	**74**	**144**	**142**	**177**

As Table 20 depicts, Control Group Givers have higher raw totals of disfluency than the Motivated Givers. On the whole, however, Control Group Givers were also more loquacious than Motivated Givers. Section 5.8.1 investigates raw word count to see whether this difference is significant. Appendix K shows these results by subject and trial for the 9 Motivated subjects and the 9 Control subjects.

5.8.1 Words

Previous studies (Bard et al., 2001; Oviatt, 1995) have shown that longer utterances give rise to higher disfluency rates. For this reason, it is important to be aware of the effects of the experimental design on raw word counts. As depicted in Figure 35 below, No Feedback conditions were the shortest. A by-subjects ANOVA revealed that Control Group Givers said more in terms of raw word count in the Dual-Feedback Modality condition (761.8 words) than they did in the No-Feedback condition (541.4) (Feedback Modality x Group: $F_1(2,32) = 4.94$, $p < .02$; Bonferroni t -test, $p < .003$, $\alpha < .003$). The same result was found for Motivated Givers: Givers in the Motivated group also said more in the Dual-Feedback Modality (731.6) than they did in the No Feedback condition (351.6) (Feedback Modality x Group: (F_1 (2,32) = 4.54, $p < .02$; Bonferroni, $t = -4.33$, $p < .003$, $\alpha < .003$). Between groups, Control Group Givers said more in the Dual-Feedback Modality (731.6) than Motivated Givers said in the No Feedback condition (351.6) (Bonferroni t -test, $p < .003$, $\alpha < .003$).

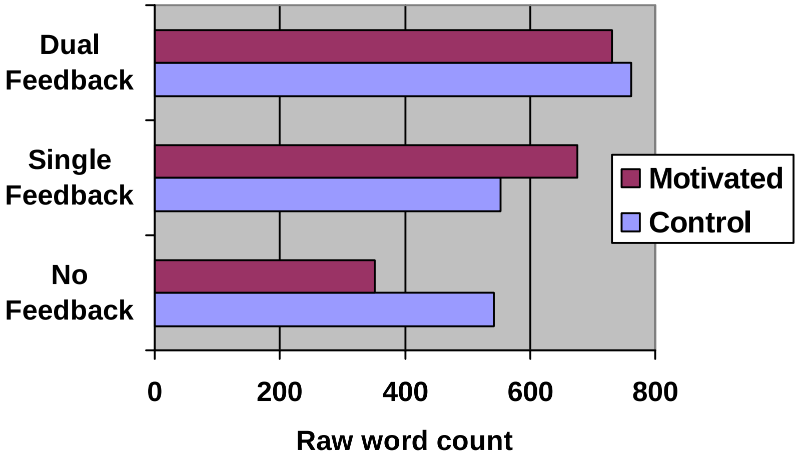


Figure 35. Mean raw word count for Motivated and Control Group Givers

The Single-Feedback Modality (Motivated: 685.4; Control: 547.9) did not differ significantly in terms of raw word count from either the No Feedback (Motivated: 352.4; Control: 530.2) or the Dual-Feedback Modality (Motivated: 735.1; Control: 763.6) for either group (Bonferroni t -test, $p < .003$, $\alpha > .003$). There was also no significant overall effect of Motivation on word count (Between-Groups: F_1 (1,16) = 0.97, $p = .760$) . Thus, it appears that Givers from both groups said more in the Dual-Feedback modality when they could interact with the Follower than they did in the No Feedback modality without the possibility of interaction.

5.8.2 Speech Rate

Since Chapter 4 revealed a significant result for speech rate, I will proceed to test speech rate for Experiment 3. Speech rate across experimental conditions was also subjected to repeated measures ANOVA. Again speech rate, equals the total Giver words per map by the total amount of time the Giver spent speaking for that map (ie. the sum of all conversational moves less the summed durations of silent and filled pause time). There were no significant differences between either Groups (Motivated vs. Control) (Group: F_1 (1,16) = .63, $p < .86$) or among Modalities (No Feedback, Visual-Only, Dual-Feedback Modality) (Feedback Modality: F_1 (2,32) = .234, $p < .95$) with respect to speech rate.

5.8.3 Transaction Rate

Chapters 3 and 4 showed that Normal Transactions are not affected by Feedback Modality. Instead, Normal Transactions were more common in Untimed conditions since Givers tended to say more when they had the time to do so. Retrieval Transactions were more common in the Dual-Feedback Modality than in the No Feedback modality (see Section 4.8.3). In this section, I will investigate whether Motivated Givers made more effort to retrieve lost Followers than Givers in the Control Group. An analysis of this sort is valuable because it can tell us whether the paradigm of offering some Givers more incentive to perform well actually worked.

Table 21. Rate of Normal Transactions for the Control and Motivated Groups. The difference between Groups is not significant.

Feedback Modality	None	Single	Dual
Control Group	16.78	16.67	17.33
Motivated Group	16.00	16.67	19.11

The rate of Normal Transactions per trial was submitted to an ANOVA for Group (2) x

Feedback Modality (3). There were no significant results for the rate of Normal Transactions (Table 21).

For Retrieval Transactions, on the other hand a Between-subjects Group effect was observed: Motivated Givers (4.74) were more likely to retrieve a lost Follower than Control Group Givers (3.04) (Between-Subjects, Group: F_1 (1,16) = 6.96, p < .02). As usual, Retrievals were more common in both the Dual-Feedback Modality (7.56) and the Single-Feedback Modality (4.06) than in the No Feedback Modality (.056) (Feedback Modality: F_1 (2,32) = 61.52, p < .001).

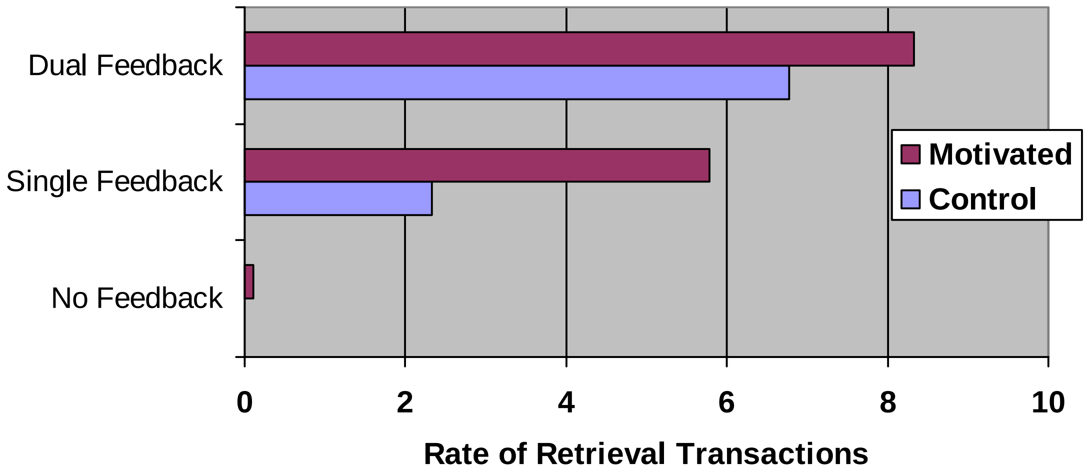


Figure 36. Rate of Retrieval Transactions for the Control and Motivated Group with respect to Feedback Modality.

Figure 36 shows the distribution of Retrieval transactions by Groups. From this graph, there is an observable difference between how often Motivated Givers retrieved compared to Control Group Givers in the Single-Feedback modality. In fact, there was a nearly significant interaction for Feedback Modality x Group: both Motivated and Control Givers made more Retrievals in the Dual-Feedback modality (Motivated: 8.33; Control: 6.78) than they did in the No Feedback modality (Motivated: .11; Control: .00) (F_1(2,32) = 3.05, p = .06; Bonferroni, t-tests, p < .003; α < .003). Post-hoc tests did not show a significant difference between the rate at which Motivated Givers made Retrievals in the Single-Feedback modality (5.78) and the rate at which Control Givers made Retrievals in the same modality (2.33) (Bonferroni, t = 2.05, p = .075, α < .003).

Overall, the results for Transaction rate in Experiment 3 support the general trend for Experiments 1 and 2. Normal Transactions are not affected by manipulations of Feedback-Modality whereas Retrieval Transactions are more common in the Dual-Feedback Modality. The result of that extra £5 was that Motivated Givers retrieved their presumably lost Followers more often than Control Group Givers did. This result would suggest that Motivated Givers are able to retrieve lost Followers because they spend more time gazing at the Follower. Since Motivated Givers did retrieve their Followers more often than Control Group Givers, we have reason to believe that the motivation manipulation worked and can therefore go on to examine the effects of motivation on disfluency. We turn to Section 5.9 to determine whether additional motivation affected general gazing patterns.

5.9 Gaze

One goal set out at the beginning of this chapter is to determine whether additional motivation actually enhanced or altered participant performance. For this experiment, we ask whether Motivated Givers generally gaze more often at the Follower's feedback than the Control Group Givers did. I attempted to answer this question using 276 'feedback episodes' from Dual-Feedback Modality trials from all 18 subjects (155 episodes for Motivated Givers, 121 episodes for Control Givers). Episodes were defined as beginning when the Giver mentions a new landmark and ending just before he introduces the next landmark on the route. If the Giver gazes at the Follower's feedback square during the episode, the entire episode is labelled 'looked at'. If the Giver's gaze fails to overlap the Follower's feedback square, the episode is labelled 'Not looked at'. The dependent variable, general gaze rate, is calculated by dividing the number of looked at episodes by the total number of feedback episodes.

A Mixed Between and Within by-subjects ANOVA (Motivation (2) x Verbal Feedback (2) x Visual Feedback (2)) where the dependent variable consisted of only the 'looked at' episodes revealed that Givers in the Motivated Group (.914) gazed more at their Followers than Givers in the Control Group (.554) (Between-Subjects Group: F_1 (1,16) = 42.44, $p < .001$). Once again, By-Material ANOVAs were not performed because it is impossible to generalize over the linguistic material surrounding different landmarks and maps under all conditions. These results show that Motivated Group Givers attend more closely to their Followers' feedback overall. An analysis of Visual Feedback x Verbal Feedback can inform us whether Motivated Givers met the predictions of the Strategic-Modelling View by looking at Followers more often on a wrong landmark.

As in Experiment 2, Givers tended to gaze more at Correct Visual Feedback (.798) than at Wrong Visual Feedback (.670) (Visual Feedback: F_1 (1,16) = 6.35, $p < .05$). There was a significant interaction between Visual and Verbal Feedback effects (F_1 (1,16) = 8.16, $p < .02$), but internal comparisons were not significant (Visual Feedback: Correct: .758; Wrong: .838; Verbal Feedback Positive: .724; Negative: .616) and there was no interaction with groups.

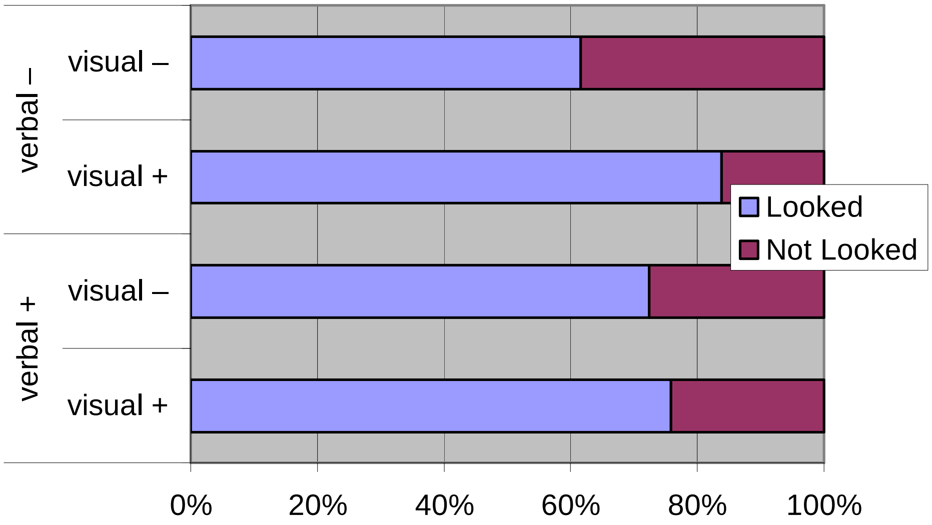


Figure 37. Proportion of feedback episodes attracting speaker gaze to feedback square: Effects of combinations of visual and verbal feedback in Dual-Feedback conditions. Post-hoc tests for this interaction were not significant.

Overall, it seems that when Givers are offered additional incentive to perform, they gaze more often at the Follower's location. Givers in both groups tended to gaze at the Follower when it was easy for them to do so, namely when the visual feedback was Correct. The predictions of the Cognitive Burden theory suggest that Givers may not be able to afford the effort to gaze at the Follower during difficulty, i.e. when the Follower is lost. The results observed here for General Gaze support results observed in Chapter 4, Section 4.9 and the predictions of the Cognitive Burden view.

5.10 Disfluency Rate

In total, there were 342 disfluencies for the Motivated Group and 787 disfluencies for the Control Group across all speakers and all conditions. The dependent variable disfluency rate was submitted to a by-subjects Mixed ANOVA for Group (2: Control vs. Motivated) and Feedback Modality (3: No Feedback, Visual-only and Dual-Feedback Modality). Overall, Givers were more disfluent in the Dual-Feedback Modality condition (.027) than they were without any feedback at all (.018) or in the Single-Feedback Modality (.021) (Feedback Modality: F_1 (2,32) = 8.66, p =

.001). These results are depicted in Figure 38. There were neither significant Groups effects nor any significant interactions.

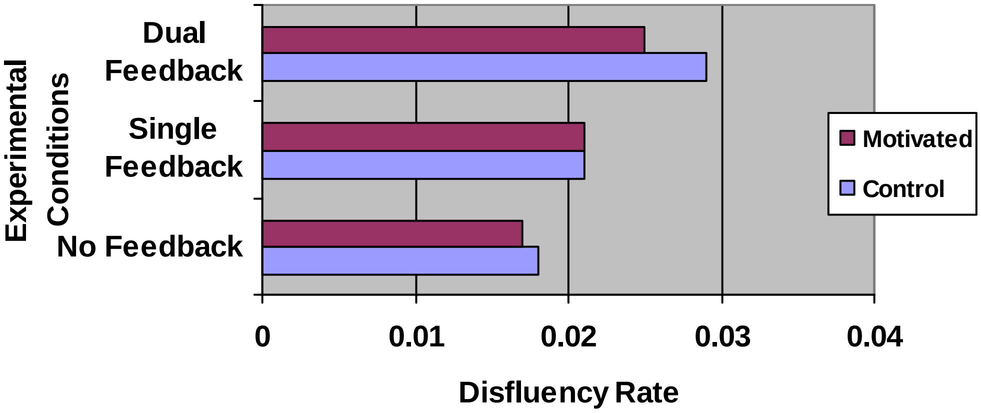


Figure 38. Overall Disfluency Rate for Control and Motivated Groups across the No Feedback, Visual-Only and Dual-Modalities

The pattern for disfluency rates suggests the added Motivation is not critical to disfluency rate. Again, as for raw word counts, the Visual-only condition did not differ statistically from the Dual-Feedback Modality condition for either group. We can therefore examine only the Dual-Feedback Modality conditions since the presence of an additional medium does not seem to affect disfluency rate in any way.

5.10.1 Disfluency Types

Previous disfluency research has found that individual types of disfluencies behave in systematic ways (Fox Tree, 1995; Levelt, 1983; Lickley, 2001). Clark & Wasow (1998) and Fox Tree (1997) predict that repetitions are linked to strategic signalling from speaker to listener. In terms of audience design, one would therefore predict higher repetition rates in more interactive circumstances. For this reason, disfluency rates of individual types were calculated and submitted to independent analyses. In total, there were 106 repetitions, 80 substitutions, 39 insertions and 117 deletions for the Motivated Group. The Control Group made 139 repetitions, 280 substitutions, 107 insertions and 74 deletions in total.

As found in Experiment 2, there was a significant Feedback Modality main effect.

Repetitions were more common in the Dual-Feedback Modality (.009) than in the No Feedback Modality (.005) (Feedback Modality: F_1 (2,32) = 3.68, p < .05). The Single-Feedback Modality (.007) did not differ significantly from either the Dual-Feedback modality or the No Feedback modality. There was no significant Group effect or any significant interactions.

In concordance with Experiment 1 and Experiment 2, deletion rate exhibits a significant effect of feedback (No Feedback .003, Visual-only .006, Dual .007; F_1 (2,32) = 5.22, p < .02). Givers made more deletions in the Dual-Feedback Modality (.007) than the No Feedback condition (.003). Motivation also seemed to affect deletion rate significantly: Motivated Givers made more deletions per word (.007) than Control Group Givers (.004) (Between Subjects: F_1 (1,16) = 8.76, p < .01). These results are shown in Figure 39.

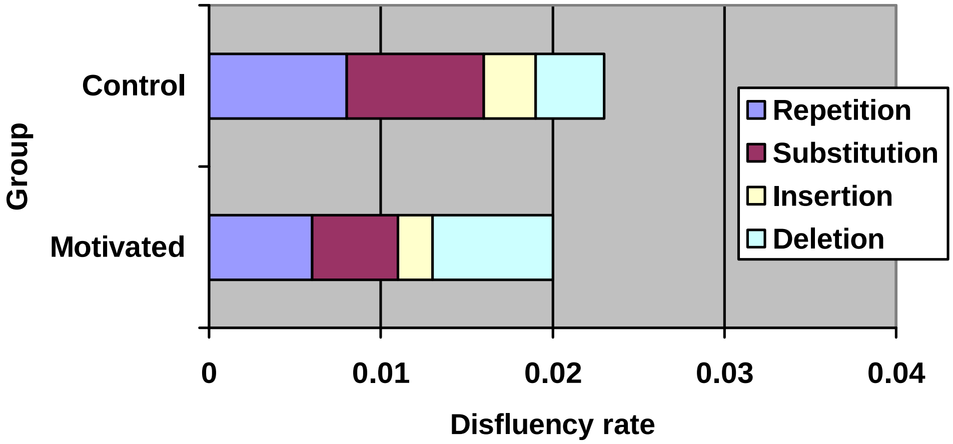


Figure 39. Disfluency Rate by Type for Motivated and Control Givers

As depicted in Figure 39, the Control Group (.008) produced more substitution disfluencies than the Motivated participants did (.005) (Between Subjects: F_1(1,16) = 6.98, p < .02). There were no other significant main effects or interactions. An ANOVA of rates of insertions failed to reveal any significant results.

Thus, deletions and substitutions both exhibit independent effects of motivation, albeit in different directions. The additional motivation increased deletion rate while it reduced substitution rate. The prediction that repetition rate would rise in interactive circumstances was not met.

5.10.2 Filled Pause Rate

Filled pause rate per word was submitted to an ANOVA for Feedback (3) x Group (2). There was a significant interaction between Feedback-Modality x Group, but post-hoc comparisons were not significant ($F_1(2,32) = 6.03$, $p < .01$; Bonferroni, *t*-tests, $p < .003$, $\alpha > .003$). Means (Motivated Group: No Feedback: .0233; Single-Feedback Modality: .015; Dual-Feedback Modality: .013; Control Group: No Feedback: .018; Single-Feedback Modality: .021; Dual-Feedback Modality: .029) are shown in Figure 40.

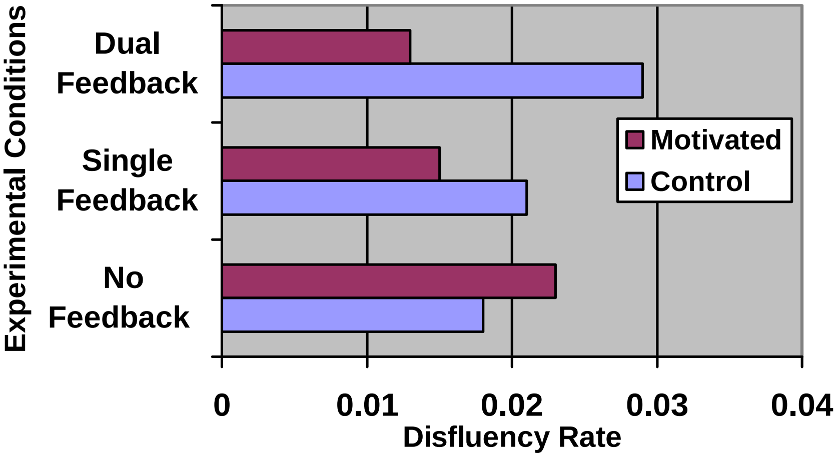


Figure 40. Filled Pause Rate for Control and Motivated Groups

Internal comparisons were not significant between the Motivated and Control Groups but notice from Figure 40 that the Motivated Givers made the most Filled Pauses in the No Feedback condition when they did not receive feedback from the Follower. Clark and Fox Tree (2002) suggest that filled pauses are used as strategic signals where *um* signifies a longer delay than *uh*. The Strategic-Modelling view predicts that Givers should make signals more often in the presence of a listener. The results observed here for the Motivated Group go against these predictions because in fact, Motivated Givers showed a tendency to make more filled pauses in the No Feedback condition than in either the Single-Feedback or Dual-Feedback modalities when they could interact more frequently with the Follower.

5.11 Disfluency Rate by Transaction Types

Chapter 4 investigated disfluency rate per word by Transaction types in order to determine whether disfluency is associated with a particular dialogue goal. Since the Control Group for the current Experiment consists of subjects from Experiment 2B, we can compute the same results in order to determine whether Motivation is a factor. Disfluency rate per Transaction (i.e. the number of disfluencies in Normal Transactions per the number of words in Normal Transactions) was submitted to an ANOVA for Transaction type (2) x Feedback-Modality (2) x Group (2). Retrieval Transactions (.033) were more prone to disfluency than Normal Transactions (.023) (Transaction Type: $F_1(1,16) = 9.58$, $p < .01$). Givers made more disfluent transactions in the Dual-Feedback Modality (.032) than in the Single-Feedback Modality (.024) (Feedback Modality: $F_1(1,16) = 4.87$, $p < .05$). There were no significant Group effects or significant interactions.

Disfluencies of different types were also submitted to independent ANOVAs. Repetition rate was higher in Retrieval Transactions (.012) than Normal Transactions (.007) (Transaction Type: $F_1(1,16) = 6.15$, $p < .05$. Control Group Givers (.012) made more substitutions than Motivated Givers (.004) (Group: $F_1(1,16) = 10.76$, $p < .01$). There was also a significant interaction between Transaction x Group, but internal comparisons were not significant (Transaction x Group: $F_1(1,16) = 6.97$, $p < .02$). ANOVAs for Insertion and Deletion rate failed to produce any significant results.

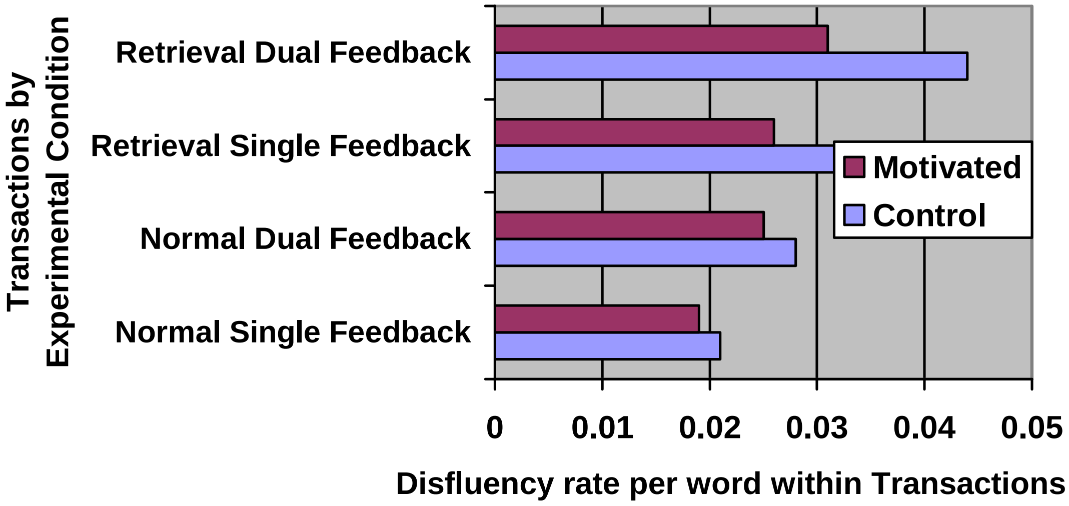


Figure 41. General Disfluency Rate in with respect to Transaction Type, Feedback Modality and Group

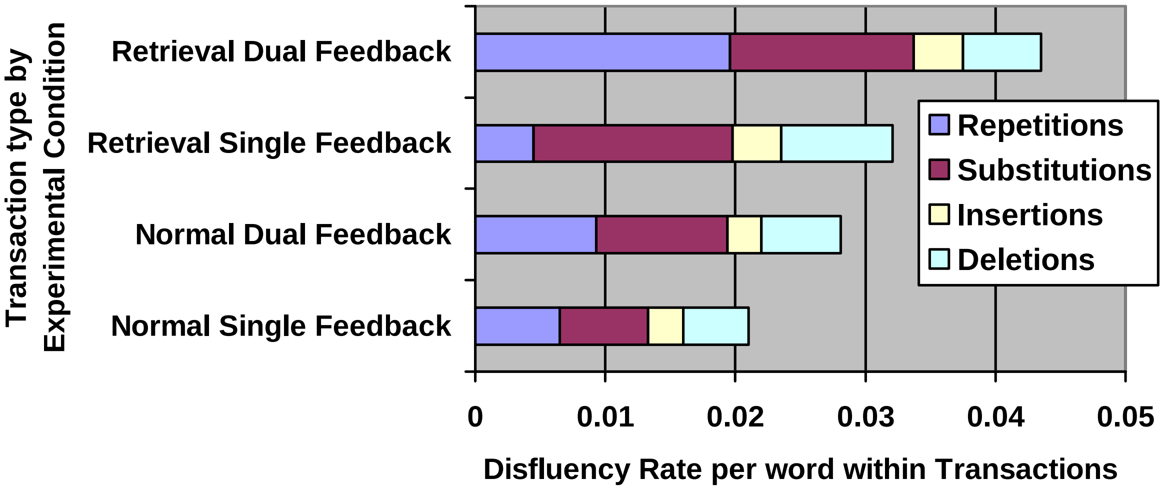


Figure 42. Rates of Disfluency Types by Transaction Type and Feedback Modality for the Control Group

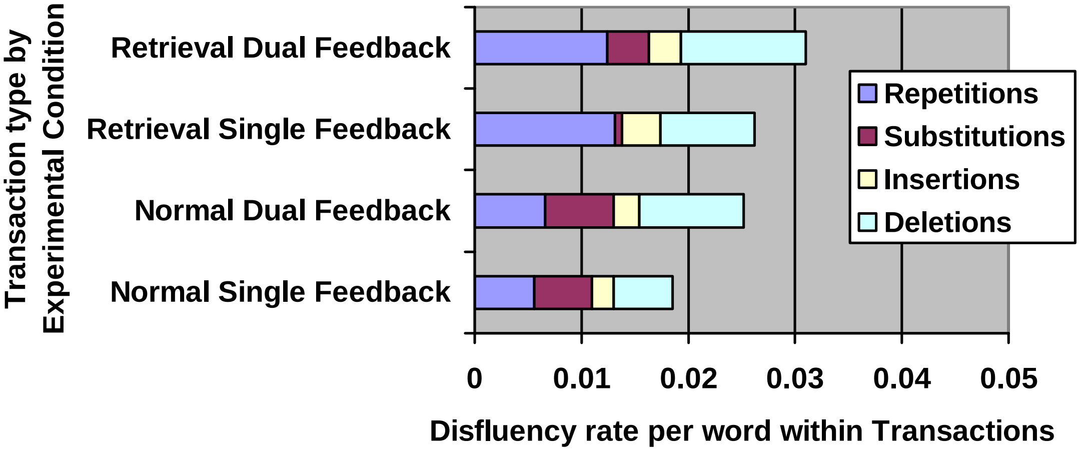


Figure 43. Rates of Disfluency Types by Transaction Type and Feedback Modality for the Motivated Group

Retrieval Transactions were more prone to disfluency than Normal Transactions in Experiment 3, as in Experiment 2. Motivation did not affect disfluent transaction rate. The only Group effect was found for substitutions rate: Control Group Givers made more substitutions than Motivated Givers. There was no effect of Motivation for deletion rate in transactions, although one was found in Section 5.10.1. Contrary to the predictions of the Strategic-Modelling view, enhanced Motivation did not increase repetition rate, suggesting that speakers do not use repetitions as signals. In the next section we turn to investigate speaker's gaze behaviour to see whether disfluency is associated with gazing for Motivated Givers.

5.12 Disfluency and Gaze with a Feedback Episode

Section 5.9 above showed that motivation increased the frequency with which Givers gazed at visual feedback. Does the additional effort of gazing at the Follower have an associated cost in terms of fluency? In order to answer this question, we looked at the number of disfluencies per feedback opportunity in 18 Dual-Feedback Modality trials (9 from Motivated Givers, 9 from Control Givers). Feedback episodes were defined as for the General Gaze analysis. An episode was 'looked at' when the Giver gazed at the Follower's feedback square. The episode was deemed 'disfluent' if the Giver was disfluent while talking about the current landmark within the episode. Disfluency rate per feedback opportunity was then calculated by dividing the number of disfluent episodes by the total episodes of that type to give the dependent variable, disfluency per opportunity. To answer the question of whether there is a cost associated with additional attention to the Follower's feedback, the dependent variable, proportion of 'looked at' episodes per total episodes, was then submitted to a Mixed Within and Between by-subjects ANOVA where Visual Feedback (2: Correct vs. Wrong), Verbal Feedback (2: Positive vs. Negative) and Group (2: Motivated vs. Control) were the independent factors.

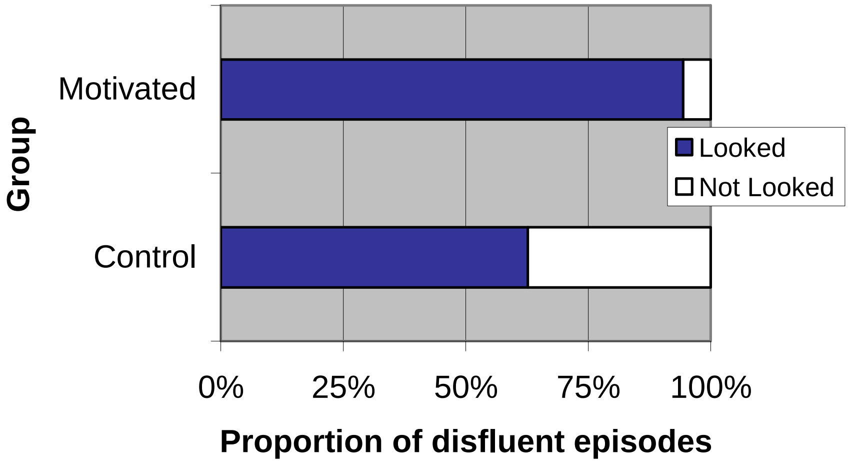


Figure 44. Proportion of disfluent feedback episodes for both the Motivated and the Control Groups per total episode opportunity with respect to whether the Giver was looking or not looking at the Follower

Motivated Givers were more disfluent when they looked (.487) at the Follower than Control Group Givers (.223) (Between Subjects Group: $F_1(1,16) = 9.54$, $p = .007$). There were no other significant interactions with Group.

As shown in Figure 45, All Givers were more disfluent when they looked at wrong visual feedback (.444) than when they gazed at 'correct visual feedback (.226) (Within Subjects Visual Feedback: $F_1(1,16) = 9.55$, $p = .007$). There was a near significant trend showing that Givers were also more disfluent following negative verbal feedback (.433) than following positive verbal feedback (.277) (Within Subjects Verbal Feedback: $F_1(1,16) = 3.99$, $p = .063$).

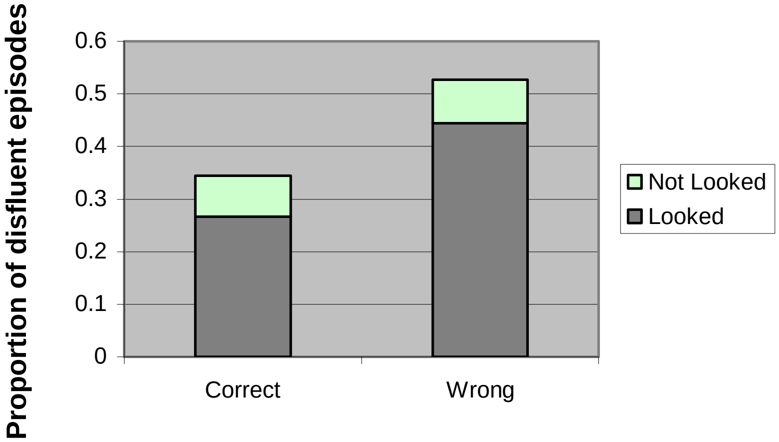


Figure 45. Proportion of disfluent episodes with respect to the Visual Feedback (correct or wrong) and Giver Attention (Looked vs. Not Looked)

Results presented in Section 5.9 showed that Motivated Givers gazed more often at their Followers than Control Givers. In addition to gazing more often, this section has shown that Motivated Givers are also more disfluent per episode than Control Givers when they gaze at the Follower. This suggests that Motivated Givers are either a) more disfluent because they are committed to the Follower and are using disfluency as a collateral signal or b) more disfluent because they are under stress when the Follower indicates that she is lost with wrong visual feedback. Were Motivated Givers attempting to use their disfluencies as signals to their Followers? We turn to the next section to answer this question.

5.13 The Function of Structural Disfluency Types

The results presented thus far have shown that Motivated Givers pay more attention to the

Follower's feedback and make more deletions than Control Group Givers. These results answer the questions of whether Givers attend more if they are motivated to do so and of how this extra caution manifests itself with respect to Giver fluency. The remaining question pertains to what the function of disfluency was. What sort of a function did the Giver fulfil with the disfluency? Was the disfluency rooted in an external or internal source? Chapter 4 investigated the function of repetitions and deletions produced in Experiment 2. Since Experiment 3 is modelled on Experiment 2 and since both repetitions and deletions increased in interactive circumstances, I will investigate the function of both repetitions and deletions in this Section. The Strategic-Modelling view would predict that Motivated Givers would be especially likely to signal to the listener through repetitions. The Cognitive Burden view predicts that Givers will assist their Followers, if they notice that they are lost but only if it is easy for the Giver to do so. Therefore, I examined 95 deletions and 85 repetitions occurring in two Feedback modalities of the Motivated Givers and 100 repetitions and 54 deletions in the two Feedback Modalities of the Control Group Givers. This data is portrayed in Table 22.

Table 22. Distribution of Repetitions and Deletions by Feedback Modality and Group

	REPETITIONS		DELETIONS	
Feedback Modality	Single	Dual	Single	Dual
Control Group	31	69	21	33
Motivated Group	42	43	32	63

Once again, the rate of planning disfluencies was submitted to an ANOVA for Disfluency Type (2) x Feedback-Modality (2) x Group (2). As found previously for Experiment 2, speaker-planning deletions (.004) were more common than planning repetitions (.002) (Disfluency Type: $F_1(1,16) = 6.14$, $p < .05$). A Group effect revealed that Motivated Givers (.004) made more disfluencies for planning reasons than Control Givers (.002) (Between-subjects, Group: $F_1(1,16) = 8.92$, $p < .01$). A near-significant interaction between Disfluency Type and Feedback Modality showed that Givers made more planning deletions occur in the Dual-Feedback Modality (.005) than they made planning deletions in the Single-Feedback Modality (.002), planning repetitions in the Dual-Feedback Modality (.002) or in the Single-Feedback Modality (.002) ($F_1(1,16) = 4.11$, $p = .06$). These results were not significant in post-hoc tests.

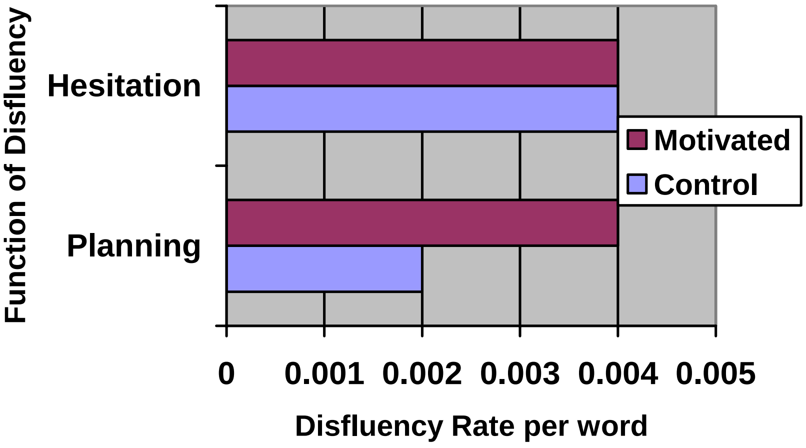


Figure 46. Distribution of repetitions and deletions by Group and Function in Experiment 3

The rate of hesitation disfluencies per word was also submitted to an ANOVA for Disfluency Type (2) x Feedback Modality (2) x Group (2). In line with the results from Experiment 2, hesitation repetitions were more common per word (.005) than hesitation deletions (.003) (F_1(1,16) = 6.67, p = .02). Additional motivation to complete the task did not affect the rate of hesitation disfluencies: there was no significant Group effect (Motivated: .004; Control: .004).

As found previously for Experiment 2, the results for Experiment 3 showed that Deletions were associated with planning reasons, whereas repetitions were associated with hesitation functions. The only Group effect was found for planning disfluencies suggesting that Motivated speakers were more prone to abandoning their own utterance to accommodate the Follower. Thus, it seems that some structural types of disfluencies can occur for planning reasons without necessarily being strategic signals in line with the Strategic-Modelling View predictions. The analysis of hesitation disfluencies showed, however, that Motivated Givers made just as many hesitation disfluencies as Control Givers did suggesting that both types of Givers made disfluencies because of a difficulty they encountered. A hesitation disfluency by definition does not necessarily occur 'for the listener', suggesting that disfluencies could fulfill two cognitive functions, a planning or hesitation function.

5.14 Discussion

In an attempt to discover when and where disfluency occurs, I have investigated different dialogue situations in the form of disfluency rate within Transactions and speaker gaze behaviour during dialogue. In line with Chapter 4, Retrieval transactions were once again more prone to disfluency. Additional incentive did not affect disfluency rate or transaction type. The fact that Retrieval transactions are more prone to disfluency suggests that the difficulty of retrieving a Follower is a source of disfluency. Motivation did seem to affect general gaze patterns, however, because Motivated Givers looked more often at the Follower feedback than Control Givers did. Thus, Motivated Givers attend to the Follower's feedback more closely, and were more disfluent per opportunity than Control Givers, suggesting that either disfluency is indicative of commitment, according to the Strategic-Modelling View, or difficulty, according to the Cognitive Burden View.

An investigation of why disfluency occurs has necessitated a comparison of a structural classification system with a cognitive classification system. As found in Experiment 2, the results suggest that the same structurally classified disfluencies appear to fulfil different cognitive functions for the speaker in different dialogue situations. Accordingly, any classification system should consider both the function and structure of disfluency. In this thesis, I have developed a classification system for discovering the function of disfluency, which I believe could assist future research.

In order to answer the questions of why disfluency occurs, the current chapter has investigated the effect that additional motivation has on a speaker during dialogue. Incentive works in some cases but not in all of them. A Motivated Giver looks at the Follower more often, retrieves lost Followers more often, abandons utterances (i.e. makes a deletion) more often, substitutes more often and is more disfluent per opportunity when looking for the Follower's location than a Control Giver. Motivated Givers were no more disfluent overall, however, than Control Group Givers when measured in terms of disfluency rate per words. Thus, overall we can conclude from these results that given additional incentive to perform well, a motivated participant will be more willing to perform difficult tasks that other subjects (e.g. controls) were not willing to perform.

As Section 5.13 shows, disfluency types seem to fulfil different behavioural functions for the speaker, intentional or otherwise. Contrary to at least some of the predictions of the Strategic-Modelling, repetitions were not associated with planning goals as frequently as deletions. As Brennan (2004) observed, attentive speakers abandoned moves when they observed from the movement of visual feedback that the current utterance was no longer relevant to the listener's new location. If any structural type of disfluency fulfils the function of a collateral signal in the sense suggested by Clark (2002), it would seem from the results presented in this thesis to be a planning deletion, or abandonment. This does not, however, mean that all disfluencies of the same structural type fulfil this role, nor does it mean that all disfluencies of different structural type fulfil this role. Repetitions are less clearly collateral signals because they tend to be made for hesitation reasons rather than planning reasons. It seems, therefore, that Givers were behaving according to the predictions of joint action. The acts of retrieving and abandonment would suggest that they took only partial responsibility for their Follower. Speakers would be expected to look more often at the listener and only offer additional assistance when they are cognitively capable of doing so and when they realise the success of the entire collaborative effort is at risk if they do not.

CHAPTER 6 – CONCLUSION

This thesis set out to address the questions of why disfluency occurs in collaborative dialogue. In order to answer these questions, I reviewed two theories from the psycholinguistic literature which attach different functions to disfluency and therefore differ in their explanation for why disfluency occurs. The Cognitive Burden View suggests that disfluencies are simply errors of an overburdened language production system and that they are not intentionally controlled. The Strategic-Modelling View, on the other hand, suggests that disfluencies occur as strategic signals from speaker to listener to signal that the speaker is committed to the utterance but is currently experiencing difficulties. Since each of these theories attach specific functions to structural types of disfluencies, I tested the predictions of each theory by observing speaker behaviour in a multi-modal setting while the speaker was disfluent. The results of this analysis showed that disfluencies described by strict structural classifications don't always perform the same functions in dialogue.

Experiment 1 investigated a baseline condition of visual feedback in order to establish a viable paradigm. Speakers retrieved the visual feedback square after noticing that it had gone astray. This is strong evidence to suggest that speakers believed the visual feedback was genuine and therefore, that the experimental paradigm worked. Further effects of trial length in words, speech rate and disfluency rate were in line with previous research which suggested that speakers say more when given more time and wordier trials are more prone to disfluency (Bard et al., 2001; Oviatt, 1995). Next, I analysed whether the experimental manipulations on feedback and time-pressure affected speaker gaze behaviour and disfluency rate. Results suggested that speakers gazed more often at the Follower when she was hovering over a correct landmark, and less when she hovered over a wrong landmark. This suggests that the speaker avoids the difficult information because of the cognitive load required on his part. In terms of disfluency rate and gaze, results showed that speakers have a high disfluency rate when they must re-orient a lost Follower, that is a Follower hovering over a wrong landmark. Speakers, according to the Strategic-Modelling view, should be most attentive when the listener is lost, if the speaker is behaving according to the principles of Optimal Design which suggest that speaker design their utterances 'for the listener'. Likewise, the evidence only partially supports the Cognitive Burden view, which predicted that speakers would be more disfluent during difficult periods, i.e. when faced with a 'lost' Follower. Finally, deletions were classified as fulfilling hesitation or planning functions. Givers were found to make more planning deletions after noticing the movement of the visual feedback square. This result supports the Middle Ground view of disfluencies which states that speakers and listeners have a joint responsibility in collaboration. When the speaker abandoned an utterance, the speaker was taking responsibility and attempting to re-align the listener.

Experiment 2 paired visual feedback with verbal feedback in order to determine whether speaker behaviour changed as a result of feedback type. For Givers who received Visual-only feedback in the Single-Feedback Modality, there was a sharp increase in trial length in the Dual-Modality once verbal feedback had been added. This sharp increase did not occur for Givers who received Verbal-only feedback in the Single-Feedback Modality suggesting that Givers relied more on verbal feedback than on Visual Feedback. Similarly, Givers retrieved more often when they had verbal feedback than in the Visual-only feedback condition. In terms of speaker gaze behaviour, Givers once again avoided gazing at their Follower when she clearly indicated that she was lost with both visual and verbal cues. This supports the Cognitive Burden view that Givers will avoid gazing at their Followers when it is difficult to do so. Finally in terms of speaker disfluency behaviour, Givers were most disfluent in the Dual-Feedback Modality in both groups, supporting the claims of both the Strategic-Modelling view and the Cognitive Burden view that speakers will be more disfluent in interactive circumstances. Once again, deletion rate increased significantly in interactive circumstances. Repetition rate also showed a significant effect of Feedback in Experiment 2, thus partially supporting the claims of Clark and Wasow (1998) that speakers use repetitions as strategic signals. A further analysis of disfluency and gaze behaviour showed that Givers were more disfluent after they had gazed at discordant feedback, i.e. correct visual, negative verbal feedback or wrong visual, positive verbal feedback, compared to concordant feedback. Since the difficulty level increases in discordant feedback, this result supports the Cognitive Burden view. In order to pinpoint the behavioural differences between deletions and repetitions, an analysis of disfluency function was also conducted. Deletions tended to occur for planning reasons whereas repetitions occurred more often for hesitation reasons, although there were some differences between Visual and Verbal Groups. The finding that repetitions occur most often for hesitation reasons is important when evaluating the Strategic-Modelling view since it predicts that repetitions are strategic signals to a listener. The findings in this thesis do not support this prediction or at least suggest that the intentionality of a repetition disfluency is not immediately apparent.

Finally, Experiment 3 tested the effect of additional incentive or motivation of the speaker to perform well. Compared to controls, Motivation was found to affect speaker attention (i.e. gaze patterns), retrieval transaction rate, deletion rate and the speaker's disfluency per opportunity.

Motivated Givers attended to the Follower's feedback more closely and retrieved lost Followers more often than Control Group Givers. Motivated Givers were also found to abandon utterances more often than Control Givers. Thus, from these results, we can conclude that motivated participants are more willing to perform difficult tasks in dialogue. This observation should be considered in future studies.

Taking all experiments into account, there seems to be mixed support for both the Cognitive Burden and Strategic-Modelling views. Speakers in all experiments were more disfluent in interactive circumstances, supporting the predictions of the Strategic-Modelling view. Speakers in Experiments 1 and 2, however, tended to avoid difficult tasks like gazing at a lost Follower and were more disfluent during complicated feedback (i.e. discordant feedback episodes), supporting the Cognitive Burden view. When structural disfluency types like repetitions and deletions were paired with functions, repetitions fulfilled a hesitation function and not a strategic, planning role as Clark has elsewhere claimed (e.g. Clark & Wasow (1998) or Fox Tree & Clark (1997). Deletions, on the other hand, tended to fulfil a planning function when the speaker observes that the feedback has found the correct landmark or has gone astray. Since it would be redundant for the speaker to continue saying the current utterance, the speaker abandons this utterance and provides more pertinent information instead. Brennan (2004), Clark (2002) and Clark and Krych (2004) have observed similar speaker behaviour. Clark and Krych (2004) have suggested that such behaviour suggests that speakers are opportunistic. This could well be the case for a subset of deletions or even of all disfluencies. As I have shown in this thesis, not all structural disfluencies of the same or different type necessarily fulfil the same function in dialogue. Regarding this matter, we can then conclude in line with Schober and Brennan (2003) that speakers may adapt in some circumstances and avoid adaptation in other circumstances.

This thesis has largely remained agnostic about the intentionality of disfluency. As described in Section 2.3.4, Chapter 2, in order for something to be considered intentional by a speaker, there must be mutual knowledge that the speaker intended to make the utterance and that the speaker intended for the listener to be aware of this intention. Determining whether a speaker had these intentions in mind when making a disfluency is an extremely difficult task since modern science is not yet capable of truly determining what a speaker had in mind even with invasive techniques. As Brennan and Schober (2003) suggest, an experimenter using an online test can better access speaker intention than an experimenter conducting a corpus analysis. Eye-tracking technology allows a researcher to see what the speaker looked at and previous research has found that speakers tend to talk about what they looked at (Brown-Schmidt & Tanenhaus, In Press; Griffin, 2005; Griffin & Bock, 2000; Tanenhaus et al., 2000). While eye-tracking has provided some useful insights in this thesis about the association between disfluency and gaze, one still has to be careful about deciding what the speaker's intentions were when they were disfluent. Therefore, I believe that the results found in this thesis elucidate potential circumstances under which a speaker could utter something which sounds like a disfluency (i.e. structurally coded deletions which fulfil a planning function) but is by no means an intentional signal. Instead, the speaker has simply stopped speaking because an external stimulus provided new and relevant information to the task and the speaker. In conjunction with Schober and Brennan (2003), this would seem to indicate that speakers use deletion-like utterances in some situations to adapt to a listener some of the time (i.e. in the case of planning deletions) but also make disfluencies out of genuine difficulty in other situations (i.e. in the case of hesitation deletions or hesitation repetitions).

The feedback manipulation used in this thesis is perhaps the most contestable part of the experiments. It is possible that Strategic-Modelling would predict no change in repetition or filled pause rate if the signals that repetitions and filled pauses send are so highly specialised that they could not occur in a simulation of eye-gaze. Experiment 1 showed no significant difference between repetition rates in a feedback and no feedback trial. Although theoretically still possible, this suggests that visual feedback alone was not enough to create the situations for such specific signals. Of course, it could also be the case that repetitions were not being used as signals regardless of the feedback manipulation. Experiment 2, which incorporated verbal feedback as well as visual, revealed that Visual Group subjects increased their rates of repetitions when provided with verbal feedback in the Dual-Feedback modality. Repetition rates in Experiment 3 were significant in an interaction with Feedback and Group. Thus, we can conclude for the present that it is possible that verbal feedback alone does not create the situation in which it is possible to discern between the two possible predictions of the Strategic-Modelling view. Once verbal feedback is added, however, repetition rates show a feedback effect, ruling out the possibility that the signalling function is too highly specialised for the present paradigm. When the function of repetitions is tested, we see that some behave like obvious planning disfluencies but that on the whole this role was generally left to deletion-like disfluencies as previously discussed.

Likewise, the feedback manipulation presented some ambiguities for the predictions of the Cognitive Burden view. Some proponents of the Cognitive Burden view suggest that dialogue is more difficult than monologue (Horton and Keysar, 1996; Barr and Keysar, 2002) and therefore disfluency rate should increase in the feedback condition compared to a no feedback condition (Bard et al., 2001). Another possibility is that the speaker simply tries harder to be understood when they have feedback, either visual or verbal or both, and is therefore more disfluent in feedback situations. What does it mean to say that the speaker 'tried harder'? In one sense, such a prediction is no different from saying that feedback itself increased the burden on the speaker. On the other hand, it may be possible that the methods used in the present thesis are not capable of discerning between a situation in which feedback alone induces difficulty and a situation in which the speaker tries harder and is therefore more disfluent. For the present, all we can conclude is that disfluency rate increased in the presence of listener feedback and admit possible explanations: feedback alone increases difficulty or perhaps the speaker simply tried harder when presented with feedback.

The MONITOR Project used a simulation of eye-gaze in a multi-modal, interactive setting in order to investigate speaker attention to a listener's feedback. The experimental results show that speaker's believe that this feedback is genuine and so there is no reason to discount the results solely on the basis of the visual feedback or experimental paradigm. Previous research has shown that disfluency types differ according to the task assigned to the speaker (Oviatt, 1995). Still, as discussed in Chapter 3, the results presented in this thesis only pertain to the specific paradigm. As Schober and Brennan (2003) stated, the most beneficial research agenda within collaborative dialogue is to observe under what circumstances speakers do and do not adapt their language usage. This thesis has shown that when presented with a surrogate for gaze, speakers use deletions to adapt to a listener in some circumstances but also make disfluencies out of difficulty in other circumstances. In order to be certain that the results found in this thesis with simulated gaze hold for face-to-face gaze, future research could conduct an experiment using face-to-face dialogue with remote eye-trackers or a video-conferencing task whilst eye-tracking both participants. This technique would still allow the same time-stamped accuracy employed in the current paradigm and one would permit face-to-face gaze between interlocutors. Since participants could perform a different task while holding the video-conference one could further investigate whether disfluency types occur with the same frequency and from the same source as observed in the current thesis.

☙ *Right, Right, Right… that's it finished finished finished finished…*
Oh my God, I'm knackered ❧

BIBLIOGRAPHY

Allwood, J., Nivre, J., & Ahlsén, E. (1990). Speech management: On the non-written life of speech. *Nordic Journal of Linguistics, 13*, 1-45.

Anderson, A. H., Bader, M., Bard, E. G., Doherty, G., Garrod, S., Isard, S., et al. (1991). The hcrc map task corpus. *Language and Speech, 34*, 352-366.

Anderson, A. H., Bard, E. G., Dalzel-Job, S., & Havard, C. (submitted). Look at me when I am listening to you: The impact of feedback in a simulation of visual team working.

Anderson, A. H., Bard, E. G., Sotillo, C., Newlands, A., & Doherty-Sneddon, G. (1997). Limited visual control of the intelligibility of speech in face-to-face dialogue. *Perception & Psychophysics, 59*(4), pp. 580-592.

Argyle, M. (1990). *Bodily communication*.London: Routledge.

Argyle, M., Alkema, F., & Gilmour, R. (1972). The communication of friendly and hostile attitudes by verbal and non-verbal signals. *European Journal of Social Psychology, 1*, 385-402.

Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*.Cambridge: Cambridge University Press.

Argyle, M., & Graham, J. A. (1977). The central europe experiment: Looking at persons and looking at things. *Journal of Environmental Psychology and Nonverbal Behaviour, 1*, 6-16.

Ariel, M. (1990). *Accessing noun-phrase antecedents*.London: Routledge/Croom Helm.

Arnold, J. E., Altmann, B., & Tanenhaus, M. (2003). *Disfluency isn't just um and uh: The role of prosody in the comprehension of disfluency.* Paper presented at the City University of New York Sentence Processing Conference, Cambridge, MA.

Arnold, J. E., Fagnano, M., & Tanenhaus, M. K. (2003). Disfluencies signal thee, um, new information. *Journal of Psycholinguistic Research, 32*, pp. 25-36.

Arnold, J. E., Tanenhaus, M., Altmann, R. J., & Fagnano, M. (2004). The old and thee, uh, new. Disfluency and reference resolution. *Psychological Science, 15*, 578 - 582.

Austin, J. (1962). *How to do things with words*. Oxford: Clarendon Press.

Bailey, K. G. D., & Ferreira, F. (2001). *Do non-word disfluencies affect syntactic parsing?* Paper presented at the Disfluency in Spontaneous Speech (DiSS'01), Edinburgh, Scotland.

Bailey, K. G. D., & Ferreira, F. (2003a). Disfluencies affect the parsing of garden path sentences. *Journal of Memory and Language, 49*, 183 -200.

Bailey, K. G. D., & Ferreira, F. (2003b). *Eye movements and comprehension of disfluent speech.* Paper presented at the City University of New York Sentence Processing Conference, Cambridge, MA.

Bailey, K. G. D., & Ferreira, F. (2005). *Don't swim, hop: The timecourse of disfluency processing.* Paper presented at the City University of New York Sentence Processing Conference, Tucson, AZ.

Bard, E. G., Anderson, A. H., Chen, Y., Nicholson, H., & Havard, C. (2004). *Let's you do that: Enquiries into the cognitive burdens of dialogue.* Paper presented at the Eighth Workshop on the Semantics and Pragmatics of Dialogue (DIALOR'04), Athens, Greece.

Bard, E. G., Anderson, A. H., Flecha-Garcia, M., Kenicer, D., Mullin, J., Nicholson, H., et al. (2003). *Controlling structure and attention in dialogue: The interlocutor vs. The clock.* Paper presented at the Proceedings of ESCOP, Granada, Spain.

Bard, E. G., 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., & Aylett, M. (2000). *Accessibility, duration, and modeling the listener in spoken dialogue.* Paper presented at the Göteborg 2000, fourth workshop on the semantics and pragmatics of dialogue, Göteborg, Sweden: Göteborg University.

Bard, E. G., Aylett, M., & Bull, M. (2000). *More than a stately dance: Dialogue as a reaction time experiment.* Paper presented at the Society for Text and Discourse.

Bard, E. G., & Aylett, M. P. (2000). *Referential form, word duration, and modeling the listener in spoken dialogue.* Paper presented at the The Twenty-third Annual Conference of the Cognitive Science Society, Edinburgh, Scotland.

Bard, E. G., & Aylett, M. P. (2001). *Referential form, word duration, and modeling the listener in spoken dialogue.* Paper presented at the The Twenty-third Annual Conference of the Cognitive Science Society, Edinburgh, Scotland.

Bard, E. G., Lickley, R. J., & Aylett, M. P. (2001). *Is disfluency just difficulty?* Paper presented at the Disfluency in Spontaneous Speech (DiSS '01), Edinburgh, Scotland.

Baron, D., Shriberg, E., & Stolcke, A. (2002). *Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues.* Paper presented at the International Conference on Spoken Language Processing, Denver, Colorado, USA.

Barr, D. J., & Keysar, B. (2002). Anchoring comprehension in linguistic precedents. *Journal of Memory and Language, 46*, 391-418.

Barwise, J., & Cooper, R. (1991). Sample situation theory and its graphical representation. In J. Seligman (Ed.), *Partial and dynamic semantics iii* (pp. 38-74). Centre for Cognitive Science, Edinburgh University: DYANA Deliverable.

Bavelas, J. B., Black, A., Lemery, C. R., & Mullett, J. (1986). I show you how you feel: Motor mimicry as a communicative act. *Journal of Personality and Social Psychology, 50*, 322-329.

Bear, J., Dowding, J., & Shriberg, E. (1992). *Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog.* Paper presented at the The 30th Annual Meeting of the Association for Computational Linguistics, Newark, DE.

Beattie, G., & Shovelton, H. (2003). *Making thought visible: The new psychology of body language.* Paper presented at the ATR Conference on Ubiquitous Experience Media, Kyoto, Japan.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M. L., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. *Journal of the Acoustical Society of America, 113*, 1001-1024.

Blackmer, E. R., & Mitton, J. (1991). Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition, 39*, 173-194.

Bock, K. (1986). Syntactic persistence in language production. *Cognitive Psychology, 18*, 355-387.

Boomer, D. S., & Laver, J. D. M. (1968). Slips of the tongue. *British Journal of Disorders of Communication, 3*, 2-12.

Branigan, H. P., Lickley, R. J., & McKelvie, D. (1999). *Non-linguistic influences on rates of disfluency in spontaneous speech.* Paper presented at the ICPhS, San Francisco.

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

Brennan, S. (2004). How conversation is shaped by visual and spoken evidence. In J. C. Trueswell & M. Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the language-as-product and language-action traditions* (pp. 95-129). Cambridge, MA: MIT Press.

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

Brennan, S., & Lockridge, C. B. (2004). *How visual copresence and joint attention shape speech planning*.Unpublished manuscript.

Brennan, S. E., & Schober, M. F. (2001). How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language, 44*, 274-296.

Brown-Schmidt, S., & Tanenhaus, M. (In Press). Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language*.

Brown, G., Anderson, A. H., Yule, G., & Shillcock, R. (1983). *Teaching talk.* Cambridge: Cambridge University Press.

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

Buxton, W. A. S., & Moran, T. (1990). Europarc's integrated interactive intermedia facility (iiif): Early experience. In S. Gibbs & A. A. Verrijn-Stuart (Eds.), *Multi-user interfaces and applications* (pp. 11-34). Amsterdam: Elsevier.

Chafe, W. (1980). The pear stories. Norwood, New Jersey: Ablex.

Carletta, J. (2005). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics, 22,* 249-254.

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

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

Clark, H. H. (1994). Discourse in production. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 985-1021). San Diego: Academic Press.

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

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

Clark, H. H., & Carlson, T. B. (1982a). Critics' beliefs about hearers' beliefs: A rejoinder. In N. Smith (Ed.), *Mutual knowledge* (pp. 52-59). London: Academic Press.

Clark, H. H., & Carlson, T. B. (1982b). Speech acts and hearers' beliefs. In N. Smith (Ed.), *Mutual knowledge* (pp. 1-37). London: Academic Press.

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

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language, 50*(1), 62-81.

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* (pp. 10-63). Cambridge: Cambridge University Press.

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

Clark, H. H., & Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive Psychology, 37*, 201-242.

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cooper, R. (1992). A working person's guide to situation theory. In S. L. Hansen & F. Sorensen (Eds.), *Topics in semantic interpretation*. Samfundslitteratur, Fredriksberg, Denmark.

Dahan, D., Tanenhaus, M. K., & Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language, 47*, 292-314.

Dahl, D. A., Bates, M., Brown, M., Fisher, W., Hunicke-Smith, K., Pallett, D., et al. (1994). *Expanding the scope of the atis task: The atis-3 corpus.* Paper presented at the the 1994 DARPA Speech and Natural Language Workshop, Princeton, NJ.

Dell, G. S., & Repka, R. J. (1992). Errors in inner speech. In B. J. Baars (Ed.), *Experimental slips and human error: Exploring the architecture of volition*.New York: Plenum Press.

Doherty-Sneddon, G., Anderson, A. H., O'Malley, C., Langton, S. R. H., Garrod, S., & Bruce, V. (1997). Face-to-face interaction and video mediated communication: A comparison of dialogue structure and co-operative task performance. *Journal of Experimental Psychology: Applied, 3*, 105-125.

Duez, D. (1982). Silent and non-silent pauses in three speech styles. *Language and Speech, 25*, 11-28.

Eberhard, K.M., Spivey-Knowlton, M.J., Sedivy, J., Tanenhaus, M.J.. (1995). Eye-movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research, 24,* 409-436.

Eklund, R. (2001). *Prolongations: A dark horse in the disfluency stable.* Paper presented at the DiSS: Disfluencies in Spontaneous Speech, Edinburgh, UK.

Eklund, R. (2004). *Disfluency in swedish human-human and human-machine travel booking dialogues.* PhD Thesis. Department of Computer and Information Science. Linköping University, Sweden.

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

Fox Tree, J., & Clark, H. H. (1997). Pronouncing 'the' as 'thee' to signal problems in speaking. *Cognition, 62*, 151-167.

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.

Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Language, 47*, 27-52.

Gale, C., & Monk, A. F. (2000). Where am I looking? The accuracy of video-mediated gaze awareness. *Perception & Psychophysics, 62*, 586-595.

Glenberg, A. M., Schroeder, J. L., & Robertson, D. A. (1998). Averting the gaze disengages the environment and facilitates remembering. *Memory and Cognition, 26*, 651-658.

Godfrey, J., Holliman, E., & McDaniel, J. (1992). *Switchboard: Telephone speech corpus for research and development.* Paper presented at the the IEEE Conference on Acoustics, Speech and Signal Processing, San Francisco, CA.

Goldman-Eisler, F.E.. (1972). Pauses, Clauses, Sentences. *Language and Speech*. *15*, 103-113.

Goodwin, C. (1981). *Conversational organization: Interaction between speakers and hearers.*New York: Academic Press.

Gregory, M. L., Joshi, A., & Sedivy, J. (2003). *Adjectives and processing effort: So, uh, what are we doing during disfluencies?* Paper presented at the City University of New York Sentence Processing Conference, Cambridge, MA.

Grice, H. P. (1957). Meaning. *Philosophical Review, 66*, 377-388.

Grice, H. P. (1968). Utterer's meaning, sentence meaning and word meaning. *Foundations of Language, 4*, 225-242.

Grice, H. P. (1989). *Studies in the ways of words.*Cambridge, MA: Harvard University Press.

Griffin, Z. (2005). The eyes are right when the mouth is wrong. *Psychological Science, 15*, 814-821.

Griffin, Z. (2004). Why look? Reasons for eye movements related to language production. In J.M. Henderson and F. Ferreira (Eds.), *The integration oflanguage, vision and action: Eye movements and the visual world.* New York: Psychology Press.

Griffin, Z., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science, 11*, 274-279.

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referrring expressions in discourse. *Language, 69*, 274-307.

Hanna, J. E., Tannenhaus, M. K., & Trueswell, J. C. (2004). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language, 49*, 43-61.

Hartsuiker, R. J., & Kolk, H. H. J. (2001). Error monitoring in speech production: A computational test of the perceptual loop theory. *Cognitive Psychology, 42*, 113-157.

Haywood, S. L. (2004). *Optimal design in language production.* University of Edinburgh, Edinburgh.

Haywood, S. L., Pickering, M., & Branigan, H. P. (2005). Do speakers avoid ambiguities during dialogue? *Psychological Science, 16*, 362-366.

Heeman, P. (1997). *Speech repairs, intonational boundaries and discourse markers: Modeling speakers' utterances in spoken dialog.* PhD Thesis.Computer Science Department. University of Rochester.

Hieke, A. E. (1981). A content-processing view of hesitation phenomena. *Language and Speech, 24*, 147-160.

Hindle, D. (1983). *Deterministic parsing of syntactic non-fluencies.* Paper presented at the 21st Annual Meeting of the Association for Computational Linguistics.

Horton, W. S., & Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition, 96*, 127-142.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition, 59*, 91-117.

Howell, P., & Young, K. (1991). The use of prosody in highlighting alterations in repairs from unrestricted speech. *The Quarterly Journal of Experimental Psychology, 43A*, 733-758.

Johnson-Laird, P. (1982a). Mutual ignorance: Comments on clark and carlson's paper. In N. Smith (Ed.), *Mutual knowledge* (pp. 40-45). London: Academic Press.

Johnson-Laird, P. (1982b). Thinking as a skill. *Quarterly Journal of Experimental Psychology, 34A*, 1-29.

Johnson-Laird, P. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*.Cambridge, MA: Harvard University Press.

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica, 26*(1), 22-63.

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, vol. 11*, 32-38.

Kowtko, J., & Price, P. J. (1989). *Data collection and analysis in the air travel planning domain.* Paper presented at the The DARPA Speech and Natural Language Workshop, Cape Cod.

Kraljic, T., & Brennan, S. E. (2005). Prosodic disambiguation of syntactic structure: For the speaker or for the addressee? *Cognitive Psychology, 50*, 194-231.

Krantz, M., George, S. W., & Hursh, K. (1983). Gaze and mutual gaze of pre-school children in conversation. *Journal of Psychology, 113*, 9-15.

Krauss, R.M. & Weinheimer, S. (1964). Changes in the length of reference phrases as a function of social interaction: A preliminary study. *Psychonomic Science*, *1*, 113-114.

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.

Krauss, R.M., Vivekananthan, P.S., & Weinheimer, S. (1968). "Inner Speech" and "External Speech": Characteristics and Communication effectiveness of socially and nonsocially encoded messages. *Journal of Personality and Social Psychology, 9*, 295-300.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*:Sage Publications.

Krippendorff, K. (1987). Association, agreement and equity. *Quality and Quantity, 21*, 109-123.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Ladd, D. R., & Campbell, N. (1991). *Theories of prosodic structure: Evidence from syllabic duration.* Paper presented at the The XIIth International Congress of Phonetic Sciences, Aix-en-Provence, France.

Langton, S. R. H., Watt, R. J., & Bruce, V. (2000). Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences, 4*, 50-58.

Laver, J. D. M. (1980). Monitoring systems in the neurolinguistic control of speech production. In V. A. Fromkin (Ed.), *Errors in linguistic performance: Slips of the tongue, ear, pen and hand.*New York: Academic Press.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition, 14*, 14-104.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.*Cambridge, MA.: The MIT Press.

Levelt, W. J. M., & Cutler, A. (1983). Prosodic marking in speech repairs. *Journal of Semantics, 2*, 205-217.

Levelt, W. J. M., & Kelter, S. (1982). Surface form and memory in question answering. *Cognitive Psychology, 14*, 78-106.

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

Lewis, D. K. (1969). *Convention: A philosophical study.*Cambridge, MA: Harvard University Press.

Liberman, M., & Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff & R. T. Oehrle (Eds.), *Language sound structure.*Cambridge, MA.: The MIT Press.

Lickley, R. J. (1994). *Detecting disfluency in spontaneous speech.* Unpublished PhD. Thesis, University of Edinburgh.

Lickley, R. J. (1995). *Missing disfluencies.* Paper presented at the ICPhS: International Congress of Phonetic Sciences, Stockholm, Sweden.

Lickley, R. J. (1996). *Juncture cues to disfluency.* Paper presented at the ICPhS: International Congress of Phonetic Sciences, Philadelphia.

Lickley, R. J. (1998). HCRC disfluency coding manual: HCRC Technical Report 100.

Lickley, R. J. (2001). *Dialogue moves and disfluency rates.* Paper presented at the DiSS: Disfluency in

Spontaneous Speech, University of Edinburgh, Scotland.

Lickley, R. J., McKelvie, D., & Bard, E. G. (1999). *Comparing human and automatic speech recognition using word gating.* Paper presented at the ICPhS satellite meeting on Disfluency in Spontaneous Speech, University California at Berkeley.

Lickley, R. J., Shillcock, R., & Bard, E. G. (1991). *Understanding disfluent speech: Is there an editing signal?* Paper presented at the Actes du XIIeme Congres International des Sciences Phonetiques, Aix-en-Provence.

Liu, Y., Shriberg, E., & Stolcke, A. (2003). *Automatic disfluency identification in conversational speech using multiple knowledge sources.* Paper presented at the Eurospeech, Geneva, Switzerland.

Liu, Y., Shriberg, E. E., Stolcke, A., & Harper, M. (2005). *Comparing hmm, maximum entropy, and conditional random fields for disfluency detection.* Paper presented at the Eurospeech, Lisbon, Portugal.

Local, J., Kelly, J., & Wells, W. G. H. (1986). Towards a phonology of conversation: Turn-taking in tyneside. *Journal of Linguistics, 22*, 411-437.

Lockridge, C. B., & Brennan, S. E. (2002). Addressees' needs influence speakers' early syntactic choices. *Psychonomic Bulletin and Review, 9*, 550-557.

MacGregor, L. J., Corley, M. C., & Donaldson, D. (2005, September 5-7). *It's.Er.The way that you say it: Hesitations in speech affect language comprehension.* Paper presented at the Architectures and Mechanisms for Language Processing (AMLaP) conference, Ghent, Belgium.

Mackay, D. (1987). *The organisation of perception and action: A theory for language and other cognitive skills.*New York: Springer.

Maclay, H., & Osgood, C. E. (1959). Hesitation phenomena in spontaneous english speech. *Word, 15*, 19-44.

MADCOW. (1992). *Multi-site data collection for a spoken language corpus.* Paper presented at the The Fifth DARPA Speech and Natural Language Workshop, Morgan Kaufmann.

Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition, 8*, 1-71.

Martin, J. G., & Strange, W. (1968). The perception of hesitation in spontaneous speech. *Perception & Psychophysics, 3*, 427-438.

McNeill, D. (1987). *Psycholinguistics: A new approach.*Cambridge, MA: Harper and Row.

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Experimental Psychology: Learning, Memory and Cognition, 49*, 201-213.

Monk, A. F., & Gale, C. (2002). A look is worth a thousand words: Full gaze awareness in video-mediated conversation. *Discourse Processes, 33*(3), 257-278.

Nakatani, C., & Hirschberg, J. (1994). A corpus-based study of repair cues in spontaneous speech. *Journal* *of the Acoustical Society of America, 95*, 1603-1616.

Nicholson, H. (2002). *Prosodic cues to repetitive repair.* Unpublished MPhil dissertation, University of Oxford.

Nicholson, H., Bard, E. G., Lickley, R. J., Anderson, A. H., Havard, C., & Chen, Y. (2005). *Disfluency and behaviour in dialogue: Evidence from eye-gaze.* Paper presented at the DiSS'05: Disfluency in Spontaneous Speech, Aix-en-Provence, France.

Nicholson, H., Bard, E. G., Lickley, R. J., Anderson, A. H., Mullin, J., Kenicer, D., et al. (2003). *The intentionality of disfluency: Findings from feedback and timing.* Paper presented at the Proceedings of DiSS '03: Gothenburgh Papers in Theoretical Linguistics, Gothenburg, Sweden.

Nooteboom, S. (1980). Speaking and unspeaking: Detection and correction of phonological and lexical errors of speech. In V. A. Fromkin (Ed.), *Errors in linguistic performance: Slips of the tongue, ear, pen and hand* (pp. 87-96). New York: Academic Press.

O'Connell, D.C. & Kowal, S. (2005). Uh and Um Revisited: Are they Interjections for Signalling Delay?. *Journal of Psycholinguistic Research, 34,* 555-576.

O'Donnell, W.R. & Todd, L. (1991). Variety in Contemporary English. (2nd Ed.). New York: Harper Collins Academic

Oomen, C. C. E., & Postma, A. (2001a). Effects of divided attention on the production of filled pauses and repetitions. *Journal of Speech, Language and Hearing Research, 44*, 997-1004.

Oomen, C. C. E., & Postma, A. (2001b). Effects of time-pressure on mechanisms of speech production and self-monitoring. *Journal of Psycholinguistic Research, 30*, 163-184.

Oviatt, S. (1995). Predicting spoken disfluencies during human-computer interaction. *Computer, Speech and Language, 9*, 19-35.

Oviatt, S., MacEachern, M., & Levow, G.-A. (1998). Predicting hyperarticulate speech during human-computer error resolution. *Speech Communication, 24*, 1-23.

Page, S. (1999). *Use of a postprocessor to identify and correct speaker disfluencies in automated speech recognition for medical dictations.* Paper presented at the ICPhS satellite meeting of Disfluency and Spontaneous Speech, University of California at Berkeley.

Pakhomov, S. (1999). *Modelling filled pauses in medical dictations.* Paper presented at the 37th Annual Meeting of the Association for Computational Linguistics, College Park, MD.

Pickering, M., & Garrod, S. (2004). Towards a mechanistic theory of dialogue: The interactive alignment model. *The Behaviorial & Brain Sciences, 27*, 169-190.

Pike, K. L. (1945). *The intonation of American English*. Ann Arbor, MI: University of Michigan Press.

Pinker, S. (2000). *The language instinct.*New York: HarperCollins Publishers Inc.

Plauche, M., & Shriberg, E. (1999). *Data-driven subclassification of disfluent repetitions based on prosodic features.* Paper presented at the ICPhS: International Congress of Phonetic Sciences, San Francisco.

Plauché, M., & Shriberg, E. (1999). *Data-driven subclassification of disfluent repetitions based on prosodic features.* Paper presented at the ICPhS: International Congress of Phonetic Sciences, San Francisco.

Postma, A. (2000). Detection of errors during speech production: A review of speech monitoring models. *Cognition, 77*, 97-131.

Postma, A., & Kolk, H. H. J. (1992). The effects of noise-masking and required accuracy on speech errors, disfluencies and self-repairs. *Journal of Speech, Language and Hearing Research, 35*, 537-544.

Postma, A., & Noordanus, C. (1996). The production of speech errors in silent, mouthed, noise-masked, and normal auditory feedback speech. *Language and Speech, 39*, 375-392.

Sanford, A. J., & Garrod, S. (1981). *Understanding written language*.Chichester: John Wiley and Sons.

Savova, G., & Bachenko, J. (2002). *Prosodic features of four types of disfluencies.* Paper presented at the Proceedings of DiSS '03: Gothenburg Papers in Linguistics, Gothenburg, Sweden.

Schegloff, E. A. (1996). Turn organization: One intersection of grammar and interaction. In E. Ochs, E. A. Schegloff & S. Thompson (Eds.), *Interaction and grammar*. Cambridge: Cambridge University Press.

Schegloff, E. A., Sacks, H., & Jefferson, G. (1977). The preference for self-correction in the organization of repair in conversation. *Language, 53*, 361-382.

Schiffer, S. (1972). *Meaning*.Oxford: Clarendon Press.

Schober, M. F. (1993). Spatial and conceptual perspective-taking in conversation. *Cognition, 47*, 4-23.

Schober, M. F., & Brennan, S. (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* (pp. 123-164). Hillsdale, N.J.: Lawrence Erlbaum Associates.

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

Scott, W. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 19*, 321-325.

Searle, J. R., Kiefer, F., & Bierswich, M. (Eds.). (1980). *Speech act theory and pragmatics*.Dordrecht: Holland / Boston: USA/ London: England: D. Reidel Publishing Company.

Selkirk, E. (1995). Sentence prosody: Intonation, stress and phrasing. In J. A. Goldsmith (Ed.), *The handbook of phonological theory* (pp. 550-569). Cambridge, MA and Oxford, UK: Blackwell.

Shriberg, E. E. (1994). *Preliminaries to a theory of speech disfluencies.* PhD. Thesis. University of California: Berkeley.

Shriberg, E. E. (1995). *Acoustic properties of disfluent repetitions.* Paper presented at the ICPhS: International Congress of Phonetic Sciences, Stockholm, Sweden.

Shriberg, E. E. (1999). *Phonetic consequences of speech disfluency.* Paper presented at the International Congress of Phonetic Sciences, San Francisco.

Shriberg, E. E. (2005). *Spontaneous speech: How people really talk and why engineers should care.* Paper presented at the Eurospeech, Lisbon, Portugal.

Shriberg, E. E., Bates, R. A., & Stolcke, A. (1997). *A prosody-only decision-tree model for disfluency detection.* Paper presented at the Eurospeech, Rhodes, Greece.

Shriberg, E. E., & Lickley, R. J. (1993). Intonation of clause-internal filled pauses. *Phonetica, 50*, 172-179.

Siegel, S., & Castellan Jr., N. J. (1988). *Nonparametric statistics for the behavioral sciences*. McGraw-Hill, second edition.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C. W., Price, P. J., et al. (1992). *ToBI: A standard for labeling English prosody.* Paper presented at the International Conference on Speech and Language Processing (ICSLP).

Smith, N. (Ed.). (1982). *Mutual knowledge*.London: Academic Press.

Snedeker, J., & Trueswell, J. C. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language, 48*, 103-130.

Sperber, D., & Wilson, D. (1987). Presumptions of relevance. *The Behaviorial & Brain Sciences, 10*, 736-754.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.). Oxford: Blackwell Publishers Ltd.

Spinos, A.M., O'Connell, D.C., Kowal, S. (2002). An empirical investigation of pause notation. *Journal of Pragmatics, 12,* 1-10.

Stifelman, L. J. (1993). *User repairs of speech recognition errors: An intonational analysis.*MIT Media Laboratory Technical Report. http://www.media.mit.edu/speech/people/lisa/user_repair.html.

Svartvik, J. & Quirk, R. (1980). A corpus of English conversation. Lund, Sweden: Gleerup.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*, 1632-1634.

Trueswell, J. C., & Kim, A. E. (1998). How to prune a garden-path by nipping it in the bud: Fast-priming of verb argument structures. *Journal of Memory and Language, 39*, 102-123.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*.New York: Academic Press.

Vertegaal, R., & Ding, Y. (2002). *Explaining effects of eye gaze on mediated group conversations: Amount or synchronization?* Paper presented at the CSCW 2002, New York.

Watanabe, M., Den, Y., Hirose, K., & Minematsu, N. (2005). *The effects of filled pauses on native and non-native listeners' speech processing.* Paper presented at the DiSS'05. Disfluency in Spontaneous Speech, Aix-en-Provence, France.

Watts, L. A., & Monk, A. F. (1996). *Remote assistance: A view of the work and a view of the face?* Paper presented at the CHI'96, New York.

Weber, R. P. (1985). *Basic content analysis*. Sage Publications.

Wells-Jensen, S. (1999). *Cogntive correlates of linguistic complexity: A cross-linguistic comparison of* *errors in speech.* State University of New York at Buffalo, Buffalo.

Wheatley, B., Doddington, G., Hemphill, C., Godfrey, J., Holliman, E., McDaniel, J., et al. (1992). *Robust automatic time alignment of orthographic transcriptions with unconstrained speech.* Paper presented at the IEEE Conference on Acoustics, Speech and Signal Processing, San Francisco, CA.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America, 91*, 1707-1717.

Wilkes-Gibbs, D., & Clark, H. H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Cognition, 31*, 183-194.

Yasnik, Y., Shattuck-Hufnagel, S., & Veilleux, N. (2005). *Gesture marking of disfluencies in spontaneous speech.* Paper presented at the DiSS'05: Disfluency in Spontaneous Speech, Aix-en-Provence, France.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*, 162-185.

APPENDIX A – PAPERS PUBLISHED BY THE AUTHOR DURING THE COURSE OF THE PHD.

Where block A= No feedback/Verbal only/verbal+visual
Where block B = no feedback/visual only/verbal+visual

APPENDIX G – ORDER OF MAPS BY SUBJECT

For A1 trials, the first three trials were timed; the last three were untimed
For A2 trials, the first three trials were untimed; the last three were timed
For B1 trials, the first three trials were untimed; the last three were timed
For B2 trials, the first three trials were timed; the last three were untimed

Subject#	Condition	trial 1	trial 2	trial 3	trial 4	trial 5	trial 6
1	A1	T	C	D	P	S	M
2	A2	M	T	C	D	P	S
3	A1	S	M	T	C	D	P
4	A2	P	S	M	T	C	D
5	A1	D	P	S	M	T	C
6	A2	C	D	P	S	M	T
7	A1	T	C	D	P	S	M
8	A2	M	T	C	D	P	S
9	A1	S	M	T	C	D	P
10	A2	P	S	M	T	C	D
11	A1	D	P	S	M	T	C
12	A2	C	D	P	S	M	T
13	A1	T	C	D	P	S	M
14	A2	M	T	C	D	P	S
15	A1	S	M	T	C	D	P
16	A2	P	S	M	T	C	D
17	A1	D	P	S	M	T	C
18	A2	C	D	P	S	M	T
1	B1	T	C	D	P	S	M
2	B2	M	T	C	D	P	S
3	B1	S	M	T	C	D	P
4	B2	P	S	M	T	C	D
5	B1	D	P	S	M	T	C
6	B2	C	D	P	S	M	T
7	B1	T	C	D	P	S	M
8	B2	M	T	C	D	P	S
9	B1	S	M	T	C	D	P
10	B2	P	S	M	T	C	D
11	B1	D	P	S	M	T	C
12	B2	C	D	P	S	M	T
13	B1	T	C	D	P	S	M
14	B2	M	T	C	D	P	S
15	B1	S	M	T	C	D	P
16	B2	P	S	M	T	C	D
17	B1	D	P	S	M	T	C
18	B2	C	D	P	S	M	T

APPENDIX H – DISFLUENCIES, WORDS and TRANSACTION COUNTS BY SUBJECT FOR THE VERBAL AND VISUAL GROUPS OF EXPERIMENT 2.

Table 1. Overall totals of transactions, disfluencies, words and average time spent on a trial in seconds for the Verbal Group of Experiment 2

	Timed			Untimed		
MEASURE	None	One	Dual	None	One	Dual
Transactions	**256**	**384**	**398**	**282**	**422**	**454**
Normal	252	263	279	276	304	324
Retrieval	0	106	116	0	106	122
Others	4	15	3	6	12	8
Words	**5235**	**8180**	**9134**	**7502**	**11417**	**12810**
Normal	5179	5261	5790	7386	7338	8769
Retrieval	0	2798	3305	0	3880	3967
Others	56	121	39	116	199	74
Time in Seconds	**121.81**	**189.33**	**214.94**	**186.66**	**277.73**	**311.66**
Disfluencies	**152**	**249**	**265**	**203**	**446**	**530**
Repetitions	59	84	87	95	205	251
Substitutions	54	91	87	58	120	122
Insertions	27	38	34	27	62	63
Deletions	12	36	57	23	59	94
Filled Pauses	**134**	**205**	**234**	**205**	**340**	**346**

Table 2. Overall totals of transactions, disfluencies, words and average time spent on a trial in seconds for the Visual Group of Experiment 2

	Timed			Untimed		
MEASURE	None	One	Dual	None	One	Dual
Transactions	**239**	**259**	**342**	**303**	**334**	**427**
Normal	232	240	243	298	296	295
Retrieval	2	18	93	0	34	123
Others	5	1	6	5	4	9
Words	**6596**	**7443**	**8553**	**9133**	**9121**	**12443**
Normal	6403	7022	5819	9038	8257	8600
Retrieval	175	411	2711	0	819	3804
Others	18	10	23	95	45	39
Time in Seconds	**149.00**	**158.93**	**188.13**	**208.63**	**219.30**	**285.98**
Disfluencies	**150**	**180**	**240**	**183**	**219**	**366**
Repetitions	60	64	91	61	57	135
Substitutions	57	75	68	73	89	118
Insertions	22	18	23	29	33	42
Deletions	11	23	58	20	40	71
Filled Pauses	**157**	**181**	**221**	**280**	**259**	**369**

Table 3. Overall disfluency count for the Verbal Group by trial and subject

OVERALL DISFLUENCIES BY SUBJECT						
	Timed			Untimed		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	2	17	12	13	19	24
2	4	10	11	6	41	41
3	7	16	9	8	10	9
4	13	22	11	7	8	15
5	4	4	7	4	4	2
6	7	8	7	5	9	14
7	12	15	29	12	33	34
8	5	13	15	11	14	10
9	8	16	14	12	9	23
10	5	13	6	2	14	15
11	9	12	14	4	15	19
12	17	13	7	13	18	17
13	10	13	26	16	31	26
14	6	13	14	13	18	30
15	29	44	47	54	153	203
16	3	4	14	5	26	13
17	5	7	12	3	7	6
18	6	9	10	15	17	29
TOTAL	**152**	**249**	**265**	**203**	**446**	**530**

Table 4. Overall disfluency count for The Visual Group by trial and subject

OVERALL DISFLUENCIES BY SUBJECT						
	Timed			Untimed		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	8	8	14	9	10	24
2	11	24	25	22	22	42
3	17	8	10	7	10	28
4	2	7	15	7	11	21
5	6	8	10	7	12	16
6	4	6	2	4	12	8
7	13	11	16	4	9	20
8	3	5	15	9	9	9
9	23	25	20	11	14	27
10	3	10	9	9	17	28
11	11	7	18	12	11	23
12	2	3	13	7	11	12
13	10	17	8	26	23	31
14	13	10	18	19	19	18
15	9	12	13	9	10	14
16	4	4	13	7	8	22
17	6	9	10	9	7	12
18	5	6	11	5	4	11
TOTAL	**150**	**180**	**240**	**183**	**219**	**366**

Table 5. Total Repetitions by Subject for The Verbal Group

OVERALL REPETITIONS BY SUBJECT						
	Timed			Untimed		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	1	4	1	6	5	7
2	1	1	4	2	20	19
3	2	4	5	0	3	2
4	3	4	1	0	0	3
5	2	1	1	3	1	0
6	2	4	0	1	1	3
7	2	6	8	9	10	8
8	2	4	2	3	4	1
9	3	5	4	3	4	7
10	2	3	3	1	9	8
11	3	3	4	2	5	8
12	6	1	2	8	3	5
13	2	4	5	3	8	9
14	3	5	7	7	5	8
15	21	30	33	38	110	145
16	1	1	2	1	9	8
17	2	3	4	2	4	2
18	1	1	1	6	4	8
TOTAL	**59**	**84**	**87**	**95**	**205**	**251**

Table 6. Total Repetitions by Subject for The Visual Group

OVERALL REPETITIONS BY SUBJECT						
	Timed			Untimed		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	3	2	3	1	1	5
2	5	18	16	13	8	22
3	8	5	7	4	2	14
4	1	0	5	2	0	1
5	2	0	4	4	6	4
6	1	1	1	0	4	2
7	6	5	9	2	3	8
8	1	2	6	2	2	3
9	13	8	12	7	4	14
10	1	4	3	1	5	8
11	2	1	5	2	1	9
12	0	1	1	3	2	3
13	2	3	1	8	9	15
14	6	6	7	6	3	6
15	2	3	4	2	3	2
16	2	2	4	1	1	10
17	3	1	2	1	2	3
18	2	2	1	2	1	6
TOTAL	**60**	**64**	**91**	**61**	**57**	**135**

Table 7. Total Substitutions by Subject for The Verbal Group

	OVERALL SUBSTITUTIONS BY SUBJECT					
	Timed			Untimed		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	1	8	5	2	4	8
2	1	4	3	0	7	14
3	1	8	2	6	5	4
4	4	12	5	3	2	6
5	2	1	4	1	1	1
6	2	3	4	4	6	7
7	7	6	13	2	8	12
8	3	3	8	7	7	6
9	3	7	8	4	5	10
10	2	4	1	0	2	2
11	5	4	0	0	3	4
12	7	6	2	3	7	4
13	3	8	8	8	12	5
14	2	5	3	5	7	11
15	4	6	5	7	22	16
16	1	1	6	1	13	4
17	2	0	6	0	2	0
18	4	5	4	5	7	8
TOTAL	**51**	**91**	**87**	**58**	**120**	**122**

Table 8. Total Substitutions for The Visual Group

OVERALL SUBSTITUTIONS BY SUBJECT						
	Timed			Untimed		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	3	3	4	5	2	13
2	2	2	5	7	6	10
3	7	3	2	2	6	9
4	1	5	5	2	8	9
5	4	7	1	0	4	5
6	2	4	1	4	6	3
7	4	4	5	1	5	6
8	1	3	3	3	6	3
9	2	8	4	3	3	5
10	2	4	2	3	7	6
11	8	2	7	7	1	9
12	1	1	5	3	5	4
13	7	11	5	8	10	11
14	5	3	5	7	9	7
15	4	8	3	6	5	7
16	1	1	4	4	4	4
17	2	2	3	6	0	4
18	1	4	4	2	2	3
TOTAL	**57**	**75**	**68**	**73**	**89**	**118**

Table 9. Total Insertions by Subject for The Verbal Group

OVERALL INSERTIONS BY SUBJECT						
	Timed			Untimed		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	0	2	1	5	2	0
2	0	2	1	2	9	6
3	2	2	2	0	1	0
4	5	3	3	2	3	1
5	0	0	1	0	2	1
6	2	0	1	0	2	1
7	3	1	7	1	9	7
8	0	3	3	0	2	3
9	1	3	2	4	0	3
10	1	3	0	1	1	1
11	1	1	2	1	3	4
12	4	5	2	1	3	3
13	3	0	2	2	4	3
14	1	1	2	1	6	7
15	3	6	3	5	8	16
16	1	2	2	1	2	1
17	0	2	0	0	1	3
18	0	2	0	1	4	3
TOTAL	**27**	**38**	**34**	**27**	**62**	**63**

Table 10. Total Insertions by Subject for The Visual Group

OVERALL INSERTIONS BY SUBJECT						
	Timed			Untimed		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	2	0	4	2	3	4
2	3	2	1	1	4	0
3	1	0	1	1	0	1
4	0	0	0	1	2	5
5	0	1	1	2	1	2
6	1	0	0	0	0	2
7	2	1	0	0	1	2
8	0	0	3	3	1	2
9	6	6	0	1	2	3
10	0	0	1	4	4	5
11	0	1	1	2	4	0
12	1	0	0	0	2	3
13	0	2	0	7	1	1
14	1	1	1	2	4	3
15	2	1	3	0	0	2
16	1	0	2	1	2	2
17	1	3	3	1	1	4
18	1	0	2	1	1	1
TOTAL	**22**	**18**	**23**	**29**	**33**	**42**

Table 11. Total Deletions by Subject for The Verbal Group

OVERALL DELETIONS BY SUBJECT						
	Timed			Untimed		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	0	3	5	0	8	9
2	2	3	3	2	5	2
3	2	2	0	2	1	3
4	1	3	2	2	3	5
5	0	2	1	0	0	0
6	1	1	2	0	0	3
7	0	2	1	0	6	7
8	0	3	2	1	1	0
9	1	1	0	1	0	3
10	0	3	2	0	2	4
11	0	4	8	1	4	3
12	0	1	1	1	5	5
13	2	1	11	3	7	9
14	0	2	2	0	0	4
15	1	2	6	4	13	26
16	0	0	4	2	2	0
17	1	2	2	1	0	1
18	1	1	5	3	2	10
TOTAL	**12**	**36**	**57**	**23**	**59**	**94**

Table 12. Total Deletions by Subject for The Visual Group

OVERALL DELETIONS BY SUBJECT						
	Timed			Untimed		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	0	3	3	1	4	2
2	1	2	3	1	4	10
3	1	0	0	0	2	4
4	0	2	5	2	1	6
5	0	0	4	1	1	5
6	0	1	0	0	2	1
7	1	1	2	1	0	4
8	1	0	3	1	0	1
9	2	3	4	0	5	5
10	0	2	3	1	1	9
11	1	3	5	1	5	5
12	0	1	7	1	2	2
13	1	1	2	3	3	4
14	1	0	5	4	3	2
15	1	0	3	1	2	3
16	0	1	3	1	1	6
17	0	3	2	1	4	1
18	1	0	4	0	0	1
TOTAL	**11**	**23**	**58**	**20**	**40**	**71**

Table 13. Total Filled Pauses by Subject for The Verbal Group

OVERALL FILLED PAUSES BY SUBJECT						
	Timed			Untimed		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	1	7	6	8	19	17
2	22	20	18	35	46	53
3	7	8	6	3	10	10
4	0	7	11	7	9	17
5	6	8	7	5	8	11
6	3	13	7	10	12	19
7	6	12	19	10	18	17
8	4	8	7	5	6	7
9	5	7	7	13	13	14
10	2	13	16	5	19	15
11	12	20	32	11	40	29
12	2	4	4	3	7	11
13	4	12	5	5	12	14
14	5	5	6	10	4	12
15	28	29	28	48	72	59
16	10	8	9	16	14	15
17	15	23	33	6	24	17
18	2	1	13	5	7	9
TOTAL	**134**	**205**	**234**	**205**	**340**	**346**

Table 14. Total Filled Pauses by Subject for The Visual Group

			OVERALL FILLED PAUSES BY SUBJECT			
	Timed			Untimed		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	24	7	14	19	13	24
2	23	24	18	35	26	26
3	9	10	5	16	20	17
4	3	5	16	9	2	22
5	1	1	0	1	9	9
6	6	7	14	7	13	17
7	1	6	3	4	3	7
8	0	3	21	10	13	39
9	19	12	4	9	19	22
10	2	2	5	8	12	17
11	3	11	7	21	14	24
12	1	10	14	5	8	9
13	12	18	12	49	40	47
14	6	3	7	7	1	9
15	15	15	19	21	20	25
16	5	5	15	7	9	17
17	2	2	2	4	4	2
18	25	40	45	48	33	36
TOTAL	**157**	**181**	**221**	**280**	**249**	**369**

Table 15. Total Word Count per trial by Subject for The Verbal Group

TOTAL RAW WORD COUNT BY SUBJECT						
	Timed			Untimed		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	240	484	386	335	584	568
2	265	428	507	513	831	873
3	441	543	657	311	508	523
4	194	503	483	278	360	434
5	251	338	333	219	386	401
6	256	406	441	334	506	783
7	366	643	716	289	527	716
8	210	529	553	722	699	647
9	490	640	534	602	577	598
10	286	482	431	293	607	629
11	331	546	890	341	839	896
12	232	340	316	379	402	392
13	312	308	387	343	583	511
14	228	330	402	459	572	938
15	516	693	870	843	1911	2328
16	236	340	379	346	570	584
17	138	311	390	129	315	291
18	243	316	459	766	640	698
TOTAL	**5235**	**8180**	**9134**	**7502**	**11417**	**12810**

Table 16. Total Raw Word Count by Subject for The Visual Group

TOTAL RAW WORD COUNT BY SUBJECT						
	Timed			Untimed		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	696	546	557	582	468	1113
2	379	515	557	647	659	917
3	393	513	572	464	508	746
4	243	366	472	475	450	855
5	416	437	372	492	582	679
6	193	206	248	279	346	399
7	352	538	614	308	331	545
8	258	293	446	537	481	631
9	572	661	770	607	710	845
10	268	246	328	424	440	749
11	370	273	351	607	676	665
12	357	450	494	596	509	661
13	455	483	479	833	693	994
14	302	265	367	428	471	444
15	381	561	639	573	583	656
16	244	268	339	376	300	451
17	388	404	433	391	411	565
18	336	418	515	514	503	528
TOTAL	**6596**	**7443**	**8553**	**9133**	**9121**	**12443**

Table 17. Total Transactions by Subject for The Verbal Group

	TOTAL TRANSACTIONS BY SUBJECT					
	Timed			Untimed		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	13	27	21	15	17	22
2	17	25	22	18	27	28
3	17	25	24	16	29	22
4	8	22	22	14	18	21
5	13	16	21	15	20	24
6	15	19	23	13	21	27
7	16	18	21	9	19	22
8	10	19	19	24	24	22
9	15	21	11	15	19	19
10	18	29	22	14	20	25
11	15	27	28	11	35	40
12	16	19	24	19	22	22
13	14	14	20	14	21	20
14	8	19	20	19	26	30
15	16	28	31	20	44	37
16	17	18	21	13	19	27
17	13	19	20	15	19	19
18	16	19	28	18	22	27
TOTAL	**257**	**384**	**398**	**282**	**422**	**454**

Table 18. Total Transactions by Subject for The Visual Group

TOTAL TRANSACTIONS BY SUBJECT						
	Timed			Untimed		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	20	15	15	18	19	30
2	16	19	16	20	24	26
3	12	12	17	13	14	19
4	12	11	24	18	16	24
5	12	10	21	17	23	27
6	7	16	21	14	23	21
7	18	22	30	14	13	24
8	15	13	18	19	20	25
9	18	22	26	16	25	24
10	13	13	17	16	10	22
11	12	9	15	18	22	20
12	12	17	21	15	17	25
13	16	13	15	22	20	23
14	14	12	15	15	19	23
15	10	13	17	19	18	29
16	11	12	20	15	14	19
17	12	13	16	15	20	25
18	9	17	18	19	17	21
TOTAL	**239**	**259**	**342**	**303**	**334**	**427**

Table 19. Total Normal Transactions by Subject for The Verbal Group

NORMAL TRANSACTIONS BY SUBJECT						
	Timed			Untimed		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	13	22	15	15	13	16
2	17	17	17	18	18	22
3	16	19	18	16	22	15
4	8	16	16	13	12	15
5	13	11	15	15	15	18
6	14	11	13	12	13	17
7	16	12	12	9	13	14
8	10	11	15	24	18	16
9	14	16	7	15	12	12
10	16	16	16	14	15	19
11	15	18	19	10	25	31
12	16	13	18	18	16	18
13	14	10	13	14	15	14
14	8	13	13	19	19	20
15	16	22	24	19	36	29
16	17	12	15	13	14	19
17	13	12	14	15	12	12
18	16	12	19	17	16	17
TOTAL	**252**	**263**	**279**	**276**	**304**	**324**

Table 20. Total Normal Transactions by Subject for The Visual Group

NORMAL TRANSACTIONS BY SUBJECT						
	Timed			Untimed		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	18	15	13	18	18	21
2	16	13	11	20	19	18
3	12	12	13	13	13	15
4	12	10	15	16	16	17
5	12	9	15	17	18	19
6	7	15	15	13	17	13
7	17	18	23	13	11	16
8	14	13	12	18	19	17
9	17	17	16	16	17	16
10	12	13	13	16	10	15
11	12	9	11	18	18	13
12	11	17	15	15	17	17
13	16	13	11	22	18	18
14	14	12	10	15	17	16
15	10	13	13	19	18	22
16	11	12	13	15	14	11
17	12	12	11	15	19	16
18	9	17	13	19	17	15
TOTAL	**232**	**240**	**243**	**298**	**296**	**295**

Table 21. Total Retrieval Transactions by Subject for The Verbal Group

RETRIEVAL TRANSACTIONS BY SUBJECT						
	Timed			Untimed		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	0	5	6	0	4	6
2	0	7	5	0	6	6
3	0	6	6	0	6	6
4	0	6	6	0	6	6
5	0	5	6	0	5	6
6	0	7	9	0	7	9
7	0	6	9	0	6	8
8	0	8	4	0	6	6
9	0	5	4	0	7	7
10	0	7	6	0	5	6
11	0	4	8	0	6	6
12	0	6	6	0	6	4
13	0	4	7	0	5	6
14	0	6	7	0	7	8
15	0	6	6	0	6	8
16	0	6	6	0	5	7
17	0	6	6	0	7	7
18	0	6	9	0	6	10
TOTAL	**0**	**106**	**116**	**0**	**106**	**122**

Table 22. Total Retrieval Transactions by Subject for The Visual Group

	RETRIEVAL TRANSACTIONS BY SUBJECT					
	Timed			Untimed		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	2	0	2	0	1	9
2	0	6	5	0	5	8
3	0	0	4	0	1	4
4	0	1	8	0	0	6
5	0	1	5	0	5	7
6	0	0	6	0	6	8
7	0	4	7	0	2	6
8	0	0	5	0	0	6
9	0	5	9	0	6	7
10	0	0	4	0	0	7
11	0	0	4	0	4	7
12	0	0	6	0	0	7
13	0	0	4	0	1	5
14	0	0	5	0	2	7
15	0	0	4	0	0	7
16	0	0	7	0	0	8
17	0	1	4	0	1	9
18	0	0	4	0	0	5
TOTAL	**2**	**18**	**93**	**0**	**34**	**123**

Table 23. Total Other Transactions by Subject for The Verbal Group

OTHER TRANSACTIONS BY SUBJECT						
	Timed			Untimed		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	0	0	0	0	0	0
2	0	1	0	0	3	0
3	1	0	0	0	1	1
4	0	0	0	1	0	0
5	0	0	0	0	0	0
6	1	1	1	1	1	1
7	0	0	0	0	0	0
8	0	0	0	0	0	0
9	1	0	0	0	0	0
10	2	6	0	0	0	0
11	0	5	1	1	4	3
12	0	0	0	1	0	0
13	0	0	0	0	1	0
14	0	0	0	0	0	2
15	0	0	1	1	2	0
16	0	0	0	0	0	1
17	0	1	0	0	0	0
18	0	1	0	1	0	0
TOTAL	**5**	**15**	**3**	**6**	**12**	**8**

Table 24. Total Other Transactions by Subject for The Visual Group

OTHER TRANSACTIONS BY SUBJECT						
	Timed			Untimed		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	1	2	0	1
5	0	0	1	0	0	1
6	0	1	0	1	0	0
7	1	0	0	1	0	2
8	1	0	1	1	1	2
9	1	0	1	0	2	1
10	1	0	0	0	0	0
11	0	0	0	0	0	0
12	1	0	0	0	0	1
13	0	0	0	0	1	0
14	0	0	0	0	0	0
15	0	0	0	0	0	0
16	0	0	0	0	0	0
17	0	0	1	0	0	0
18	0	0	1	0	0	1
TOTAL	**5**	**1**	**6**	**5**	**4**	**9**

APPENDIX I – TRANSCRIPT FOR AUDIO EXAMPLE ON CD-ROM OF OUTLYING SUBJECT’S TYPICAL DISFLUENCIES

Play 15t-vsvr-u.2.wma on CD-ROM

Subject 15, Dual Feedback, Untimed Condition

“Ehm…so if you…I’ll tell you w-….see if you see if you can try and ehm if you can look see if you move…see if you…see if you right…if you stop where you are right now okay”

APPENDIX J – TRANSCRIPTS OF DUAL-FEEDBACK MODALITY AND EXAMPLES OF PLANNING AND HESITATION DISFLUENCIES

NOTE: The reliability judgements reported in Section 3.12 were done on these and other similar materials.

Clip 1, Subject 2, Visual Group, Experiment 2
Example of Hesitation Repetition
PLAY s2-both-clip1-rep.avi **ALONG WITH THIS TRANSCRIPT**

233.24	IF feedback arrives at Overgrowngully		
246.56	3.52 of MUTUAL GAZE at Overgrowngully	IG starts gazing at Overgrowngully	
246.903	begin normal_transaction	begin instruct_move	You loop round the overgrown gully , um ,
250.08	end of MUTUAL GAZE	IG stops gazing at Overgrowngully	
251.36	IG starts gazing at Giraffesnorth		
251.744	**begin disfluency (r)**		**<R to the**
252.405	**end disfluency (r)**		**to the R>** left-hand side, to the right-hand side of it .
254.594	end instruct_move	begin instruct_move	And then head north.
255.793	end normal_transaction	end instruct_move	
255.793	begin normal transaction	begin interactive move	Um, do you have any giraffes at all
256.68	IG stops gazing at Giraffesnorth		
259.32	IF feedback leaves Overgrowngully		

Clip 2, Subject 4, Visual Group, Experiment 2
Example of Planning Deletion
PLAY s4-both-clip2-del.avi **ALONG WITH THIS TRANSCRIPT**

66.88	IF feedback arrives at Rocks	0.2 of MUTUAL GAZE at Rocks	
67.08	end of MUTUAL GAZE	IG stops gazing at Rocks	
67.097	end explain_move	begin explain_move	so it's d just down from the rift valley
68.32	IG starts gazing at Rift Valley		
68.52	IG stops gazing at Rift Valley		
69.36	IG starts gazing at Outlawshideout		
69.52	IG stops gazing at Outlawshideout		
69.582	end explain_move	begin interactive_move	
70.627	**begin disfluency (d)**		**<D just**,
70.80	0.32 of MUTUAL GAZE at Rocks	IG starts gazing at Rocks	
70.96	**end disfluency (d)**		**D>** I think you are looking at the right spot
71.12	end of MUTUAL GAZE	IG stops gazing at Rocks	
71.92	IG starts gazing at Noose		
71.96	IG stops gazing at Noose		
72.00	2.48 of MUTUAL GAZE at Rocks	IG starts gazing at Rocks	
72.582	end interactive_move	begin instruct_move	we will loop underneath those rocks
74.48	end of MUTUAL GAZE	IG stops gazing at Rocks	IG starts gazing at StonecreekN
74.88	IG stops gazing at StonecreekN		
74.92	0.4 of MUTUAL GAZE at Rocks	IG starts gazing at Rocks	
75.271	end instruct_move	begin instruct_move	and head up diagonally left towards the stone creek
75.32	end of MUTUAL GAZE	IG stops gazing at Rocks	IG starts gazing at StonecreekN
75.56	0.24 of MUTUAL GAZE at Rocks	IG stops gazing at StonecreekN	IG starts gazing at Rocks
75.80	end of MUTUAL GAZE	IG stops gazing at Rocks	IG starts gazing at StonecreekN
76.28	IG stops gazing at StonecreekN	IG starts gazing at Whitewater	
77.04	IG stops gazing at Whitewater	IG starts gazing at StonecreekN	
78.84	IG stops gazing at StonecreekN	IG starts gazing at Whitewater	
80.432	end normal_transaction	end instruct_move	

Clip 3, Subject 13, Motivated Subject from Experiment 3
Example of Hesitation Deletion
PLAY s13-both-clip3-del.avi **ALONG WITH THIS TRANSCRIPT**

38.64	IF feedback arrives at Ropebridge	Negative verbal feedback started		
38.68	4.16 of MUTUAL GAZE at Ropebridge	IG starts gazing at Ropebridge		
42.433	begin normal_transaction	begin explain_move		
42.84	end of MUTUAL GAZE	IG stops gazing at Ropebridge	IG starts gazing at Flamingoes	
43.00	0.48 of MUTUAL GAZE at Ropebridge	IG stops gazing at Flamingoes	IG starts gazing at Ropebridge	
43.451	**begin disfluency (d)**		**<D** ehm we are	
43.48	end of MUTUAL GAZE	IG stops gazing at Ropebridge		
43.72	IG starts gazing at Fallen Pillars			
43.88	0.32 of MUTUAL GAZE at Ropebridge	IG stops gazing at Fallen Pillars	IG starts gazing at Ropebridge	
43.935	**end disfluency (d)**		**D>**	
44.20	end of MUTUAL GAZE	IG stops gazing at Ropebridge	IG starts gazing at Fallen Pillars	
44.201	end explain_move	begin instruct_move	Now move straight eh from there towards the right-hand side of the page	
44.44	1.72 of MUTUAL GAZE at Ropebridge	IG stops gazing at Fallen Pillars	IG starts gazing at Ropebridge	
46.16	end of MUTUAL GAZE	IG stops gazing at Ropebridge	IG starts gazing at Flamingoes	IG stops gazing at Flamingoes
46.36	1.12 of MUTUAL GAZE at	IG starts gazing at		

	Ropebridge	Ropebridge				
47.48	end of MUTUAL GAZE	IG stops gazing at Ropebridge				
47.68	IG starts gazing at Flamingoes					
47.92	0 of MUTUAL GAZE at Ropebridge	end of MUTUAL GAZE	IG stops gazing at Flamingoes	IG starts gazing at Ropebridge	IG stops gazing at Ropebridge	IG starts gazing at Flamingoes
47.96	0.4 of MUTUAL GAZE at Ropebridge	IG stops gazing at Flamingoes	IG starts gazing at Ropebridge			
48.36	end of MUTUAL GAZE	IG stops gazing at Ropebridge				
48.52	IG starts gazing at Waterfall					
49.40	IG stops gazing at Waterfall					
49.68	0.4 of MUTUAL GAZE at Ropebridge	IG starts gazing at Ropebridge				
50.08	end of MUTUAL GAZE	IG stops gazing at Ropebridge				
50.36	IG starts gazing at Fallen Pillars					
50.44	IG stops gazing at Fallen Pillars					
50.88	IG starts gazing at Waterfall					
50.96	1.28 of MUTUAL GAZE at Ropebridge	IG stops gazing at Waterfall	IG starts gazing at Ropebridge			
51.715	end instruct_move	begin explain_move	towards the right-hand side of the page			
52.24	end of MUTUAL GAZE	IG stops gazing at Ropebridge				
52.40	IG starts gazing at Waterfall					
52.88	0.44 of MUTUAL GAZE at Ropebridge	IG stops gazing at Waterfall	IG starts gazing at Ropebridge			
53.32	end of MUTUAL GAZE	IG stops gazing at Ropebridge	IG starts gazing at Flamingoes	IG stops gazing at Flamingoes		
53.76	IG starts gazing at					

	Waterfall		
54.08	1 of MUTUAL GAZE at Ropebridge	IG stops gazing at Waterfall	IG starts gazing at Ropebridge
55.08	end of MUTUAL GAZE	IG stops gazing at Ropebridge	IG starts gazing at Flamingoes
55.349	end explain_move	begin explain_move	from there straight towards the right-hand side of the
55.44	IG stops gazing at Flamingoes		
55.68	IG starts gazing at Fallen Pillars		
55.80	IG stops gazing at Fallen Pillars		

Clip 4, Subject 10, Motivated Group
Example of Planning Repetition
PLAY s10-clip4-rep.avi **ALONG WITH THIS TRANSCRIPT**

140.28	IG starts gazing at Stones		
142.00	IF feedback arrives at Ancientruins	Negative verbal feedback started	“No, not with you”
142.832	begin retrieval_transaction	begin interactive_move	well where
143.308	**begin disfluency (r)**		**<R you're**
143.798	**end disfluency (r)**		**you're R>** looking now I've got some ancient ruins
146.293	end interactive_move	begin explain_move	You want to pass up the left-hand side of that
149.219	end retrieval_transaction	end explain_move	
156.76	IF feedback leaves Ancientruins	Negative verbal feedback ended	
158.92	IG stops gazing at Stones		

APPENDIX J - DISFLUENCIES, WORDS and TRANSACTION COUNTS BY SUBJECT FOR THE MOTIVATED AND CONTROL GROUPS OF EXPERIMENT 3.

Table 1. Overall disfluencies for Experiment 3 by Subject and Feedback Modality

OVERALL DISFLUENCIES BY SUBJECT						
	MOTIVATED			CONTROL		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	2	17	34	9	10	24
2	7	11	19	7	10	28
3	7	4	5	7	12	16
4	8	8	14	4	9	20
5	5	8	26	11	14	27
6	8	11	20	12	11	23
7	6	19	15	26	23	31
8	5	18	9	9	10	14
9	2	23	21	9	7	12
TOTAL	**50**	**119**	**163**	**94**	**106**	**195**

Table 2. Overall repetitions for Experiment 3 by Subject and Feedback Modality

OVERALL REPETITIONS BY SUBJECT						
	MOTIVATED			CONTROL		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	0	2	6	1	1	5
2	3	1	7	4	2	14
3	2	0	1	4	6	4
4	3	0	1	2	3	8
5	1	5	6	7	4	14
6	2	3	3	2	1	9
7	1	9	9	8	9	15
8	2	8	3	2	3	2
9	0	14	9	1	2	3
TOTAL	**14**	**42**	**45**	**31**	**31**	**74**

Table 3. Overall substitutions for Experiment 3 by Subject and Feedback Modality

OVERALL SUBSTITUTIONS BY SUBJECT						
	MOTIVATED			CONTROL		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	2	3	8	5	2	13
2	1	0	4	2	6	9
3	1	1	0	0	4	5
4	3	4	3	1	5	6
5	4	2	10	3	3	5
6	2	2	6	7	1	9
7	1	8	2	8	10	11
8	0	1	1	6	5	7
9	2	5	4	6	0	4
TOTAL	**16**	**26**	**38**	**38**	**36**	**69**

Table 4. Overall insertions for Experiment 3 by Subject and Feedback Modality

OVERALL INSERTIONS BY SUBJECT						
	MOTIVATED			CONTROL		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	0	2	4	2	3	4
2	2	5	2	1	0	1
3	1	0	2	2	1	2
4	1	1	5	0	1	2
5	0	0	1	1	2	3
6	1	2	2	2	4	0
7	1	1	0	7	1	1
8	2	1	0	0	0	2
9	0	3	0	1	1	4
TOTAL	**8**	**15**	**16**	**16**	**13**	**17**

Table 5. Overall deletions for Experiment 3 by Subject and Feedback Modality

OVERALL DELETIONS BY SUBJECT						
	MOTIVATED			CONTROL		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	0	10	16	1	4	2
2	1	5	6	0	2	4
3	3	3	2	1	1	5
4	1	3	5	1	0	4
5	0	1	9	0	5	5
6	3	4	9	1	5	5
7	3	1	4	3	3	4
8	1	8	5	1	2	3
9	0	1	8	1	4	1
TOTAL	**12**	**36**	**64**	**9**	**26**	**33**

Table 6. Overall filled pauses for Experiment 3 by Subject and Feedback Modality

OVERALL FILLED PAUSES BY SUBJECT						
	MOTIVATED			CONTROL		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	11	14	3	19	13	24
2	12	9	10	16	20	17
3	1	1	5	1	9	9
4	2	2	3	4	3	7
5	0	4	1	9	19	22
6	0	2	15	21	14	24
7	5	5	4	49	40	47
8	26	29	19	21	20	25
9	7	11	14	4	4	2
TOTAL	**64**	**77**	**74**	**144**	**142**	**177**

Table 7. Total raw word count for Experiment 3 by Subject and Feedback Modality

	TOTAL RAW WORD COUNT BY SUBJECT					
	MOTIVATED			CONTROL		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	214	607	922	582	468	1113
2	452	552	581	464	508	746
3	283	421	444	492	582	679
4	379	696	745	308	331	545
5	314	379	920	607	710	845
6	530	966	761	607	676	665
7	203	546	403	833	693	994
8	294	489	453	573	583	656
9	401	1381	1217	391	411	565
TOTAL	**3070**	**6037**	**6446**	**4857**	**4962**	**6808**

Table 8. Overall transactions for Experiment 3 by Subject and Feedback Modality

	OVERALL TRANSACTIONS BY SUBJECT					
	MOTIVATED			CONTROL		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	16	23	35	18	19	30
2	19	24	26	13	14	19
3	14	18	21	17	23	27
4	20	29	33	14	13	24
5	18	20	34	16	25	24
6	22	35	30	18	22	20
7	14	30	24	22	20	23
8	21	25	19	19	18	29
9	16	36	37	15	20	25
TOTAL	**160**	**240**	**259**	**152**	**174**	**221**

Table 9. Total Normal Transactions for Experiment 3 by Subject and Feedback Modality

NORMAL TRANSACTIONS BY SUBJECT						
	MOTIVATED			CONTROL		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	15	15	21	18	18	21
2	18	15	19	13	13	15
3	13	16	14	17	18	19
4	18	17	20	13	11	16
5	17	16	22	16	17	16
6	19	26	21	18	18	13
7	12	21	16	22	18	18
8	18	17	14	19	18	22
9	14	25	25	15	19	16
TOTAL	**144**	**168**	**172**	**151**	**150**	**156**

Table 10. Total Retrieval Transactions for Experiment 3 by Subject and Feedback Modality

RETRIEVAL TRANSACTIONS BY SUBJECT						
	MOTIVATED			CONTROL		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	0	6	12	0	1	9
2	0	6	5	0	1	4
3	0	0	5	0	5	7
4	0	8	11	0	2	6
5	0	1	11	0	6	7
6	1	7	7	0	4	7
7	0	8	7	0	1	5
8	0	6	5	0	0	7
9	0	10	12	0	1	9
TOTAL	**1**	**52**	**75**	**0**	**21**	**61**

Table 11. Total Other Transactions for Experiment 3 by Subject and Feedback Modality

OTHER TRANSACTIONS BY SUBJECT						
	MOTIVATED			CONTROL		
SUBJECT	No Feedback	Verbal-only Feedback	Dual Feedback	No Feedback	Verbal-only Feedback	Dual Feedback
1	1	2	2	0	0	0
2	1	3	2	0	0	0
3	1	2	2	0	0	1
4	2	4	2	1	0	2
5	1	3	1	0	2	1
6	2	2	2	0	0	0
7	2	1	1	0	1	0
8	3	2	0	0	0	0
9	2	1	0	0	0	0
TOTAL	**15**	**20**	**12**	**1**	**3**	**4**

Proceedings of DiSS'03, Disfluency in Spontaneous Speech Workshop. 5–8 September 2003, Göteborg University, Sweden.
Robert Eklund (ed.), Gothenburg Papers in Theoretical Linguistics 89, ISSN 0349–1021, pp. 15–18.

The intentionality of disfluency: Findings from feedback and timing

Hannele Nicholson[1], Ellen Gurman Bard[1], Robin Lickley[2],
Anne H. Anderson[3], Jim Mullin[3], David Kenicer[3] & Lucy Smallwood[3]

[1] University of Edinburgh, Edinburgh, Scotland
[2] Queen Margaret University College, Edinburgh, Scotland
[3] University of Glasgow, Glasgow, Scotland

Abstract

This paper addresses the causes of disfluency. Disfluency has been described as a strategic device for intentionally signalling to an interlocutor that the speaker is committed to an utterance under construction [14, 21]. It is also described as an automatic effect of cognitive burdens, particularly of managing speech production during other tasks [6]. To assess these claims, we used a version of the map task [1, 11] and tested 24 normal adult subjects in a baseline untimed monologue condition against conditions adding either feedback in the form of an indication of a supposed listener's gaze, or time-pressure, or both. Both feedback and time-pressure affected the nature of the speaker's performance overall. Disfluency rate increased when feedback was available, as the strategic view predicts, but only deletion disfluencies showed a significant effect of this manipulation. Both the nature of the deletion disfluencies in the current task and of the information which the speaker would need to acquire in order to use them appropriately suggest ways of refining the strategic view of disfluency.

1. Introduction

Disfluency is known to be more common in dialogue than in monologue [19]. Explanations for this fact fall into two categories. One ties disfluency to active strategies for cultivating common ground, the accumulating knowledge that interlocutors are mutually conscious of sharing [9, 13, 21], while the other sees disfluency as an accidental result of cognitive burdens [6], which necessarily increase when a speaker must process a listener's utterances while composing his or her own.

In the strategic view, disfluency is one of a number of intentional strategies which speakers employ to maintain mutuality. Clark and Wasow [14] argue that repetition disfluencies are strategically deployed to signal ongoing difficulty in producing an utterance to which the speaker is nonetheless committed. Evidence of prosodic cues that signal strategic intention has been obtained for repetitive repair [21].

In the alternate view, conversation is a cognitively taxing process and competition is high for production resources [3, 4, 9, 15, 16]. A speaker must design the sub-goals of any task which a dialogue helps the interlocutors to pursue, plan the sections of the dialogue which correspond to these goals, and attend to the contributions of the interlocutor, while micro-planning his/her own utterances [4, 5]. Disfluencies may occur when this burden becomes so great that errors in planning or production are not detected and edited covertly before articulation begins. Increases in disfluency accompanying increased complexity of any of the cognitive functions underlying dialogue are taken to support this view. Long utterances, which tend to be more complex than short, certainly tend to be disfluent more often [14]. Bard and her colleagues have shown that even with utterance length taken into account, production burdens correlate with disfluency: formulating multi-reference utterances and initiating new sections of the dialogue both tend to encourage disfluency. In contrast, no characteristics of the prior interlocutor utterance have any independent effect on disfluency rate. This account of disfluency joins other models of dialogue phenomena in ascribing to the speaker's own current needs many of the behaviours which are often thought to be adaptations to a developing model of the listener's knowledge [See 2, 3, 4, 5, 8, 20].

This paper presents the first group of results from a series of experiments designed to discover whether speakers are more concerned with attending to their listeners' knowledge or completing their own production tasks. The experiments use a variant of the map task [1, 11]. In the original task, players have before them versions of a cartoon map representing a novel imaginary location. The Instruction Giver communicates to the Instruction Follower a route pre-printed on the Giver's map. The current series uses only Instruction Givers and manipulates both time-pressure and feedback from a presumptive Follower.

The time-pressure variable contrasts instructions composed in the Giver's own time with a time-limited condition. If disfluencies are a basic signaling device and important to the conduct of a dialogue, then this manipulation will not affect them. If disfluencies are failures of planning, time-pressure should increase their rate of occurrence. If, on the other hand, disfluencies are a luxury, a rhetorical device available to speakers but not required for the process of maintaining mutual knowledge, then they may be more common when interlocutors have the time to indulge in them, that is, in the untimed condition.

The feedback variable contrasts monologue map tasks, supposedly transmitted to a listener in another room, with tasks for which there is minimal feedback in the form of a square projected on the map to represent the direction of the Follower's gaze. If modeling the listener's knowledge is critical to the process of dialogue, then this is the most important kind of feedback, for it tells one interlocutor what the other knows about the map and how s/he interprets the instructions. If speakers treat these tasks as interactive, and if disfluency is an intentionally helpful signal, then disfluency should be more common in this condition than in pure monologue. For example, repetition disfluency should be induced by the availability of the listener [14].

The interactions of these two manipulations are of particular interest. A pure strategic model demands a main effect of feedback but would sit well with enhanced rates of disfluency in the feedback condition with time pressure, where most difficulties would arise. A pure cognitive difficulty model predicts enhanced rates of disfluency under time pressure, but particularly again where feedback and time-pressure both add to the speaker's cognitive burdens. Associated with the cognitive difficulty model are a set of results which could support a hybrid view: that listener-centric behaviour in dialogue is a luxury [15, 16] which will be abandoned when the speaker has more pressing tasks to pursue. This model predicts that disfluencies will appear at a higher rate where feedback makes the task interactive and where ample time permits the consideration of the listener's needs.

2. Method

2.1. Task

Disfluencies are obtained from the MONITOR corpus currently under collection [7]. This corpus employs a variant of the map task [1, 11]. In this version of the MONITOR task, subjects are seated before a computer screen displaying a map of a fictional location which includes a route from a marked start-point to buried treasure. Labelled landmarks and map designs are adapted from the HCRC Map Task Corpus [1]. Subjects are requested to help a distant listener reproduce the route. Subjects' instructions were recorded onto the video record by a close-talking microphone and their gaze direction was recorded by a screen-mounted eye-tracker. At the beginning of each trial, the tracker was calibrated.

2.2. Experimental Design

The experiment crossed feedback (2) and time-pressure (2). In the no feedback conditions, subjects saw only the map. In the feedback condition, a small moving square was superimposed on the map and subjects were told that this represented the current direction of their Instruction Follower's gaze. Unbeknownst to the subjects, there was no actual Follower. The feedback gaze-square followed a pre-programmed sequence. It remained on the landmarks determining the route until the first two or three had been successfully negotiated. Subsequently, feedback gaze wandered off-course at least once every other landmark The pattern of incorrect gaze-responses corresponded roughly to the distribution of landmarks which did not match across Giver and Follower maps in [1]. In four cases in each map, the feedback square did not go to the intended landmark, but instead moved to a second, but distant, copy of that landmark or to a space on the map which would have hosted a landmark on the Follower's version of the corresponding HCRC map. In each case, once the subject had introduced the next route-critical landmark, an experimenter in another room advanced the feedback gaze square to its next scheduled target. The square moved about its target landmark in a realistic fashion, with sorties of random radius and angle.

Crossed with feedback was the time-pressure variable. In half of the trials, speakers were permitted only one minute to complete the task; otherwise time was unlimited.

Subjects with normal uncorrected vision were recruited from the Glasgow University community. All were paid for their time. All encountered all 4 conditions. Four different basic maps were used, counter-balanced across conditions over the whole design. Subjects were eliminated if any single map trial failed to meet criteria for feedback or capture quality. The feedback criterion demanded that the experimenter advance the feedback square between the introduction of the pertinent landmark and the onset of the following instruction in all cases where where the feedback was scheduled to be errant and in 70% where the square's movement was scheduled to be correct. The capture criterion demanded that at least 80% of the eye-tracking data was intact. Fifty-four subjects were run before 24 remained with valid sessions in all conditions and with a balanced design in total.

3. Results

3.1. Dialogue Structure

Each monologue was transcribed verbatim and then coded for transaction [12]. A transaction is a block of speech in task-oriented dialogue which accomplishes a task sub-goal. Accordingly, in this task Normal transactions are periods of standard instruction giving. Review transactions recount the route negotiated thus far. Overviews describe the route or map in general. Irrelevant transactions are all off-task remarks.

A fifth type of transaction, Retrievals, was identified in the present monologues and can be used to show that the feedback conditions were in fact interactive. In a Retrieval the speaker neither gives new instructions nor reviews the route but instead moves the presumed IF to a previously named landmark where s/he should be but apparently is not. Figure 1, which divides Transactions by type in each of the four conditions, shows that Retrievals occurred in the two feedback conditions (13% of all Transactions in Feedback-Timed; 18% in Feedback-Untimed) but very rarely otherwise (0.8% of all No Feedback Timed Transactions and 0.3% of No Feedback Untimed: by-subjects 2 × 2 repeated measures ANOVA main effect for Feedback, $F_1(1,23) = 25.84$, $p < .001$). The imbalance suggests that Retrievals are unlikely to be mere clarifications, independent of the IF's behaviour. Since each speaker encountered 4 off-route gaze locations per dialogue, the average number of Retrieval transactions per dialogue, 1.58 for Feedback Timed; 2.58 for Feedback Untimed, shows fairly good uptake of the feedback square's 'mistakes'. The effect of Time-pressure approached significance ($F_1(1,23) = 4.12$, $p = .054$). but only because of an increase in Retrievals in Feedback conditions (interaction: $F_1(1,23) = 5.40$, $p = .029$).

As Figure 1 also shows, Retrievals do not follow the general trends for volume of transactions. Both Normal transactions and total number of transactions are more numerous in the Untimed conditions (11.40 Normal transactions, 13.83 in total per trial) than in the Timed (9.63 Normal, 11.27 total)

($F_1(1,23) = 5.77$, $p = .025$ for normal; $F_1(1,23) = 9.95$, $p < .01$, overall), with no effect of feedback. Other transaction types were unaffected by the experimental variables.

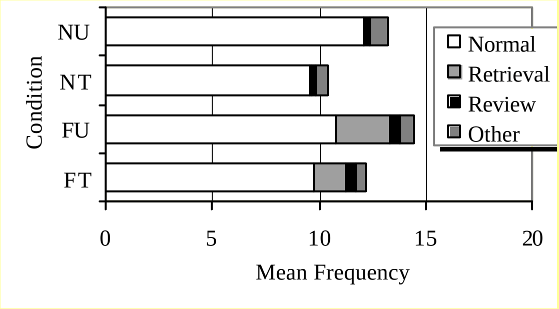


Figure 1: Mean numbers of transactions per trial by type and experimental condition (N = No Feedback; F = Feedback; T = Timed; U = Untimed).

3.2. Words

Word counts included whole and part-words. Again results show less speech with time-pressure (224 words/trial on average) than without (319): ($F_1(1,23) = 33.69$, $p < .001$). There was a non-significant tendency for speakers to resist the effect of time-pressure more with feedback (FT: 238 words/trial; FU: 316) than without (NT: 209; NU: 320): ($F_1(1,23) = 3.31$ $p = .082$).

3.3. Disfluencies

Disfluencies were first labeled according to the system devised by Lickley [18]: as repetitions, insertions, substitutions or deletions. The disfluency coder used Entropic/Xwaves software to listen, view and label disfluent regions of speech. Spectrograms were analyzed whenever necessary. Each word within a disfluent utterance was labeled as belonging to the onset, reparandum, repair, or continuation [17].

Because disfluencies are more common in longer utterances [3, 14, 21], raw disfluency counts may reflect only opportunities for disfluency. To provide a measure of disfluency rate, we divided the number of disfluencies in a monologue by its total number of fluent words, that is by the total number of words less the words in reparanda.

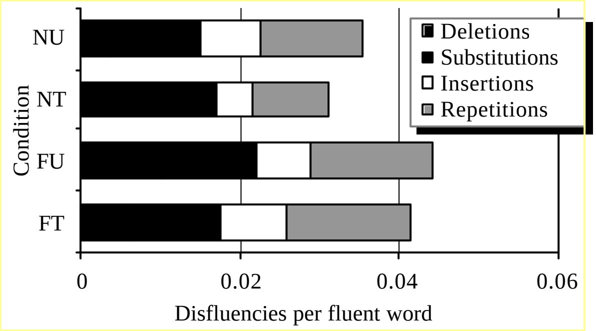


Figure 2: Rates of disfluency by type and experimental condition

The data in Figure 2 display a pattern which would be predicted from an strategic model of disfluency: Speakers were more disfluent in conditions with feedback (0.044) than in conditions without feedback (0.034), ($F_1(1,23) = 8.66$, $p = .007$), but were unaffected by time pressure ($F_1(1,23) = 1.87$, $p = .185$) or by any interaction ($F_1(1,23) < 1$). Because transaction-initial utterances are prone to disfluency, the effects were recalculated with number of transactions in the trial as a covariate. Again, only feedback affected disfluency ($F_1(1,22) = 11.33$, $p < .003$).

3.4. Disfluency Type

Figure 2 also displays the breakdown of disfluencies by type across experimental conditions. Only the rate of deletions showed any significant effect of feedback: an increase in the feedback conditions (.008) over no feedback (.004): ($F_1(1,23) = 14.61$, $p = .001$; $F_1(1,22) = 14.24$, $p = .001$ with transactions as covariate). There was no overall effect of time pressure on deletion ($F_1(1,23) = 2.44$ $p > .10$), though there was a non-significant tendency ($F_1(1,23) = 3.59$, $p = .071$; $F_1(1,22) = 3.62$, $p = .070$ with transactions as covariate) towards the 'disfluency as luxury' pattern: deletions tended to be more common in Feedback Untimed (0.010) than in Feedback Timed (0.007) trials, with no corresponding effect of time pressure in the No Feedback conditions (0.004 in both cases). No other type of disfluency and no combination of other types showed significant effects, though the rate of all non-deletion disfluencies was numerically higher (0.035) with feedback than without (0.030) ($F_1(1,23) = 3.21$, $p = .086$).

4. Discussion and Conclusions

The literature provided us with two major proposals for the causes of disfluency. One suggests that interlocutors intentionally employ disfluencies to warn each other of local difficulty. An interactive situation should encourage more disfluency, and if the signal function is critical, it should be maintained or even increase as the speaker's difficulties are augmented with increasing time pressure. An alternative view suggests that disfluency is an accident of heightened cognitive burden. If so, time pressure should promote disfluency particularly when feedback complicates the speaker's task. A third prediction stresses the fragility of listener-centric behaviour. If disfluency is listener-centric and all such behaviour is at best an option available to speakers when time or attention permit, disfluencies should be more frequent when speakers are not under time pressure but are interacting with listeners.

The experiment reported above successfully manipulated the interactive quality of the speaker's task and the pressure to complete it efficiently. Feedback in the form of a visual representation of a presumptive listener's gaze changed speakers' strategic treatment of the route communication task. A novel type of transaction, provides circumstantial evidence that subjects took seriously the task of tracking and redirecting their listener's gaze when it appeared to have strayed off-course. Retrievals were almost exclusive to the Feedback trials. Time pressure affected how much subjects said, with fewer transactions and fewer words under the one-minute limit.

With the manipulations effective in altering speakers' behaviour, we can return to the predictions for disfluency rate.

At first glance, disfluency seems to operate as an important strategic tool, with higher rates in the conditions with feedback and no effect of time-pressure. Yet, when disfluencies are subdivided by type, only deletion disfluencies were significantly more common in feedback trials. This fact is not just a result of sparse data in certain disfluency sub-types. Taken together, all the other kinds of disfluency still failed to respond robustly to feedback. Deletions alone support the strategic view.

Subject 10. Feedback Untimed	
Start	*Utterance*
70.4340	ehm go around and do a big circle ehm like just do a big loop down, **not**
71.4250	oh sorry there was
72.1388	<breath
72.2730	two stone creeks
72.4504	breath>
75.1890	ehm so yeah you're in the right place
Subject 19. Feedback Timed	
Start	*Utterance*
55.6070	and then you take a right across the farmed land
56.4686	< breath
56.7157	breath>
57.8160	**doing a s-**
58.8550	no you go right right at the farmed land

Figure 3: Deletion examples. Deletion disfluency in boldface.

It cannot yet be said that they support it conclusively. First, there was a nearly significant interaction of the type which would be predicted if disfluency were a luxury: disfluency rates were highest in the untimed feedback trials rather than in the timed, where there ought to have been more problems to report. Though we are unable to conclude definitively that deletions result from some optional rhetorical strategy, their content invites further investigation.

The examples in Figure 3 are typical. Subject 10 appears to be abandoning an utterance because he encountered difficulties in reading the map, and resumed with more accurate instructions. His deletion marks 'Giver failure'. Subject 19, on the other hand, interrupts the flow of speech and begins anew because the feedback gaze square did not move in the correct direction. This is an instance of 'Follower failure': the 'Follower's' action appears to have induced the subject to abandon an instruction which the Follower was in no position to obey.

Though deletions are indicators of interaction, it would be difficult to see them as signalling commitment to an utterance, as is thought to be the case for repetitions [14]. Instead, by abandoning an utterance, the speaker is expressing either the inadequacy of his/her own description or inappropriacy of the Follower's response. Whether the two functions are equally likely in both timing conditions we do not yet know.

It is plain, however, that both of these actions would require visual attention beyond what is needed for tracking the route to the next landmark and describing it. Our preliminary analyses of the eye-tracking data captured during these trials indicate that subjects' gaze primarily at the landmarks which are critical to the route [7]. The operations which appear to underlie deletions would produce two different patterns of off-route speaker gaze: scanning the map in the case of Giver failures and monitoring the feedback square's location in the case of Follower failures. If digressions are more common with feedback than without, and if they predominantly track the feedback square, then we may have a visual substrate for Follower failure deletions. If digressions are more common in untimed trials than in timed, then time to acquire the knowledge which underlies any deletion may be the real luxury afforded by our paradigm. Exactly how such a luxury is used – for better scanning of the map or tracking of the interlocutor, we do not yet know. At present, we are examining Giver gaze data to determine which patterns accompany disfluency.

5. Acknowledgements

The authors would like to thank Maria Luisa Flecha-Garcia and Yiya Chen who devised and administered the transcription and dialogue coding systems. This work was supported by EPSRC Research Grant GR/R59038/01 to Ellen Gurman Bard and and GR/R59021/01 to Anne H. Anderson.

6. References

[1] Anderson, Anne H., Miles Bader, Ellen Gurman Bard, Gwyneth Doherty, Simon Garrod, Steve Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Cathy Sotillo, Henry S. Thompson, and Regina Weinert, 1991. The HCRC Map Task Corpus. *Language and Speech***,** vol. 34, pp. 352–366.

[2] Anderson, Anne H., Ellen Gurman Bard, Cathy Sotillo, Alison Newlands & Gwyneth Doherty-Sneddon, 1997. Limited visual control of the intelligibility of speech in face-to-face dialogue. *Perception and Psychophysics*, vol. 59(4), pp. 580–592.

[3] Bard, Ellen Gurman, Anne H. Anderson, Cathy Sotillo, Matthew Aylett, Gwyneth Doherty-Sneddon & Alison Newlands. 2000. Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language*, vol. 42, pp. 1–22.

[4] Bard, Ellen Gurman, Matthew Aylett & Matthew Bull. 2000. More than a stately sance: Dialogue as a Reaction Time experiment. *Proceedings of the Society for Text and Discourse.*

[5] Bard, Ellen Gurman & Matthew Aylett, 2001. Referential Form, Word duration, and Modelling the Listener in Spoken Dialogue. *Proceedings of the 23rd Annual Conference of the Cognitive Science Society.*

[6] Bard, Ellen Gurman, Matthew Aylett & Robin Lickley,2002. Towards a Psycholinguistics of dialogue: Defining Reaction time and Error Rate in a Dialogue Corpus. *EDILOG 2002. Proceedings of the 6th workshop on the semantics and pragmatics of dialogue.* Edinburgh: The University of Edinburgh.

[7] Bard, Ellen Gurman, Anne H. Anderson, Marisa Flecha-Garcia, David Kenicer, Jim Mullin, Hannele B.M. Nicholson, Lucy Smallwood & Yiya Chen, 2003. Controlling Structure and Attention in Dialogue: The Interlocutor vs. the Clock. *Proceedings of ESCOP, 2003,* Granada, Spain.

[8] Barr, Dale J. & Boaz Keysar, 2002. Anchoring comprehension in linguistic precedents. *Journal of Memory and Language*, vol. 46, pp. 391–418.

[9] Brennan, Susan. & Herbert H. Clark, 1996. Conceptual Pacts and Lexical choice in Conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, vol. 22(6), pp. 1482–1493.

[10] Brown, P. & Gary S. Dell, 1987. Adapting production to comprehension – the explicit mention of instruments, *Cognitive Psychology*, vol 19, pp. 441–472.

[11] Brown, Gillian, Anne H. Anderson, George Yule, Richard Shillcock, 1983. *Teaching Talk*. Cambridge: Cambridge University Press.

[12] Carletta, Jean, Amy Isard, Steve Isard, Jacqueline Kowtko, Gwyneth Doherty-Sneddon, and Anne H. Anderson, 1997. The reliability of dialogue structure coding scheme. *Computational Linguistics*, vol. 23, pp. 13–31.

[13] Clark, Herbert H. and Catherine. Marshall, 1981. Definite reference and mutual knowledge. In Aravind K. Joshi, Bonnie L. Webber, and Ivan A. Sag (eds.), *Elements of discourse understanding*. Cambridge: Cambridge University. Press.

[14] Clark, Herbert H. & Thomas Wasow, 1998. Repeating words in Spontaneous Speech. *Cognitive Psychology,* vol. 37, pp. 201–242.

[15] Horton, W. & Boaz Keysar, 1996. When do speakers take into account common ground? *Cognition,* vol. 59, pp. 91–117.

[16] Keysar, Boaz, 1997. Unconfounding common ground. *Discourse Processes*, vol. 24, pp. 253–270

[17] Levelt, Willem J.M., 1989. Monitoring and self-repair in speech, *Cognition*, vol. 14, pp. 14–104.

[18] Lickley, Robin J. 1998. HCRC Disfluency Coding Manual *HCRC Technical Report* 100. `http://www.ling.ed.ac.uk/~robin/maptask/disfluency-coding.html`

[19] Oviatt, Sharon, 1995. Predicting disfluencies during human-computer interaction. *Computer Speech and Language,* vol. 9, pp. 19–35.

[20] Pickering, Martin & Simon Garrod, in press, Towards a mechanistic theory of dialogue: The interactive alignment model. *Behavioral & Brain Sciences*.

[21] Plauché, Madelaine & Elizabeth Shriberg, 1999. Data-Driven Subclassification of Disfluent Repetitions Based on Prosodic Features. *Proceedings of the International Congress of Phonetic Sciences,* vol. 2, pp. 1513–1516, San Francisco.

Disfluency & Behaviour in Dialogue: Evidence from Eye-Gaze

Hannele Nicholson[1], *Ellen Gurman Bard*[1], *Robin Lickley*[2],
Anne H. Anderson[3], *Catriona Havard*[3] *& Yiya Chen*[1]

[1] University of Edinburgh, Edinburgh, Scotland
[2] Queen Margaret University College, Edinburgh, Scotland
[3] University of Glasgow, Glasgow, Scotland

Abstract

Previous research on disfluency types has focused on their distinct cognitive causes, prosodic patterns, or effects on the listener [9, 12, 17, 21]. This paper seeks to add to this taxonomy by providing a psycholinguistic account of the dialogue and gaze behaviour speakers engage in when they make certain types of disfluency. Dialogues came from a version of the Map Task, [2, 4], in which 36 normal adult speakers each participated in six dialogues across which feedback modality and time-pressure were counter-balanced. In this paper, we ask whether disfluency, both generally and type-specifically, was associated with speaker attention to the listener. We show that certain disfluency types can be linked to particular dialogue goals, depending on whether the speaker had attended to listener feedback. The results shed light on the general cognitive causes of disfluency and suggest that it will be possible to predict the types of disfluency which will accompany particular behaviours.

1. Introduction

Types of disfluency distinguished by their form are also distinguishable by other characteristics. Repetition disfluencies are the most common in spontaneous speech [21]. In a pioneering paper, Maclay & Osgood showed that repetitions precede content words more often than function words [22]. Repetitions have been linked to strategic signalling commitment to both listener and utterance [10, 12]. The prosodic cues for repetitions are linked to certain strategies in dialogue [25]. Savova showed, however, that the prosodic cues to repetitions differ from the cues to a substitution, providing support for the notion that disfluency types have distinct sources in the cognitive processes underlying the production of speech in dialogue [26].

It is already clear that disfluencies of different types cause different processing problems for the listener. While repetitions cause less disruption than false starts [a kind of deletion disfluency] for a word recognition task, [13], repetitions are more difficult for trained transcribers to detect than false starts of the same length [20].

Disfluency has been linked to cognitive causes by Levelt [17], who proposes that some disfluencies occur for covert cognitive reasons while other disfluencies are overt corrections. Lickley found that disfluency types vary systematically across turn types whereby turns that involve planning typically involve more self-corrections than utterances which are responses to queries [18]. Replies to queries, on the other hand, tend to involve more filled pauses (ums, uhs) and repetitions in order to buy time [18]. Thus, it seems that certain types of disfluencies have already been linked to certain dialogue behaviours.

More recently, psycholinguistic studies of a speaker's eye-gaze at a visual array have revealed that speakers look at objects involved in the process of speech perception and production. [15, 28]. Speakers who made a speech error when performing a simple object naming task had spent just as long gazing at the object as they did when they named it fluently. Apparently, then, disfluency did not result from either long or hasty examination of the object to be named. Disfluency does not appear to be a measure of perceptual problems per se.

Instead, disfluency is related to the cognitive burdens of production [5]. We will use disfluency to discover whether there is a cognitive cost involved in taking up information needed to pursue a dialogue task. We will then show that this cost is put to good use: the locations of disfluencies reveal that they are appropriate responses to the information that speakers have garnered.

The information in question underpins what is thought to be a crucial task in dialogue: each participant must maintain a model of her interlocutors' knowledge so as to adjust to their mutual knowledge both what she says and how she says it. Most views of dialogue now assume that speakers will take some interest in indications both of the listener's knowledge about the domain under discussion and of the listener's satisfaction with the communication just made. Clark and Krych [9], for example, propose that speakers monitor listeners' faces for all manner of feedback, much as they track listeners' utterances. Horton and Gerrig [16] acknowledge the costs of this operation, suggesting that complete uptake and application of listener information could prove to be taxing in some cases, so that utterances will be less perfectly designed for the audience as the cognitive burden increases.

To determine whether garnering cues to listener knowledge is indeed costly to production, we use a variant of the map task [2, 7]. As in the original task, players have before them versions of a cartoon map representing a novel imaginary location. The Instruction Giver communicates to the Instruction Follower a route pre-printed on the Giver's map. The present experiment manipulates time-pressure and the modality or modalities in which a distant confederate delivers pre-scripted feedback to the speaker's instructions. Verbal feedback affirms comprehension of some instructions and declares general incomprehension of others. Visual feedback, in the form of a simulated listener-eyetrack projected onto the map, may correctly go to the named map landmark or wrongly advance to another. Where both modalities are used, their feedback may be concordant or discordant across modalities. Scripted and simulated responses are used to control the conditions under which speakers are operating. Genuine speaker eye-gaze is tracked.

We use eyetracks, rather than sight of the speaker's direction of gaze, to represent listener feedback for two reasons. First, simulated gaze is much easier to control than genuine gaze on the part of the confederate. Second, though facial expressions and direction of gaze have real value, tasks with a visual component produce remarkably little inter-interlocutor gaze [[1,3,11]]. To allow simultaneous performance of the task and uptake of listener information, the listener's 'eyetrack' was superimposed on the map (See Figures 1 and 2).

The present paper will examine two kinds of disfluency diistinguished by previous research, repetitions and deletions. In the current definition, a repetition is produced when the speaker repeats verbatim one or more words with no additions, deletions, or re-ordering, as in (1)

(1) Now you want to ***go*** **go** just past the tree

Repetitions are thus a single faulty attempt at communicating the same message in the same form. In contrast, a deletion has occurred whent the speaker interrupts an utterance without restarting or substituting syntactically similar elements, as in (2)

(2) A MOVE 36 You need to be just under…
A MOVE 37 Do you have a White Mountain?

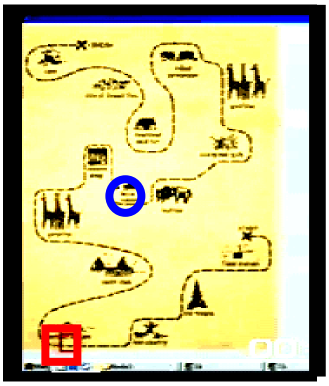
Thus, deletions abandon one communicative act in favour of another.

In this setting, there seem to be two distinguishable predictions. Clark and Krych [9] predict good uptake of all visual cues to listener knowledge and suitable application of the information. Horton and Gerrig [16] predict that the more complex the input, the more difficult will be both uptake of cues and the production of suitable speech. Thus there should in principal be an increase in dsfluency if speakers observe negative visual feedback ('follower gaze' at wrong landmarks) and if there ar conflicts between verbal and visual feedback.

1.1. Task and procedure

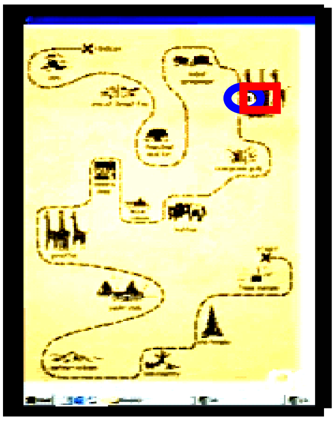
All the materials come from an experiment which used conversations between subject Instruction Givers and a confederate Instruction Follower. Each subject was greeted individually with the confederate. Each subject was naïve to the status of the confederate and during post-experimental debriefing, none reported any suspicions. Both subject and confederate were told that whoever took the role of Instruction Giver should guide the Instruction Follower, from a marked start-point to buried treasure. Subject and confederate then 'negotiated' that the subject would be Giver and the two were taken to separate rooms. The Giver was seated 60 cm from a flat screen monitor displaying the map. Labelled landmarks and map designs were adapted from the HCRC Map Task Corpus [2]. Eye tracking movements were recorded using a non-invasive Senso-Motor Instruments remote eye-tracking device placed on a table below the monitor. Eye movements were captured with Iview version 2 software. The tracker was re-calibrated at the beginning of each trial. Speech was recorded in mono using Asden HS35s headphone- microphone combination headsets. Video signals from the eye tracker and the participant monitor were combined and recorded in Mpeg with Broadway Pro version 4.0 software.

Feedback from the confederate took two forms. Visual feedback consisted of a simulated eyetrack, a small red square advancing from landmark to landmark once each landmark was named, and showing saccades of random length and direction. The visual feedback was under the control of the experimenter, who advanced the feedback square to its next programmed position when the Giver first mentioned a new route-critical a landmark. When feedback was scheduled to be wrong, the square moved to a landmark that had not been named. When feedback was to be correct, the feedback square advanced to the landmark just named. Similarly, verbal feedback came from the confederate subject who read pre-scripted responses. Just as with the visual feedback, the confederate provided verbal feedback when the speaker uttered the first mention of the landmark in question. Figures 1 and 2 illustrate possible events.



Instruction Follower: 'Yes, got it.'

Figure 1. Discordant feedback. Circle = Giver's gaze; Square = Follower's feedback (wrong location).



Instruction Follower: 'Okay, that's fine'

Figure 2. Concordant feedback. Circle = Giver's gaze; Square = Follower's feedback (correct location).

1.2. Experimental Design

The experiment crossed feedback modality (3), single modality group (2), and time-pressure (2). In the *No Feedback* conditions, subjects saw only the map. In the *Single-Modality* condition, subjects in the Verbal Group got verbal feedback only, while those in the Visual Group had only visual feedback. Finally, in the *Dual-Modality* condition, all subjects received both visual and verbal feedback. The two modalities might be discordant or concordant. *Concordan*t feedback consisted on average of 8 instances of positive verbal and correct visual feedback, and 6 instances of negative verbal and wrong visual feedback per map. In each map, *discordant* feedback included roughly 3 instances of negative verbal and correct visual feedback, and 6 instances of positive verbal and wrong visual feedback. This design is portrayed in Table 1. In half of the trials, speakers under *time-pressure* had three minutes to complete the task; in *untimed* dialogues there was no time limit.

Table 1. The relationship between the Experimental Groups and the various Feedback Modalities.

Experiment	Feedback Modalities		
	None	Single	Dual
Verbal Group	None	**Verbal**	Verbal + Visual
Visual Group	None	**Visual**	Verbal + Visual

Thirty-six subjects with normal uncorrected vision were recruited from the Glasgow University community. All were paid for their time. All encountered all 6 conditions. Six different basic maps were used, counter-balanced across conditions over the whole design. Subjects were eliminated if any single map trial failed to meet criteria for feedback or capture quality. The feedback criterion demanded that the experimenter advance the feedback square between the introduction of the pertinent landmark and the onset of the following instruction in all cases where the feedback was scheduled to be errant and in 70% where the square's movement was scheduled to be correct. The capture criterion demanded that at least 80% of the eye-tracking data was intact. Subjects were also eliminated if on debriefing they revealed any suspicions about the nature of the interlocutor.

2. Results

2.1. Baseline effects: Words

Since the opportunities for disfluency increase with increasing amount of speech, it is important to note effects of the experiment's design on word counts. Word counts for whole and part-words show less speech with time-pressure (425 words/trial on average) than without (579): (F_1(1,34) = 24.38, p < .001). Visual Group Single-Modality trials (459 words) were shorter than the corresponding Dual-Modality trials (590 words) with no corresponding change for Verbal subjects (Feedback Modality x Group: (F_1(2,68) = 8.65 p < .001; Bonferroni: t = -6.4, p < *.001*). Since Dual-Modality Conditions do not differ between groups (Verbal: 616, Visual: 590), we can use this condition to examine the relationships between disfluency and gaze or dialogue events.

We also examined speech rate across the experimental conditions. To calculate speech rate we divided the Giver words per map by the total Giver speaking time for the map (the summed durations of all conversational moves less the summed durations of both simple and filled pauses). Time-pressure had no significant effect on speech rate. The interaction between Feedback Modality and Group (F_1(2,68) = 4.87, p < .02) presented in Table 2, is due only to a difference between the No-Feedback (.34) and Dual-Modality (.30) conditions for the Verbal Group (Bonferroni p = .004). Again Dual Modality conditions are alike.

Table 2. Speech rate (Words/Total speaking time) means from Feedback Modality x Group interaction

Experiment	Feedback Modalities		
	None	Single	Dual
Verbal Group	.340	.303	.304
Visual Group	.344	.343	.340

2.2. Baseline effects: Gaze

In order to test for the relationship between disfluency and Giver gaze, it was necessary to determine whether all conditions in which a Giver might gaze at a feedback square actually did succeed in directing the Giver's attention to the square. To check for overlap of gaze between Giver and 'Follower', the video record of feedback and Giver Gaze were analyzed frame by frame for the landmark at which each was directed. When Follower Gaze and Giver Gaze were on the same landmark, the Giver was considered to be looking at the feedback square. Here we report the number of feedback episodes [task sub-portions containing in feedback] in which *any* frame contained an instance of gaze at the feedback square].

Givers did not make use of all their opportunities by any means (Figure 3). Nor did they use their opportunities equally (Visual feedback x Verbal feedback: F_1(1,34) = 7.70, p < .01). Strangely enough, Givers used fewest opportunities in an important concordant condition, the one in which the Follower was clearly lost: the Follower square was hovering over a wrong landmark while the Follower was simultaneously providing negative verbal feedback (verbal- vis-: .366). These attracted less gaze than another concordant condition – when the Follower needed no help because she was in the right place and said so (verbal+ vis+: .511). Similarly Givers looked less when the Follower was lost but claimed not to be (verbal+ vis-: .448) than when she was correct but claimed to be lost (verbal- vis+:.591) (Bonferroni t-tests at .008). A simple description says that speakers are most likely to track listeners, the listener's location falls under their own gaze, which is occupied by the things they are describing. Apparently, spekaers prefer not to go off-route to learn the whereabouts of an errant follower.

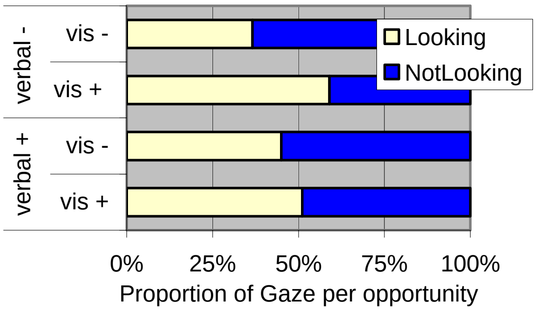


FIGURE. 3 Proportion of feedback episodes attracting speaker gaze to feedback square: Effects of combinations of visual and verbal feedback in dual channel conditions

2.3. Disfluencies Overall

The first author labeled disfluencies according to the system devised by Lickley [19] as repetitions, insertions, substitutions or deletions. She used Entropic/Xwaves software to listen to, view and label disfluent regions of speech. Spectrograms were analyzed whenever necessary. Each word within a disfluent utterance was labeled as belonging to the reparandum, the interregnum, or the repair. A reparandum involves speech that is either overwritten, expunged or retraced in the repair [19]. Repairs typically 'replace' the error in the reparandum. Since deletions are typically abandoned utterances, they have no repair [19, 27].

Because disfluencies are more common in longer utterances [6, 10, 25] we divided the number of disfluencies in a monologue by its total number of words, yielding disfluency rate as a dependent variable.

Disfluency rates were submitted to a by-subjects ANOVA for Group (2) (Verbal vs. Visual), Time-pressure (2) (timed vs. untimed) and Feedback Modality (3) (none, Single-Modality, Dual-Modality). The baseline No-Feedback conditions differed between Verbal and Visual groups (Group * Modality: F_2(2,68) = 5.21, p < .01; Bonferroni, t = 2.94, p < .02). This difference can be explained by a single subject in the Verbal Group who was an outlier in terms of disfluency. Because of this subject, there was no effect of Feedback Modality within the Verbal Group, while the Visual Group showed the expected increase in rate of disfluency between No Feedback and Single- (Bonferroni t = -4.12, p = .001) or Dual-Modality conditions (Bonferroni t = -5.77, p < .001). Since Single and Dual Modality conditions did not differ, we can proceed to examine only the Dual Modality conditions in the expectation that conflicting feedback (only found in Dual Modality) *per se* is not an overall cause of disfluency.

2.4. Disfluency Types: Repetitions v Deletions

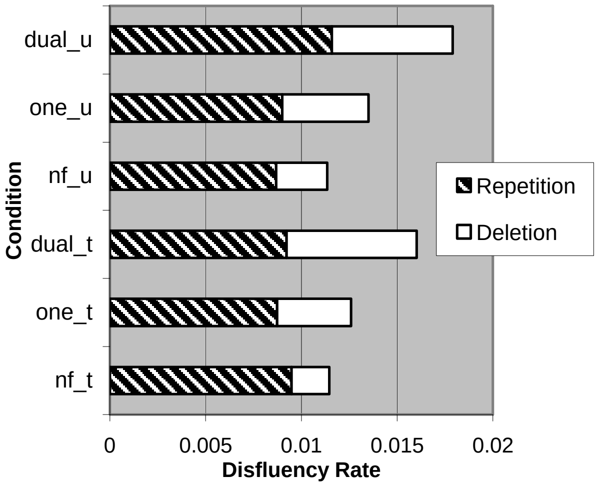


Figure 4. Rates of disfluency by type and experimental condition for the Verbal and Visual Groups combined. nf = no feedback, one = Single-Modality feedback, dual = Dual modality feedback; t = timed, u = untimed.

An initial investigation of deletions and repetitions begins to separate them. Figure 4 displays their distributions across experimental conditions. Independent analyses were done for each type of disfluency; that is one analysis within deletions only and one within repetitions only.

As found in [23], only deletion rate showed any significant effect of feedback: Deletion rate rose significantly with each additional feedback modality (No Feedback .002, Single-Modality .004, Dual-Modality .007; $F_1(2,68) = 21.00$, $p < .001$; all Bonferroni t-values < .01). There were no effects of time-pressure on deletion rate and no significant interactions.

For repetitions on the other hand, an interaction between Time-pressure and Group ($F(1,34) = 6.27$, $p < .02$) revealed that subjects were more disfluent in the untimed condition (.012) of the Verbal Group than they were anywhere else in either the Verbal or the Visual Group, timed or untimed, though the internal comparisons were not significant.

3.5 Disfluency & Eye-Gaze

Within the Dual-Modality condition, the experimental design contrasted positive and negative feedback in the two modalities. However, the modalities are concordant or discordant only if the Giver actually takes up both visual and verbal feedback. The tendency for more speech in conditions with verbal feedback suggests that subjects were attending to what the confederate Follower said. Eye-tracking enabled us to tell when the Giver had actually looked at the Follower's visual feedback. As Figure 3 made plain, Givers do not take up the same proportion of concordant and discordant feedback. They gazed most at one kind of discordant feedback (negative verbal + correct visual) and least at a concordant condition (negative + wrong visual feedback).

To look for disfluency in truly vs potentially concordant and discordant situations, we examined disfluency per feedback opportunites in concordant and discordant situations contrasting those in which Givers did or did not look at Follower feedback. In fact, Givers who attended to discordant feedback from the Follower encountered subsequent fluency problems. The number of disfluencies per feedback opportunity was greatest following a discordant feedback episode in which the Giver had actually gazed at the Follower feedback square (.333), a significantly higher rate than following a concordant feedback episode which had drawn the Giver's attention (.205) (Bonferroni $t = -3.51$, $p = .001$ within by-subjects Group (2) x Giver attention (looking v not looking) x Concordance of modalities (concordant v discordant: $F_1(1,34) = 7.24$, $p = .01$). None of the other pairwise comparisons was significant.

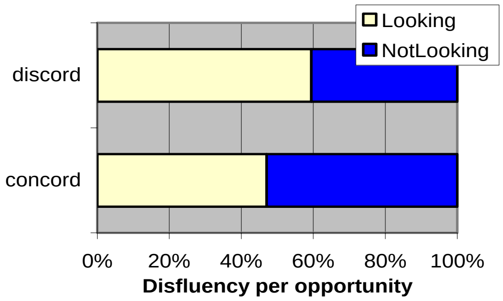


Figure 5. Rate of repair disfluencies per concordant or discordant feedback opportunity with respect to whether the Giver was either looking or not looking at the Follower. The difference is significant when the Giver looked at the Follower.

3.6 Disfluency Type, Gaze & Motivation

So far we have seen that speakers' gaze behaviour is not randomly distributed. It follows certainly problems (a Follower on-route who claims not to be) and ignores others (a Follower off-route who claims to be on-route). We have also seen that on those occasions when an instruction Giver actually takes in enough information to see what is amiss, he or she is more likely to speak disfluently. The question we ask here is whether these disfluencies are part of well formed communicative processes. If the information taken in by examination of listener feedback is properly processed by the speaker, what s/he says disfluently will be something appropriate to the situation. To determine whether this is really the case, it was necessary to classify utterances by their goal or motivation. To do this, the first author examined all 564 repetitions and 280 deletions occurring in the Dual Modality feedback condition.

The first stage of this process was to identify an interval for analysis. All dialogues were coded according to the HCRC Conversational-Game-Move coding scheme [8]. In this system, each turn is decomposable into conversational Moves, or sub-units of the dialogue. For example, a speaker might 'Instruct' by giving directions or 'Align' when noting that the Follower has gone astray. Analyses began with the Move that carried the disfluency. The coder searched backwards from the Interruption Point of the disfluency to the most recent Giver Move introducing a new landmark. The start time was considered to be the Giver's first mention of a new landmark while the end time was the Interruption point of the disfluency or for deletions, the end of the repair.

The second stage was to identify Giver gaze behaviours within these intervals. The gaze record of the speaker for this time-span was then checked and disfluency was coded as 'Looking' if there were any overlaps of Giver and Follower Gaze from the introduction of the landmark to the end of the disfluency. All others were coded 'Not Looking'.

Third, each disfluency was classified by Motivation, the content of the repair. Repetitions necessarily occur within the same dialogue Move, while deletions are almost always a single abandoned Move, so that the repair effectively lies in the next Move. Motivations were classified under two major goals: either the speaker was 'confirming' that the Follower was at a correct or incorrect landmark or the speaker was 'reformulating' by adding, elaborating, or correcting information being transmitted. Examples of goal and disfluency combinations are given in Table 3 below.

Table 3. Examples of disfluencies by goal and type. For repetitions, both reparandum and repair appear in bold text. For deletions, just the reparandum appears in bold text since the repair is effectively non-existent.

Disfluency Type	Dialogue Goal	
	Confirmation	Reformulation
Repetition	'**That's, That's** just fine	'Eh you travel directly ehm sort of **north…north** and east'
Deletion	'So loop around the waterfall **over**….Yeah, there'	'Um **can you si-** …it's to the left of that'

Since appropriate confirmation of position should depend on the Giver actually determining where the Follower was, we would expect confirmations to accompany gaze at the follower. Since the arrival of the Follower at the goal or her movement off route should complete the execution of a series of instructions, all the Giver need do is cease instructing and declare the Follower to be right or wrong. Accordingly, deletion disfluencies are appropriate: in this view they mark a sequence of instructing, checking, and, finally, abandoning any ongoing instruction for a new a phase in the dialogue.

Our second goal category, reformulation, can also repair communication problems but by elaborating the material serving the current goal. Typically [14], speakers have to look away from their interlocutors when formulating complex material. Also on the grounds of complexity, we might expect not looking and reformulating to accompany repetition disfluencies [10].

Analyses of Giver's Gaze (2: looking vs. not looking), Motivation (2: confirmation vs. reformulation), Disfluency Type (2: repetition vs. deletion) and Time-pressure (2: timed vs. untimed) showed part of this pattern.

We predicted that reformulations would attract repetition disfluencies and confirmations would attract deletions. As Figure 6 illustrates, numerically repetitions (confirmation = 0.083; reformulation = 0.403) and deletions (confirmation = 0.245; reformulation = 0.186) worked as predicted (F_1(1,34) = 59.60, p < .001). The predicted effect of Motivation, however, was significant only for repetitions (F_1 (1,34) = 124.17, p < .001).

We predicted that looking at the feedback square would yield confirmations and not looking would accompany reformulations. In fact, only when Givers did not gaze at the Follower's square was the prediction met: there was a higher rate of reformulations than confirmations (Gaze x Motivation: *F(1,34) = 9.27, p < .01*, Bonferroni *t* at *p* = .008.).

Since we have an association between reformulations and repetitions, and one just reported between reformulations and not looking at the interlocutor, we tested for the effects within repetitions and deletions separately. Though the Giver tended not to look at the Follower square during repetition disfluencies, the trend is weak because it appears to hold only in the Verbal Group (Disfluency Type x Gaze: *F(1,34) = 3.59, p = .067;* Gaze x Motivation x Experiment: *F(1,34) = 8.62, p < .006;* Bonferroni at *p* = .001). For deletion disfluencies, the effect of gaze depends on motivation: deletions classified as confirmations were, as we predicted, more common when the Giver took the opportunity to look at the Follower (Bonferroni at *p* = .008), whereas deletions classed as reformulations showed an insignificant tendency to be more common when the Giver was not looking at the Follower (Motivation x Gaze: *F(1,34) = 8.61, p < .01).* Thus, there were associations between disfluency type and motivation type and between disfluency-motivation combination and gaze.

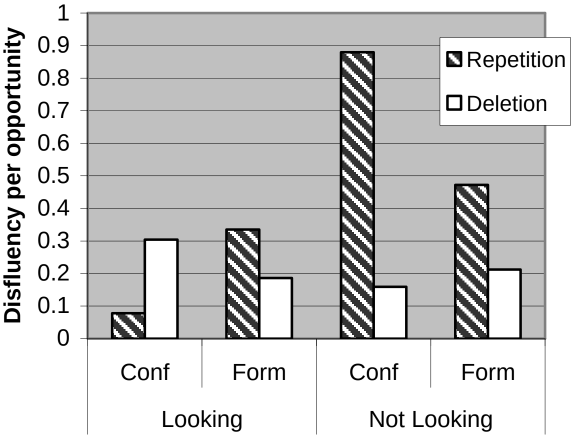


Figure 6. Rates of Repetitions and Deletions per opportunity with respect to Behaviour type, either confirmation (Conf) or reformulation (Form) and Gaze. The difference is significant for Repetititions but not for Deletions.

3. Discussion and Conclusions

Although the visual feedback provided the Giver with the Follower's exact location at any point during the interaction, this information had a cost. The Giver tended to gaze away from the Follower's location. Gaze aversion during difficulty is a common phenomenon found in conversational analysis and gaze studies [14, 15], and we find that gaze itself makes for production difficulty: speakers are more disfluent if they look at the follower feedback. Furthermore, Givers tended not to look at concordant negative feedback which clearly indicated trouble, though they did look at discordant feedback when the Follower was easily found – on the landmark being described.

When a Giver noticed this discordance, disfluency often occurred as result, presumably because the speaker was burdened with resolving the conflicting verbal and visual signals and in a sense handling the Follower's confusion. Disfluency, it seems, tend to co-occur first with uptake of the speaker's whereabouts and misalignment in dialogue, as predicted in [24]

If speakers are committed to tracking and accommodating listeners' knowledge [9, 10], and if repetitions indicate commitment to listener and message, Givers should visually attend to their Followers whilst making a repair: a committed speaker might be expected to assist a Follower who is clearly in difficulty by looking at the Follower's feedback and tailoring any following utterances to them. Instead, repetitions tended to associate with reformulation and thus by reformulation to gaze aversion during critical need. Looking at the follower instead accompanied deletions, as the Giver abandoned a Move in order to confirm or deny the listener's progress. Thus, it seems deletions, or false starts were associated with attending to the Follower but not with commitment to the utterance.

The present paper has added a psycholinguistic and dialogue perspective to the taxonomy of disfluency. We found that speakers are disfluent in different ways depending upon the dialogue task in which they are currently engaged. The nature of listener feedback and the Giver's uptake of information about the listener both had effects.

4. Acknowledgements

The authors thank Yiya Chen and Catriona Havard for running the experiment and administering the gaze and transcription coding. This work was supported by EPSRC Research Grant GR/R59038/01 to Ellen Gurman Bard and GR/R59021/01 to Anne H. Anderson.

5. References

[1] Anderson, Anne H., Ellen Gurman Bard, Cathy Sotillo, Alison Newlands, & Gwyneth Doherty-Sneddon. 1997. Limited Visual Control of the Intelligibility of Speech in Face-to-Face Dialogue. *Perception and Psychophysics,* 59 (4), pp. 580-592.

[2] Anderson, Anne H., Miles Bader, Ellen Gurman Bard, Gwyneth Doherty, Simon Garrod, Steve Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Cathy Sotillo, Henry S. Thompson, & Regina Weinert, 1991. The HCRC Map Task Corpus. *Language and Speech***,** vol. 34, pp. 352–366.

[3] Argyle, Michael & R Ingham. 1972. Gaze, mutual gaze and proximity. *Semiotica*, 6. pp. 289-304.

[4] Bard, Ellen Gurman, Anne H. Anderson, Marisa Flecha-Garcia, David Kenicer, Jim Mullin, Hannele Nicholson, Lucy Smallwood & Yiya Chen, 2003. Controlling Structure and Attention in Dialogue: The Interlocutor vs. the Clock. *Proceedings of ESCOP, 2003*, Granada, Spain.

[5] Bard, Ellen Gurman, Robin J. Lickley, & Matthew P. Aylett. 2001. Is Disfluency just Difficulty? *Proceedings of DiSS'01*, Edinburgh.

[6] Bard, Ellen Gurman, Anne H. Anderson, Cathy Sotillo, Matthew Aylett, Gwyneth Doherty-Sneddon & Alison Newlands. 2000. Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language*, vol. 42, pp. 1–22.

[7] Brown, Gillian, Anne H. Anderson, George Yule, Richard Shillcock, 1983. *Teaching Talk*. Cambridge: Cambridge University Press.

[8] Carletta, Jean, Amy Isard, Steve Isard, Jacqueline Kowtko, Gwyneth Doherty-Sneddon, & Anne H. Anderson, 1997. The reliability of dialogue structure coding scheme. *Computational Linguistics*, vol. 23, pp. 13–31.

[9] Clark, Herbert H. & Meredyth A Krych. 2004. Speaking while monitoring addresses for understanding. *Journal of Memory and Language.* vol. 50, Issue 1, pp. 62-81.

[10] Clark, Herbert H. & Thomas Wasow, 1998. Repeating words in Spontaneous Speech. *Cognitive Psychology*, vol. 37, pp. 201–242.

[11] Exline, Ralph V., P. Jones, & K. Maciorowski. 1977. *Race, affiliation-conflict theory and mutual vision attention during conversation.* Paper presented at the meeting of the American Psychological Association.

[12] Fox Tree, Jean & Clark, Herbert H.. 1997. Pronouncing 'the' as 'thee' to signal problems in speaking. *Cognition. 62. pp. 151-167*

[13] Fox Tree, Jean. 1995. The effects of false-starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory & Language.* 34. pp. 709-738.

[14] Glenberg, Arthur M, Jennifer L. Schroeder & David A. Robertson. 1998. Averting the gaze disengages the environment and facilitates remembering. *Memory and Cognition.* Vol. 26, (4). pp. 651-658

[15] Griffin, Zenzi M., 2005. The Eyes are right when the Mouth is Wrong. *Psychological Science*. Vol 15, number 12, pp. 814-821

[16] Horton, William S. & Richard J. Gerrig. 2005. The impact of memory demands on audience design during language production. *Cognition*, vol. 96. pp. 127-142.

[17] Levelt, Willem J.M., 1989. Monitoring and self-repair in speech, *Cognition*, vol. 14, pp. 14–104.

[18] Lickley, Robin J. 2001. Dialogue Moves and Disfluency Rates. *Proceedings of DiSS '01, ISCA Tutorial and Workshop*, University of Edinburgh, Scotland, UK, pp. 93-96.

[19] Lickley, Robin J. 1998. HCRC Disfluency Coding Manual *HCRC Technical Report* 100. `http://www.ling.ed.ac.uk/~robin/maptask/disfluency-coding.html`

[20] Lickely, Robin J. 1995. Missing Disfluencies. *Proceedings of ICPhS,* Stockholm, vol. 4. pp. 192-195.

[21] Lickley, Robin J. 1994. *Detecting Disfluency in Spontaneous Speech.* PhD. Thesis, University of Edinburgh.

[22] Maclay, Howard & Charles E. Osgood. 1959. Hesitation phenomena in spontaneous English speech. *Word,* 15, pp. 19-44.

[23] Nicholson, Hannele, Ellen Gurman Bard, Robin Lickley, Anne H. Anderson, Jim Mullin, David Kenicer & Lucy Smallwood, 2003. The Intentionality of Disfluency: Findings from Feedback and Timing. *Proc. Of DiSS'03, Gothenburg Papers in Theoretical Linguistics 89. pp.15-18*

[24] Pickering, Martin & Simon Garrod, 2004, Towards a mechanistic theory of dialogue: The interactive alignment model. *Behavioral & Brain Sciences*. 27 (2), pp. 169-190.

[25] Plauché, Madelaine & Elizabeth Shriberg, 1999. Data-Driven Subclassification of Disfluent Repetitions Based on Prosodic Features. *Proceedings of the International Congress of Phonetic Sciences,* vol. 2, pp. 1513–1516, San Francisco.

[26] Savova, Guergana & Joan Bachenko. 2002. Prosodic features of four types of disfluencies. *Proceedings of DiSS'03.* Gothenburg University, Sweden. pp. 91-94.

[27] Shriberg, Elizabeth. 1994. Preliminaries to a Theory of Speech Disfluencies. PhD Thesis. University of California at Berkeley.

[28] Tanenhaus, Michael K., Michael J. Spivey-Knowlton, Kathleen M. Eberhard, Julie Sedivy. 2000. Integration of visual and linguistic information in spoken language comprehension. *Science.* 268, pp. 1632-1634.

APPENDIX AA. – INSTRUCTION SHEETS AND CONSENT FORM FROM THE MONITOR PROJECT

INSTRUCTIONS GIVEN TO SUBJECTS IN EXPERIMENT 1

Instructions:

Your map was drawn by an explorer in order to provide a route to some treasure buried at the finish point. Your task is to explain to another person (in a separate audio-linked room) as accurately as possible the route shown on your map. The other person has a similar map, but with no start point, finish point, or route drawn on it. They will draw the route on their map using with respect to your instructions. The two maps were drawn by different explorers, thus some of the landmarks on the map may differ slightly.

The task will be repeated on four different maps. On some maps, you will be provided with an indicator showing you where the other person is looking on their map. On two of the maps you will have a time limit of ***one minute*** to complete your instructions. You will be given a 30 second warning such that you can gauge how long you are taking. On the other maps, there will be no time limit.

You are free to terminate the experiment at any stage, and have your data destroyed if you feel in any way uncomfortable. Simply advise the experimenter that you wish to do so.

Should you have any questions, please ask the experimenter before the session starts.

INSTRUCTIONS GIVEN TO SUBJECTS IN EXPERIMENT 2

Instructions:

Your map was drawn by an explorer in order to provide a route to some treasure buried at the finish point. Your task is to explain to another person (in a separate audio-linked room) as accurately as possible the route shown on your map. The other person has a similar map, but with no start point, finish point, or route drawn on it. They will draw the route on their map using with respect to your instructions. The two maps were drawn by different explorers, thus some of the landmarks on the map may differ slightly.

The task will be repeated on six different maps. On some maps, you will be provided with an indicator showing you where the other person is looking on their map. On three of the maps you will have a time limit of ***two minutes*** to complete your instructions. You will be given a one-minute warning such that you can gauge how long you are taking. On the other maps, there will be no time limit.
On some of the maps, you will be provided with a two-way audio link, such that you can receive verbal feedback from the other person, but on the other maps, you will communicate only through a one-way audio link. The experimenter will advise you as to the conditions of each map before each trial starts.

You are free to terminate the experiment at any stage, and have your data destroyed if you feel in any way uncomfortable. Simply advise the experimenter that you wish to do so.

Should you have any questions, please ask the experimenter before the session starts.

INSTRUCTIONS GIVEN TO SUBJECTS IN EXPERIMENT 3

Instructions:

Your map was drawn by an explorer in order to provide a route to some treasure buried at the finish point. Your task is to explain to another person (in a separate audio-linked room) as accurately as possible the route shown on your map. The other person has a similar map, but with no start point, finish point, or route drawn on it. They will draw the route on their map using with respect to your instructions. The two maps were drawn by different explorers, thus some of the landmarks on the map may differ slightly.

The task will be repeated on three different maps. On some maps, you will be provided with an indicator showing you where the other person is looking on their map. You will also be able to converse with the other person on some maps.

In order to use these recordings, we need 'perfect' descriptions. This means that if you describe the route in such a way that the other person doesn't make any mistakes, we will double your money to £10 per hour.

The experimenter will advise you as to the conditions of each map before each trial starts.

You are free to terminate the experiment at any stage, and have your data destroyed if you feel in any way uncomfortable. Simply advise the experimenter that you wish to do so.

Should you have any questions, please ask the experimenter before the session starts.

Monitor Map Task Experiment

The experiment you are about to take part in will be audio recorded. During the experiment we will also be measuring your eye gaze. All data collected will be treated with confidentiality and your anonymity will be maintained at all times. Please sign the consent form below to say that you are aware of this, and that you understand that you may leave the experiment at any time if you are not entirely comfortable.

CONSENT FORM

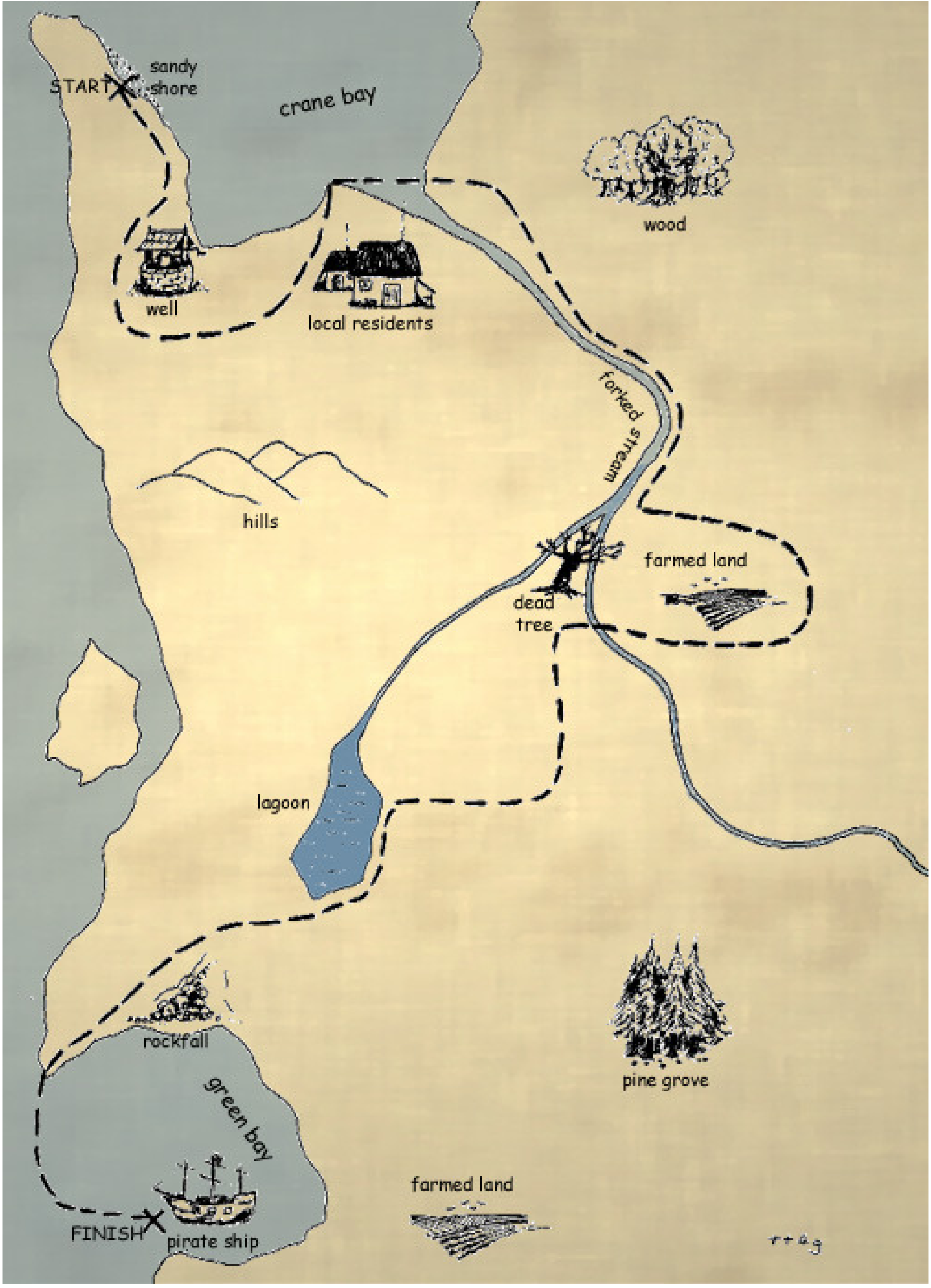
I (please print name)______________________________________give consent to take part in the Monitor Map Task Experiment as described to me above.

Signed _______________________________

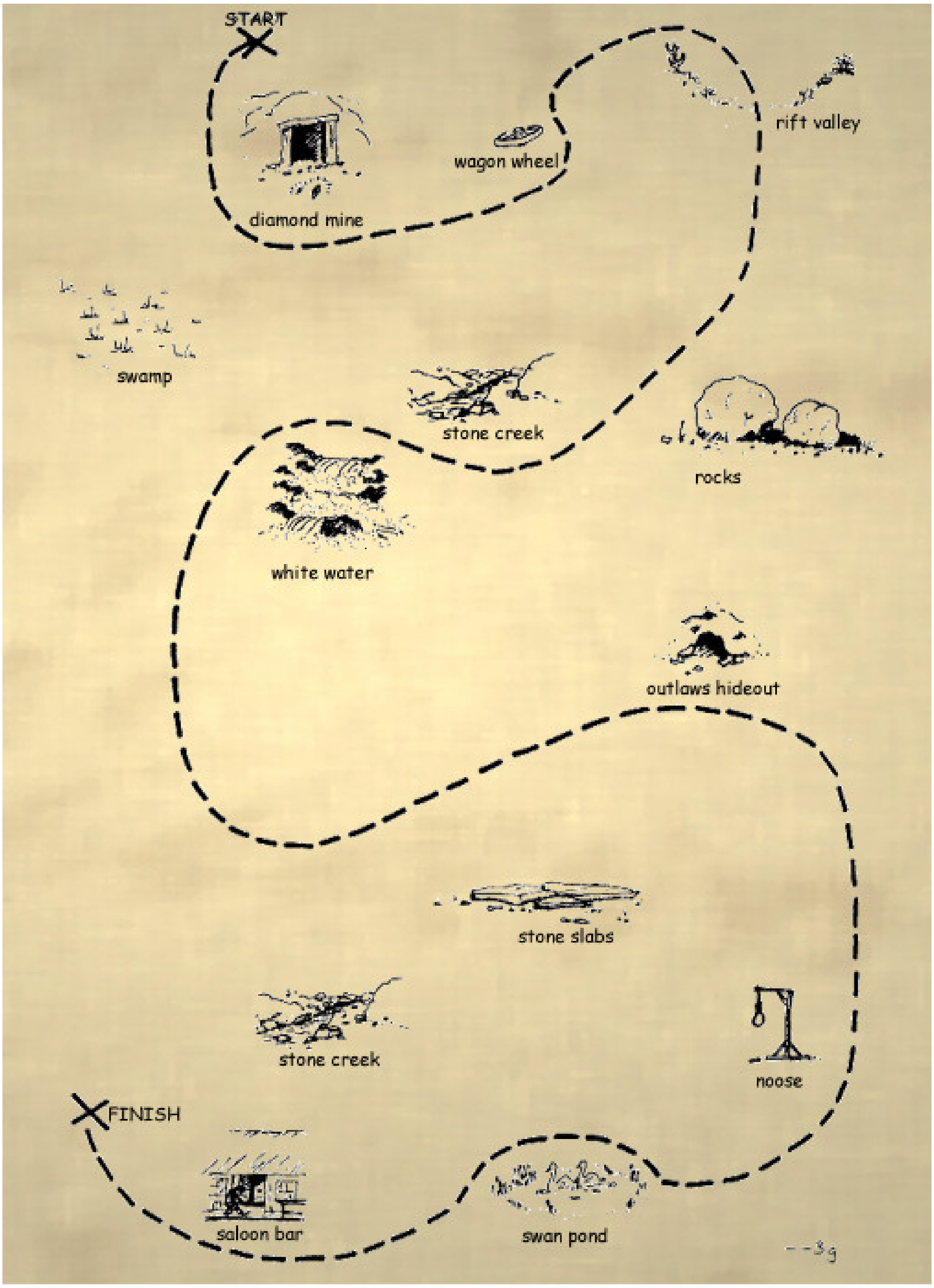
Date ______________________

APPENDIX B - MAPS IN EXPERIMENT 1

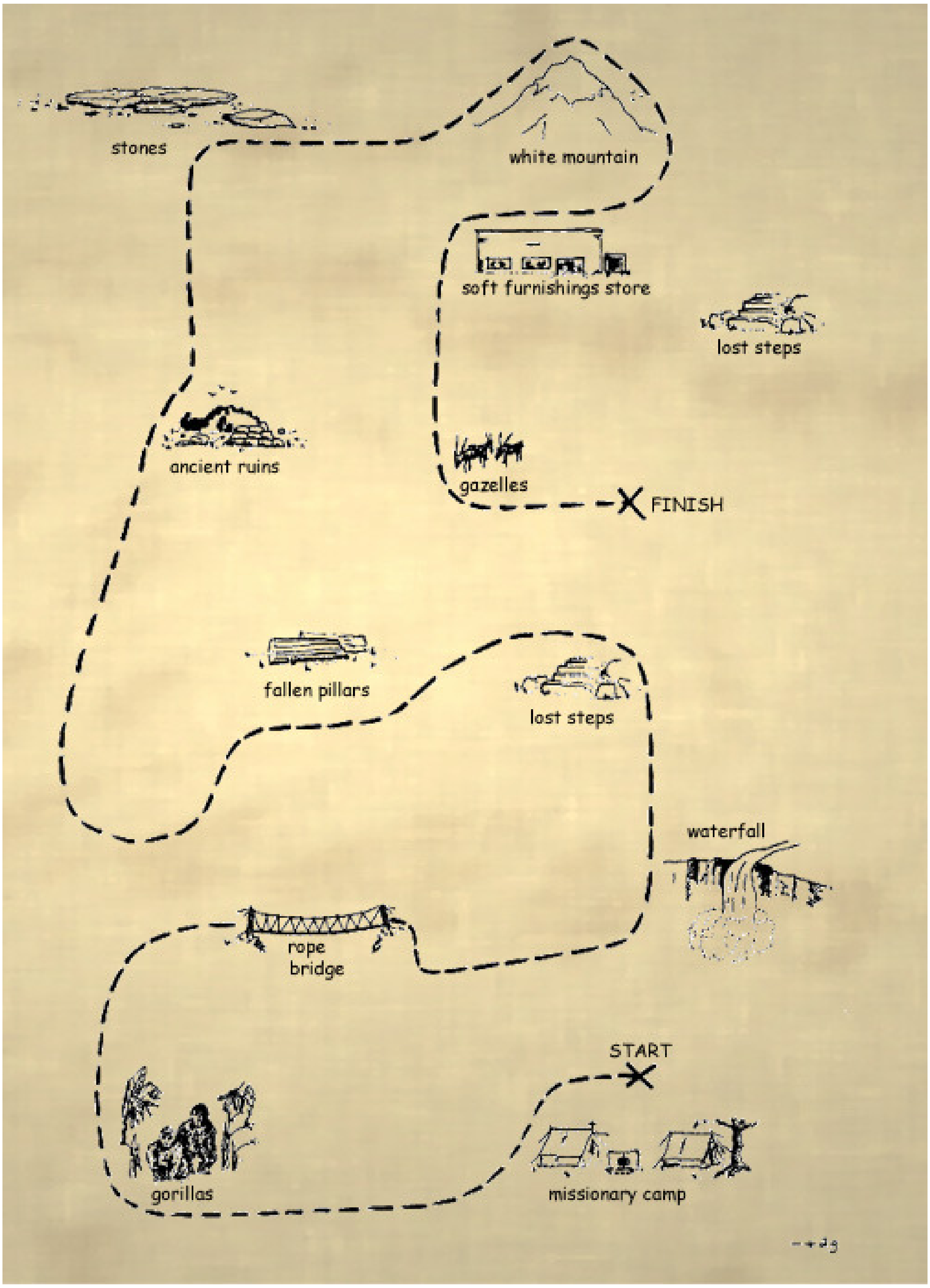
Crane Bay Map



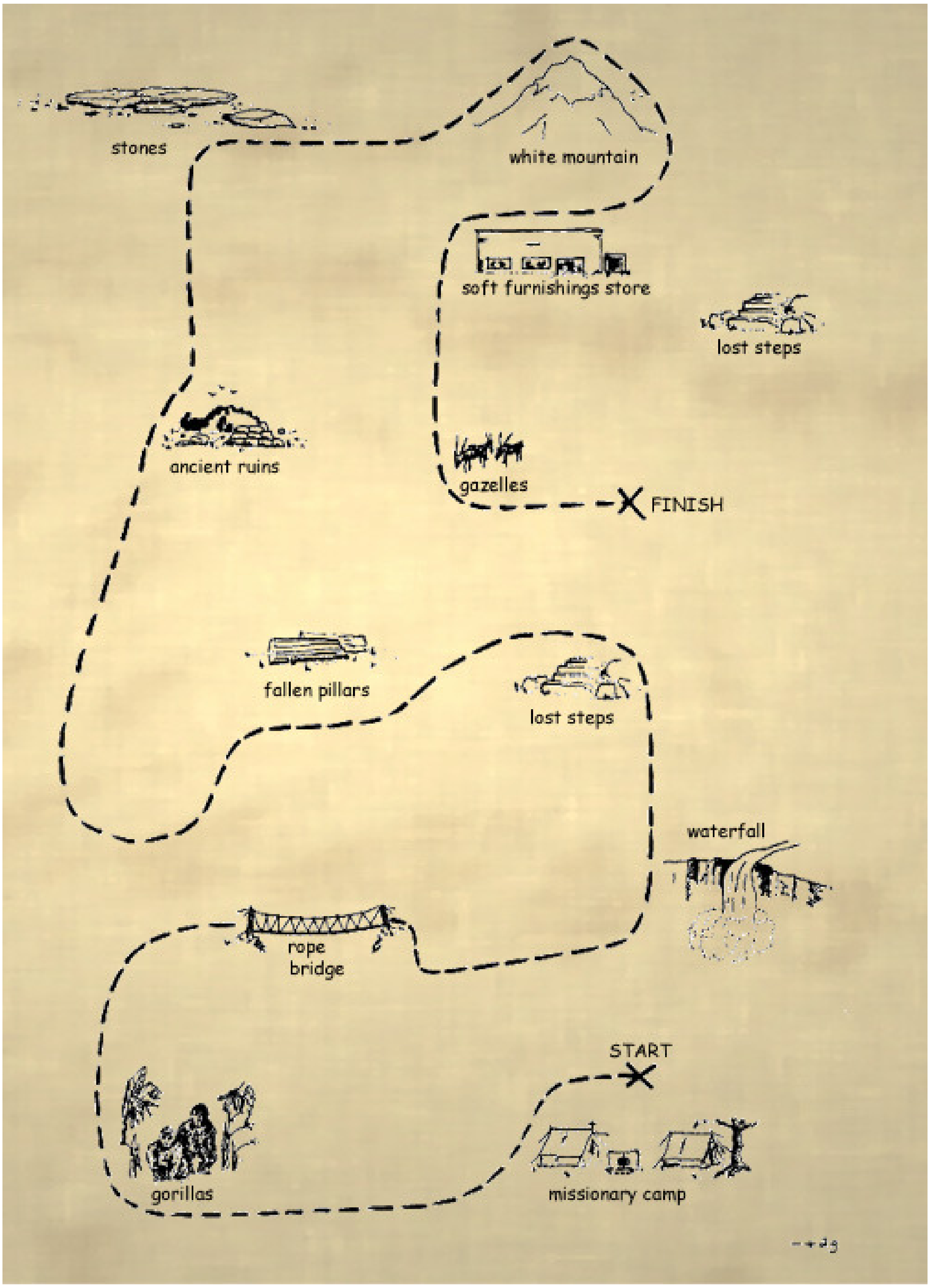
Diamond Mine Map



Mountain Map



Telephone Kiosk Map



APPENDIX D – EXAMPLES OF PLANNING AND HESITATION DELETIONS FOR EXPERIMENT 1

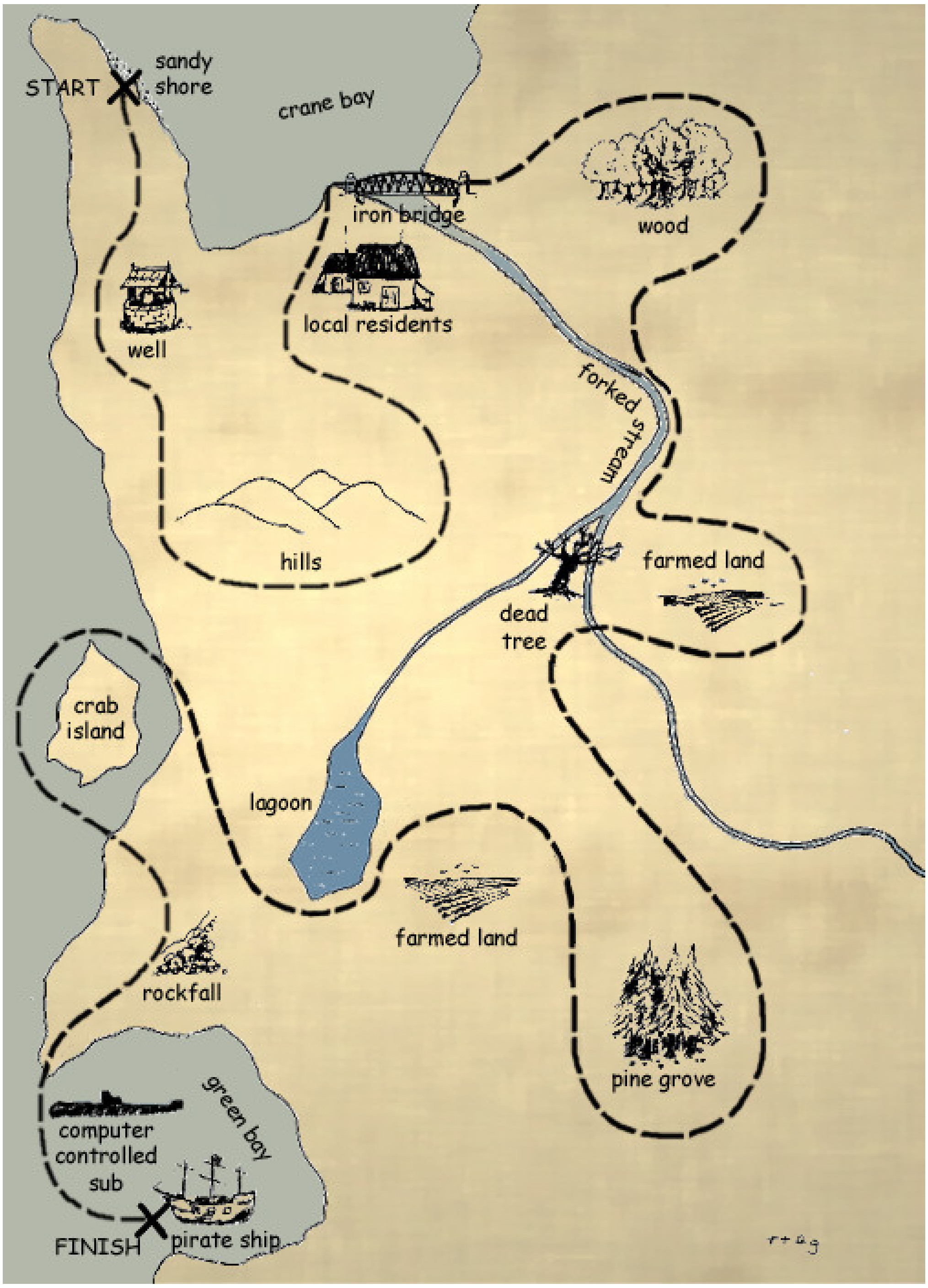
		Planning Deletions
Speaker	**Map**	**Transcription**
s19	Crane Bay	if you go to the well, if you look ... that's it yeah...that's that's the start
s19	Crane Bay	doing a s ... no you go right right at the farmed land
s5	Crane Bay	Ehm down the bottom bi- … you look like you're looking in the wrong place
s10	Diamond Mine	and then go...head...no, not right around the diamond mine
s11	Diamond Mine	Now you want to kee ... yeah anticlockwise round it
s3	Diamond Mine	The outlaws hideout and then ... yep ... go there
s10	Mountain	Oh sorry the bot the…yeah that lost steps
s18	Mountain	that's right…no you were right before
s19	Mountain	there we go, that's the st ... that's exac ... <breath breath> oh no wait sorry you had it.
s9	Telephone Kiosk	towards the farmer's gate which is to the…yeah, that's right there
s7	Telephone Kiosk	the dead tree…and then…yeah
s2	Telephone Kiosk	and then come up…can you see the dead tree?

Hesitation Deletions

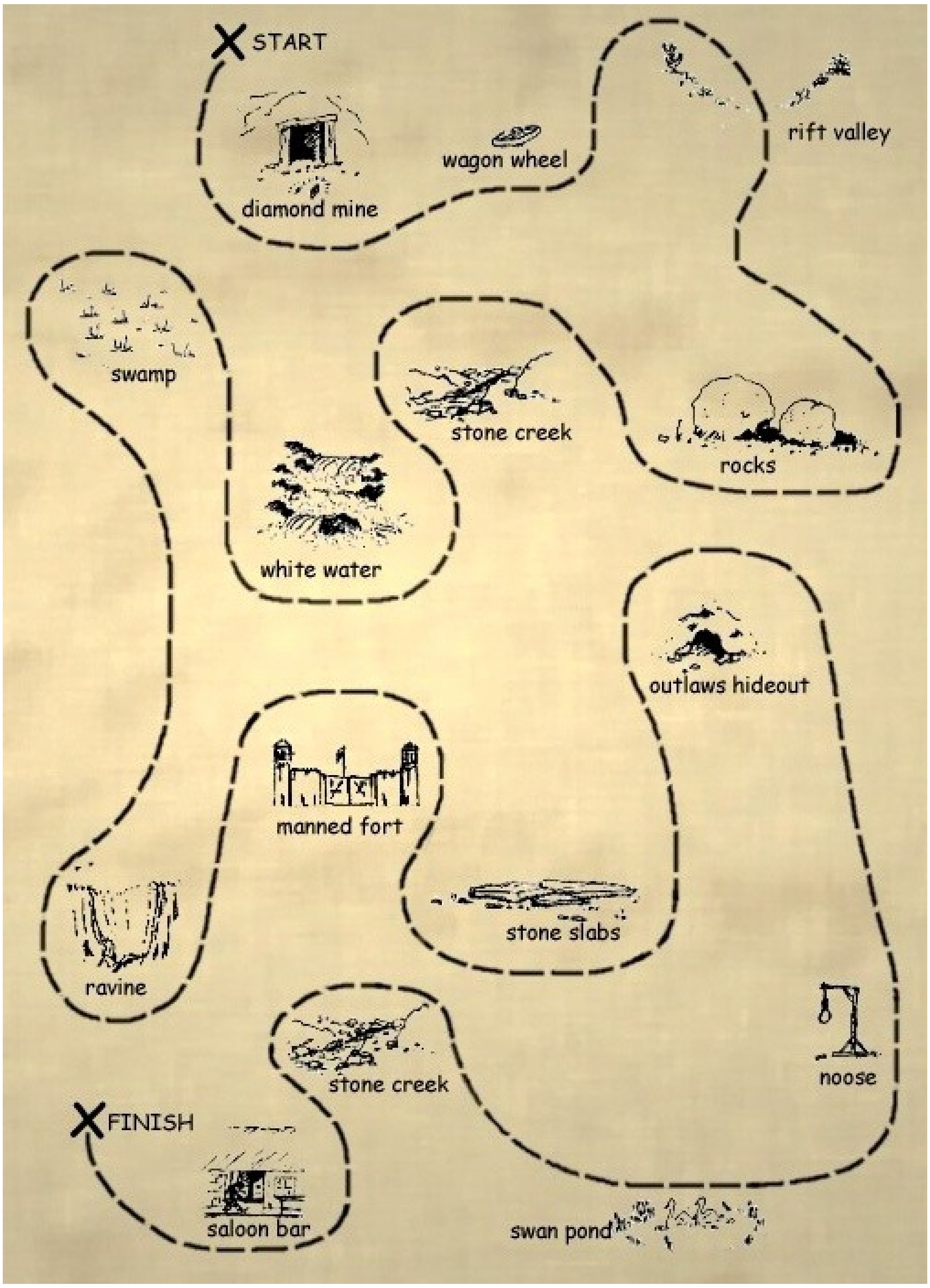
Speaker	**Map**	**Transcription**
s10	Crane Bay	then cross over the water <breath> ehm go ehm trace the line of like ... ehm like just follow the line ... follow the shore of
s5	Crane Bay	Ehm you go straight over the top uh well it's jus- sorry it's horizontal
s9	Crane Bay	So, when we get past…we we go above this farmed land
s10	Diamond Mine	the bottom stone creek t- go right round
s15	Diamond Mine	Right okay I- from the upper left-hand corner go
s15	Diamond Mine	If you can turn west you sh- uh ... there's a swan pond
s16	Mountain	keep...then turn up and go past ... have the ancient ruins
s12	Mountain	is it a little bit south but then w- curved up east past the waterfall
s10	Mountain	ehm go around and do a big circle ehm like just do a big loop down, not, oh sorry there was two stone creeks
s9	Telephone Kiosk	the picture in the right-hand si-…uh…to the…in the centre of the map
s7	Telephone Kiosk	And then to f- …and then just to the left of the great viewpoint
s2	Telephone Kiosk	and then go sort of diagonally sort of a slight eh slight ... slightly diagonally eh up to the right

APPENDIX E – MAPS USED IN EXPERIMENT 2 and EXPERIMENT 3

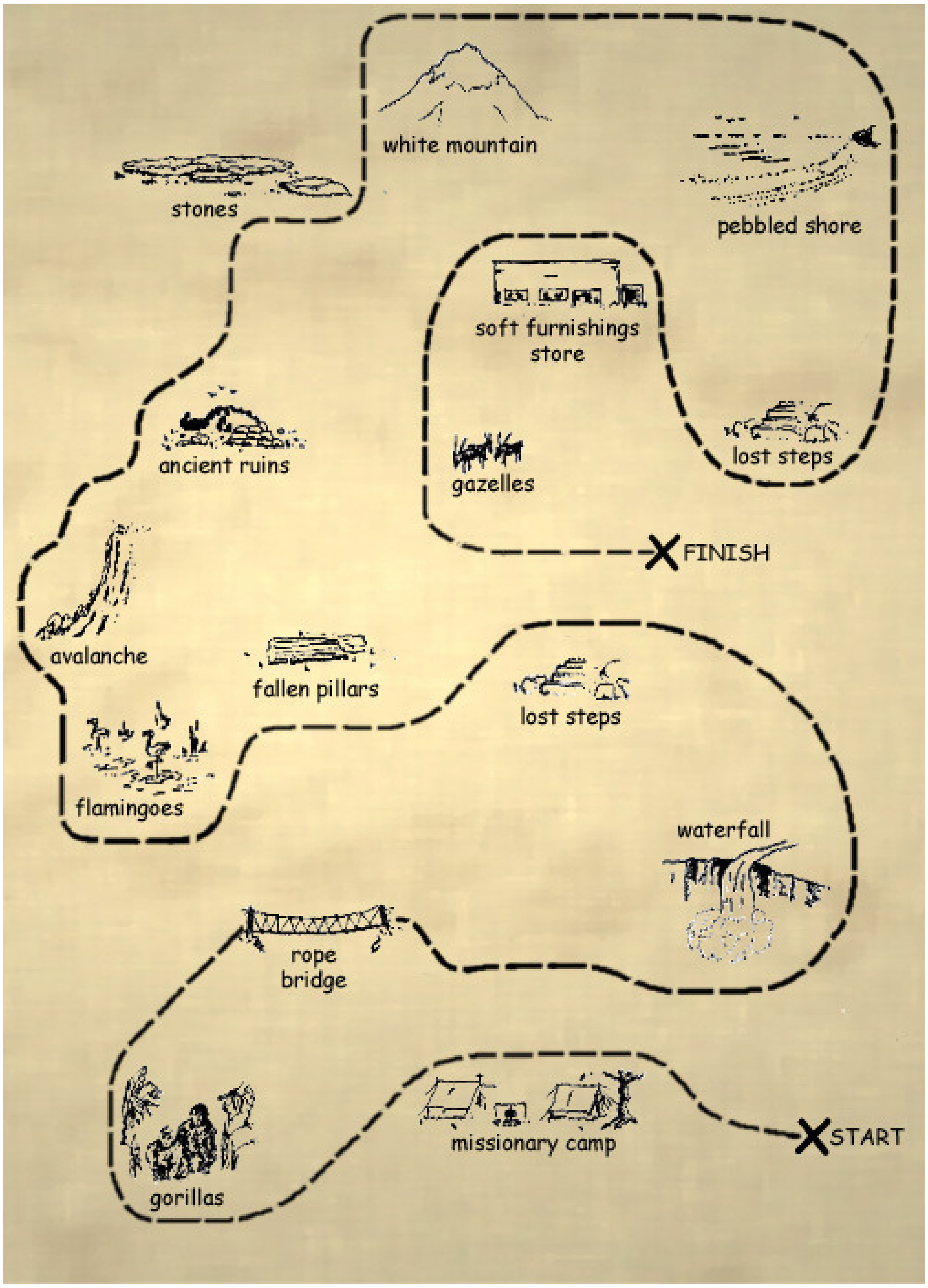
CRANE BAY MAP



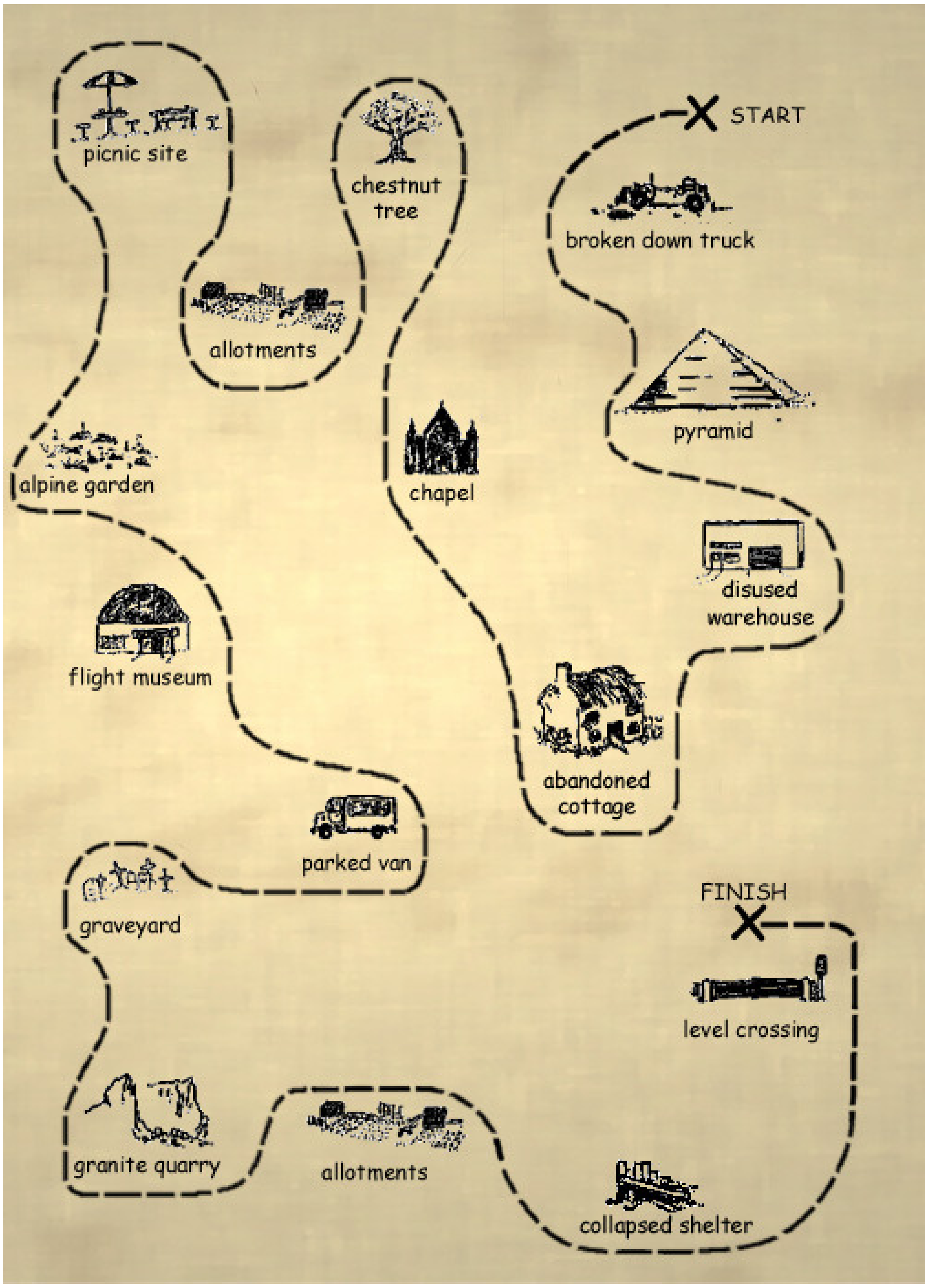
DIAMOND MINE MAP



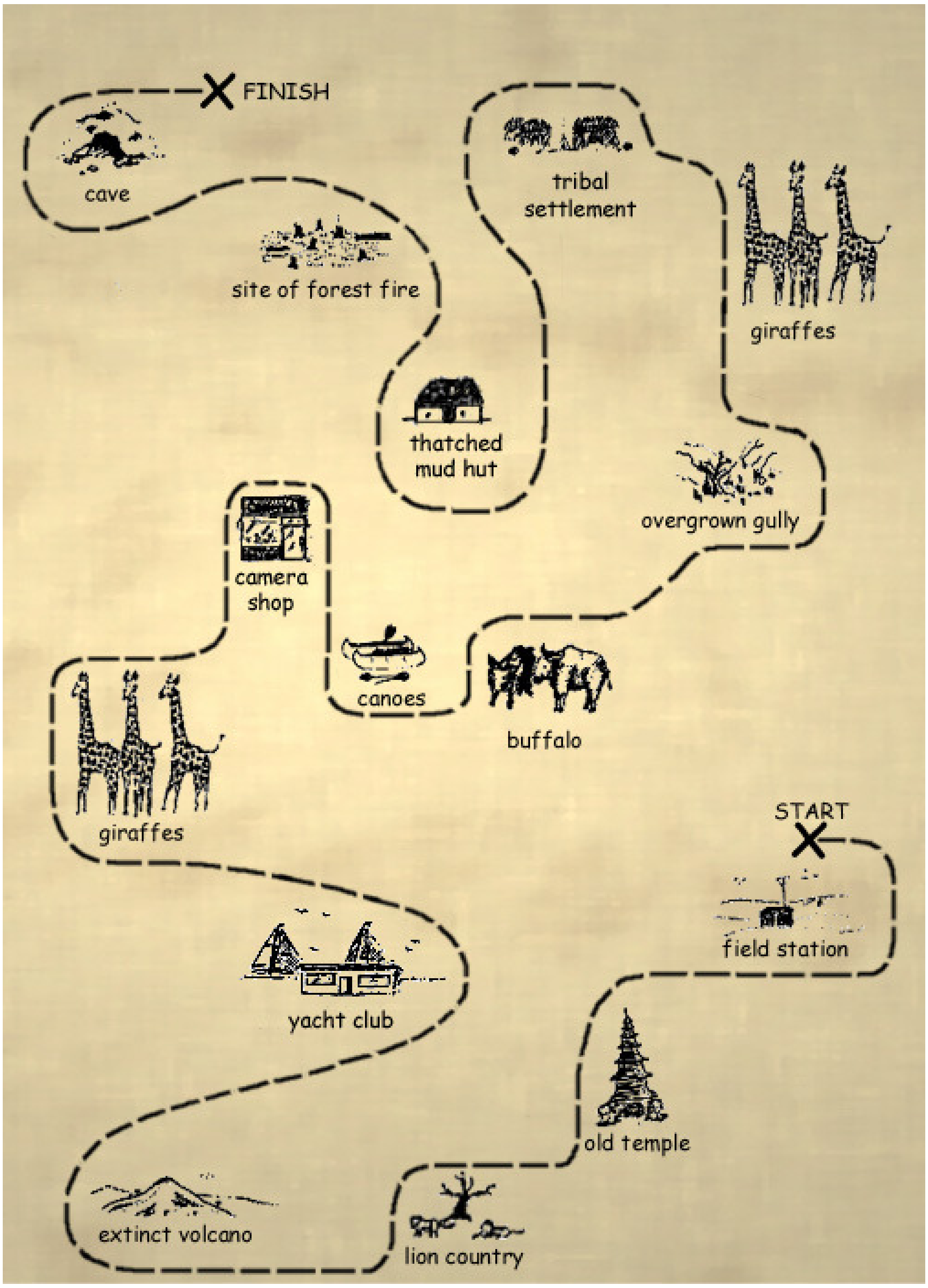
MOUNTAIN MAP



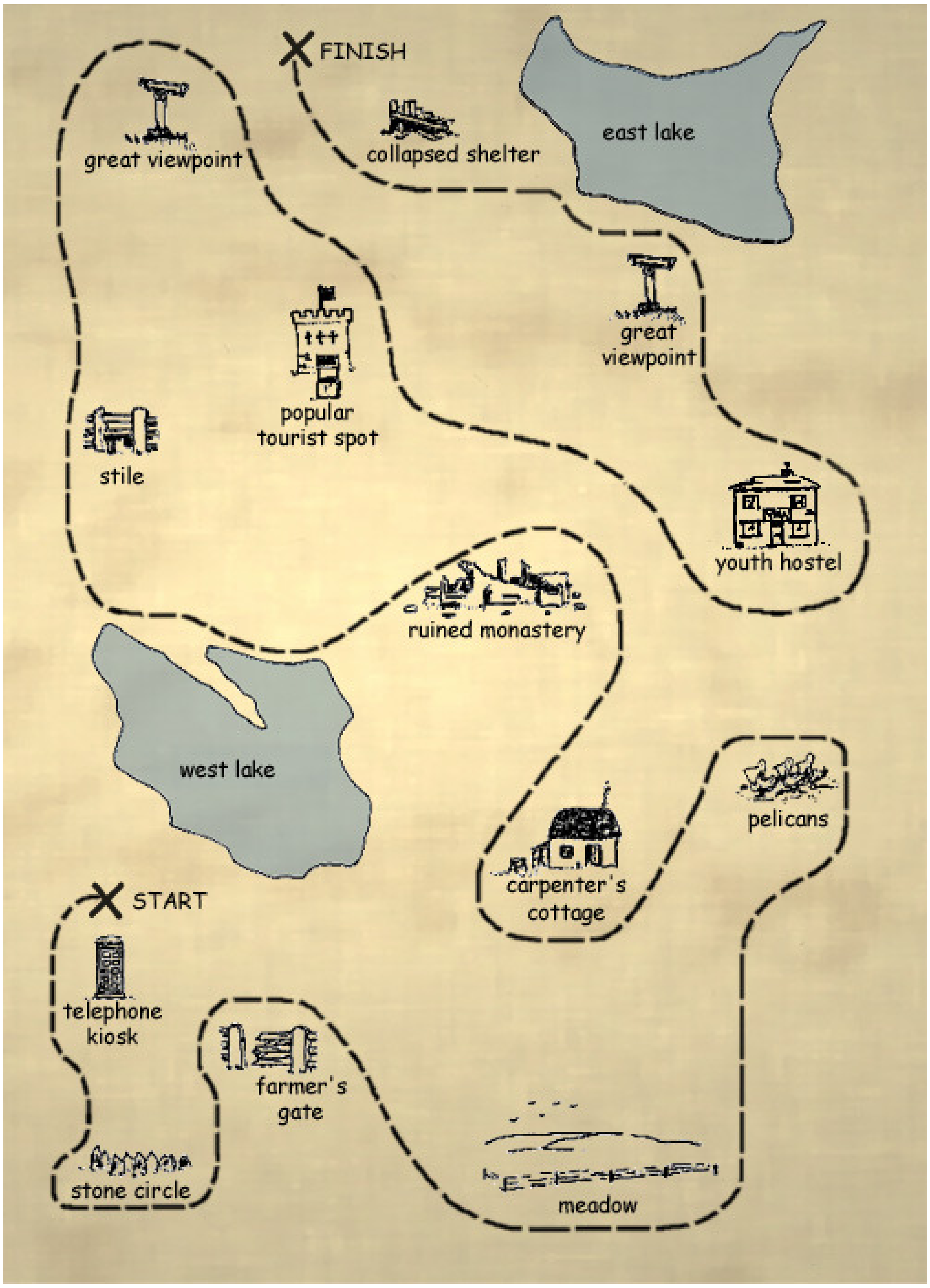
PYRAMID MAP


START
broken down truck
pyramid
disused warehouse
abandoned cottage
chapel
chestnut tree
allotments
picnic site
alpine garden
flight museum
parked van
graveyard
granite quarry
allotments
collapsed shelter
level crossing
FINISH

SAFARI MAP


FINISH
cave
tribal settlement
giraffes
site of forest fire
thatched mud hut
overgrown gully
camera shop
canoes
buffalo
giraffes
START
field station
yacht club
old temple
extinct volcano
lion country

TELEPHONE KIOSK MAP



APPENDIX F. – SCHEDULE OF VISUAL AND VERBAL FEEDBACK FOR EXPERIMENTS 2A, 2B and 3.

Map:- ***Crane Bay***

LM	**Verbal Response**	**Visual FB**
Start / Sandy Shore:	*Ok got that.*	Correct
Well:	*Ok, yes.*	Correct
Hills:	*Yep, fine*	Correct
Local Residents:	*Can't see it*	Correct
Iron Bridge:	I *don't see it*	Wrong
Wood:	*Okay, fine*	Correct
Forked Stream:	*Got it.*	Wrong
Farmed Land 1:	*Don't know where you mean.*	Wrong
Dead Tree:	*Okay, got it.*	Correct
Pine Grove:	*Ok, got that*	Wrong
Farmedland 2:	*Can't see it.*	Correct
Lagoon:	*Yep, got it.*	Wrong
Crab Island:	*Ok, I'm with you*	Correct
Rock Fall:	*No, not with you.*	Correct
CCSub[1]:	*Stop, where's that?*	Wrong
Pirate ship / Finish:	*Yes, ok.*	Correct

[1] CCSub = Computer Controlled Submarine

Map:- *Diamond*

LM	**Response**	**Visual FB**
Start:	*Ok got that.*	Correct
Diamond Mine:	*Ok, yes.*	Correct
Wagon Wheel:	*No, not with you.*	Correct
Rift Valley:	*Ok Got it.*	Correct
Rocks:	*Got it.*	Wrong
Stone Creek1:	*Don't know where you mean.*	Wrong
White Water:	*Yes, ok.*	Correct
Swamp:	*Ok, that's fine.*	Wrong
Ravine:	*Yes, ok.*	Correct
Manned Fort:	*Yes that's fine.*	Wrong
Stone Slabs:	*Don't know where that is.*	Correct
Outlaw's Hideout:	*Yep, got it.*	Correct
Noose:	*Stop! Where's that?.*	Wrong
Swan Pond:	*No, not with you.*	Correct
Stone Creek2	*Ok, got it.*	Correct
Saloon Bar:	*Don't know where you mean.*	Wrong
Finish:	*Right, got it.*	Correct

Map:- ***Mountain***

LM	**Response**	**Visual FB**
Start:	*Ok got that.*	Correct
Missionary Camp:	*Ok, yes.*	Correct
Gorillas:	*Yes, that's fine.*	Wrong
Rope Bridge:	*No, not with you.*	Correct
Waterfall:	*Got it.*	Correct
Lost Steps:	*Don't know where you mean.*	Wrong
Fallen Pillars:	*Ok, got it.*	Correct
Flamingos:	*Yes that's fine.*	Wrong
Avalanche	*Yep, ok*	Correct
Ancient ruins:	*No, not with you.*	Correct
Stones:	*Yep, got it.*	Wrong
White Mountain:	*Ok, got it.*	Correct
Pebbled Shore	*Stop! Where's that?*	Wrong
Lost Steps2:	*Yes, that's fine.*	Correct
Soft Furnishings Store:	*Don't know where you mean.*	Correct
Gazelles:	*Nope, not with you.*	Wrong
Finish:	*Right, got it.*	Correct

Map:- *Pyramid*

LM	**Response**	**Visual FB**
Start:	*Ok got that.*	Correct
Broken down truck:	*Ok, yes*	Correct
Pyramid:	*Yes that's fine..*	Correct
Disused warehouse:	*Yep, ok.*	Wrong
Abandoned cottage:	*No, not with you.*	Correct
Chapel:	*Ok, yes*	Correct
Chestnut Tree:	*Yes, ok*	Wrong
Allotments 1:	*I don't see it.*	Wrong
Picnic Site:	*Ok, got it*	Correct
Alpine Garden:	*Yep, ok.*	Wrong
Flight museum:	*No, not with you.*	Correct
Parked Van:	*Ok, yes*	Correct
Graveyard:	*I don't know where you mean*	Correct
Granite Quarry:	*Stop, Where's that?*	Wrong
Allotments 2:	*Ok, got it.*	Correct
Collapsed Shelter	*Nope, not with you.*	Wrong
Level Crossing:	*Ok, fine*	Correct
Finish:	*Right, got it.*	Correct

Map:- *Safari.*

LM	**Response**	**Visual FB**
Start:	*Ok got that.*	Correct
Field Station:	*Ok, yes*	Correct
Old Temple:	*No, not with you.*	Correct
Lion country:	*Ok, yes.*	Wrong
Extinct Volcano:	*Got it.*	Correct
Yacht club:	*Ok, got it*	Wrong
Giraffes1:	*Don't know where you mean?*	Wrong
Camera shop:	*Yes, ok*	Correct
Canoes:	*Ok, got it.*	Wrong
Buffalo:	*Yes that's fine.*	Correct
Overgrown Gully	*Don't see it.*	Correct
Giraffes 2:	*Okay, fine.*	Correct
Tribal Settlement:	*No, not with you.*	Wrong
Thatched Mud Hut	*Yes, got it.*	Correct
Site of forest fire:	*I can't see that*	Correct
Cave:	*Stop, where's that?.*	Wrong
Finish:	*Ok, yes.*	Correct

Map:- *Telephone Kiosk*

LM	**Response**	**Visual FB**
Start:	*Ok got that.*	Correct
Telephone kiosk:	*Ok, yes.*	Correct
Stone Circle:	*No, not with you.*	Correct
Farmer's Gate:	*Ok, Got it.*	Correct
Meadow:	*Got it.*	Wrong
Pelicans:	*Ok, found it.*	Correct
Carpenter's cottage:	*Yep, got it.*	Wrong
Ruined Monastery:	*Don't see it.*	Correct
West Lake:	*Right, ok.*	Correct
Stile:	*Yes that's fine.*	Wrong
Great View point 1:	*I don't see it*	Wrong
Popular Tourist Spot:	*Yep, got it.*	Correct
Youth Hostel:	*Stop, where's that?*	Wrong
Great View Point2:	*Yep, found it.*	Correct
East Lake:	*Can't see it.*	Correct
Collapsed Shelter:	*Don't see it.*	Wrong
Finish:	*Right, got it.*	Correct