Testing the roles of disfluency and rate of speech in the coordination of conversation

IAN R. FINLAYSON

A thesis submitted in partial fulfilment of requirements for the degree of
Doctor of Philosophy

QUEEN MARGARET UNIVERSITY

2014

Declaration

I hereby declare that this thesis is of my own composition, and that it contains no material previously submitted for the award of any other degree. The work reported in this thesis has been executed by myself, except where due acknowledgement is made in the text.

IAN R. FINLAYSON

Abstract

This thesis is concerned with two different accounts of how speakers coordinate conversation. In both accounts it is suggested that aspects of the manner in which speech is performed (its disfluency and its rate) are integral to the smooth performance of conversation.

In the first strand, we address Clark's (1996) suggestion that speakers design hesitations, such as filled pauses (e.g. *uh* and *um*), repetitions and prolongations, to signal to their audience that they are experiencing difficulties during language production. Such signals allow speakers to account for their use of time, particularly when they experience disruptions during production. The account is tested against three criteria, proposed by Kraljic and Brennan (2005), for evaluating whether a feature of speech is being designed: That it be produced with regularity, that it be interpretable by listeners, and that its production varies according to the speaker's communicative intention. While existing literature offers support for the first two criteria, neither an experiment with dyads nor analyses of dialogue in the Map Task Corpus (MTC; Anderson et al., 1991) found support for the third criterion. We conclude that, rather than being signals of difficulty, hesitations are merely symptoms which listeners may exploit to aid comprehension.

In the second strand, we tested Wilson and Wilson's (2005) oscillator theory of the timing of turn-taking. This suggests that entrainment between conversational partners' rates of speech allow them to make precise predictions about when each others' turns are going to end, and, subsequently, when they can begin a turn of their own. As a critical test of the theory, we predicted that speakers who were more tightly entrained would produce more seamless turn-taking. Again using the MTC, we found no evidence of a relationship between how closely entrained speakers were and how precisely they timed the beginning of their turns relative to the ends of each others' turns.

Acknowledgements

I would like to acknowledge several funding sources that have supported me throughout my Ph.D. The research presented in this thesis was funded by a Queen Margaret University Research Degree Bursary. Both an award from Queen Margaret University's Vice-Chancellor's Fund and a Grindley Grant from the Experimental Psychology Society helped fund my attendance at the 2012 CUNY Conference. Finally, I am grateful to the organisers of CUNY 2012 for a Student Presenter Award.

There are many people at both Queen Margaret University and the University of Edinburgh who have played a part in getting me to this point. First and foremost, I would like to thank Dr. Robin Lickley and Dr. Martin Corley. Collectively, I am grateful for their supervision, their support, their enthusiasm, their advice on more topics then I could possibly list, and for providing valuable feedback on countless drafts of chapters, manuscripts and abstracts. This would be a poorer thesis without the guidance that they both provided throughout the past four years. Individually, I am grateful to Robin for his assistance with securing the funding that made all of this possible, and to Martin for not only providing a lab and an office to work in but also for first setting me down this road six years ago.

For discussions about various topics that appears in this thesis, I am grateful to Prof. Martin Pickering and Dr. Patrick Sturt. I would also like to thank Dr Jean Carletta and Jonathan Kilgour, without whose help I might still be trying to work out how to extract tokens from the Map Task Corpus. I am grateful to Sam Miller for lending his voice to Experiment 1, and to Ewan Keith and Genevieve Warriner-Gallyer for acting as experimenter and confederate in Experiment 2.

For most of my studies, I was fortunate enough to be surrounded by many people who made the Ph.D. process all the more enjoyable. To name but a few, I am grateful to the four people that I got to share an office with, Elli Drake, Ollie

Stewart, Dr. Paul Brocklehurst and Sam Miller, whose conversations both on- and (particularly!) off-topic made coming into the office especially worthwhile. I would also like to thank past and present Edinburgh Disfluency Group regulars, in particular Dr. Ian Cunnings, for various stimulating discussions, many of which I hazily recall taking place over some very good beer.

Finally, I would like to thank my family and in particular my mother, Barbara, for their endless support, not just in the past four years but also in the twenty-four that preceded them.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Conversation is often likened to ballroom dancing. A joint activity where the tight coordination between participants leads to a product that is greater than the sum of its parts. While psycholinguistics has typically tended to focus on what people say, and what people understand, recent years have seen a growing interest in the ways in which people say what they say.

In this thesis, we will test two theories concerned with how speakers manage conversation. In both of these theories it is argued that aspects of the way in which people speak (their disfluency and their rate of speech) play important roles in the coordination of turn-taking.

In Clark's (1996) account of language use, conversational partners have an obligation to account for the ways in which they are using each others' time. When a person is speaking, it is clear how time is being used. Often, however, spontaneous speech is disrupted. When this happens, speakers may regularly pepper their speech with silences, filled pauses (such as *uh* and *um*), repetitions and repairs. While these disfluencies have traditionally been viewed as symptoms of difficulties that arise while planning speech, Clark has extensively argued an alternative viewpoint where certain disfluencies are actually used by speakers to account for their use of time.

One reason why a speaker may wish to account for their use of time when their speech is disrupted is to prevent interlocutors from interpreting the delay that accompanies the disruption as the end of the speaker's turn. If a partner was to believe that the current speaker had said all that they wanted to then they may begin a new turn of their own. By producing a hesitation, such as an *um* or a repetition, it is argued that a speaker can signal to their interlocutor that despite the disruption they intend to resume speaking, allowing them to continue their turn.

With their turn retained and the disruption overcome, the speaker will eventually finish saying what it was that they intended to say. At this point, the "floor" is open for another participant in the conversation to produce a response by taking a new turn. It has been observed in many studies of turn-taking that the process by which one speaker's turn ends and another speaker's turn begins is almost seamless, with often no perceivable gap between turns. In their oscillator theory of turn-taking, Wilson and Wilson (2005) have argued that these rapid turn exchanges are achieved by conversational partners making precise predictions about when each others' turns will end. They argue that the ability to make such precise predictions about timing arise through a process where the rates at which each partner speaks become similar during conversation.

1.1 Thesis structure

The thesis is divided into three parts. Firstly, we explore one means by which speakers hold onto their conversational turns when their speech is disrupted. Secondly, we examine how subsequent speakers time the beginning of a new turn. Finally, we present the conclusions of this thesis.

In the first part we will evaluate the claim that hesitations are designed by speakers in order to manage conversations. In Chapter 2, literature on the production and comprehension of disfluencies will be reviewed. This chapter serves not only to introduce the subject of disfluency, but also to provide evidence to suggest that if certain hesitations are being designed to be signals, as Clark suggests, then these signals have a reliable meaning which is readily interpretable by listeners.

Chapter 3 introduces the Map Task Corpus (MTC; Anderson et al., 1991), which is not only used to provide one source of evidence for evaluating the claim that certain hesitations are designed, but is also used to test several predictions derived from Wilson and Wilson's account of the timing of turn-taking. This chapter will also discuss the statistical framework that is used for analyses of both experimental and corpus data throughout the thesis.

Repetitions are one type of hesitation that have been argued to have a function in the management of conversation. However, unlike filled pauses, there is relatively little evidence to suggest that repetitions have any effect on listeners' linguistic processing. In Chapter 4, we present an experiment which used a change detection paradigm to investigate whether repetitions have an effect on attention, and consequently on the granularity of semantic representations for the words that they precede. The findings of this experiment provide little evidence to suggest that hearing a disfluent repetition has any effect on listeners' attention.

The final chapter of this part, Chapter 5, presents two studies that tested the claim that hesitations are designed. Firstly, with an experiment that compared the production of hesitations in monologue and dialogue. Secondly, with a set of analyses of hesitations in the MTC which explored whether they are sensitive to manipulations of the situation in which dialogue occurs in a manner that suggests that they are being designed. Neither of these studies provide evidence consistent with the claim that hesitations were being designed: Speakers were no more likely to produce hesitations in dialogue than in monologue, while only manipulations that had direct consequences on the cognitive burden experienced by speakers were found to influence the likelihood that they would be disfluent.

In the second part of the thesis our attention will shift to the process by which conversational partners take turns to speak in conversation. In Chapter 6 we will introduce Wilson and Wilson's theory, as well as reviewing two other prominent theories of turn-taking which have influenced it. In Chapter 7 we will present a further series of analyses of the MTC which tested three predictions, derived from Wilson and Wilson's (2005) theory, about the relationship between rate of speech and the timing of turn-taking. While, consistent with the first two predictions, partners were found to speak at similar rates throughout a conversation and a relationship was observed between the rate at which a turn was spoken and the interval that proceeded it, we found no support for the critical third prediction that there should be a relationship between how similar partners spoke and the seamlessness of their turn exchanges.

In the third part, Chapter 8 will discuss the findings of both empirical strands, and present the conclusions of the thesis.

Part I

Holding onto a turn

Chapter 2

The production and comprehension of disfluencies

This chapter reviews existing literature on the production and comprehension of disfluencies. By providing background for the phenomena that will be the focus of much of the thesis, it is intended that this chapter establishes there is evidence consistent with the claim that one class of disfluencies, hesitations, are designed by speakers to have communicative function.

After introducing the five types of disfluency that will appear in the empirical chapters of this part of the thesis, we will outline Herbert Clark's (1996) account of language use, which has informed much of the discussion about the possible communicative role of hesitations. We then introduce some early studies of the role of one particular type of hesitation, filled pauses (e.g. *uh* and *um*), during turn-taking.

One difficulty with establishing whether or not hesitations are being designed is that the difficulties in language production which Clark argues that they signal could instead simply be the symptomatic cause of hesitations. In order to overcome this difficulty, we will assess the claim that hesitations are designed against three criteria formalised by Kraljic and Brennan (2005). After introducing these criteria, and their previous applications, we will review the literature on the production and comprehension of hesitations, which, when taken together, suggests that at least some hesitations may meet at least some of these criteria.

2.1 What are disfluencies?

The apparent ease with which humans achieve verbal communication disguises a complex set of processes required for conversational partners to produce mutually understandable speech. A speaker must conceptualise, plan, and articulate utterances which express the thoughts in their own mind in terms that can be understood in the minds of others. Despite the proficiency at producing spoken language that humans display, speech rarely proceeds perfectly smoothly: filled pauses (e.g. such as *uh* and *um*), repetitions, prolongations, silent pauses, and repairs litter spontaneous speech. These *disfluencies*, “phenomena that interrupt the flow of speech and do not add propositional content to an utterance” (Fox Tree, 1995, p. 709), are commonplace in spontaneous speech. Studies of spontaneous conversation, as well as of task-orientated dialogues, have reliably shown that speakers are disfluent approximately 6 times for every 100 words they produce (e.g. Bortfeld, Leon, Bloom, Schober, & Brennan, 2001; Fox Tree, 1995; Shriberg, 1994).

As will later be apparent, despite many types of disfluency being associated with difficulties that occur during speech, not all of these appear to result from the same difficulties. For example, the types of problems that may cause a speaker to *uh* may not cause them to produce a repetition. Additionally, instances of an individual type of disfluency may reliably differ. For example, in certain situations a speaker may *um*, rather than *uh*. With these issues in mind, this section will introduce each type of disfluency and discuss systematic differences in the forms that they take.

2.1.1 Filled pauses

For those outside of the discipline of psycholinguistics (and for many inside, perhaps), filled pauses (also known as fillers) may be the archetypal example of disfluency. While the literature on filled pauses frequently refers to *uh* and *um*, there is in fact much variation between the realisations of filled pauses, both between dialects and between languages, for example *este* in Spanish and *ano* in Japanese (for a summary, see Clark & Fox Tree, 2002).

Clark and Fox Tree report that in each of the eight languages they surveyed, there are at least two forms of filled pause which form a contrast to each other. They suggest that often the two forms contain central vowels, with a nasal coda that appears to be optional. They argue that filled pauses are used by speakers to signal that they are about to experience a delay in speaking, with the presence of the coda marking a distinction in the detail of what the filled pause is signalling. In a corpus of spoken dialogue (the London-Lund corpus; hereafter LLC; Svartik & Quirk, 1980) they found that the silences which followed filled pauses with a nasal coda were longer than those which followed filled pauses without this coda. Based on this finding, Clark and Fox Tree argue that *uh* and *um* act as signals that a speaker is experiencing either a “minor” or a “major” delay, respectively.

Clark and Fox Tree’s findings have not gone unchallenged. One observation that has been made is that, rather than using an objective measure of the durations of silences, the lengths of pauses in the corpus were annotated using perceptual units of prosodic stress. For the purposes of establishing durations, Clark and Fox Tree simply counted the numbers of these units. While the subjectivity of the annotation has been pointed to as a weakness of their study (e.g., by O’Connell & Kowal, 2005), it may actually represent a benefit. Different speakers, speaking at different rates, may produce different lengths of minor and major delays. As different speakers may produce different numbers of filled pauses, with different ratios of uhs to ums, the relationship between filled pauses and silences may be confounded by differences in speech rate (for example, if a fast speaker, frequently producing short pauses, produced many more uhs than ums). The use of a subjective measure, which takes into account rate of speech, instead allows measures of duration to be obtained which control for differences between speakers of these sorts.

A second challenge has come from O’Connell and Kowal’s investigation of filled pauses in media interviews with Hillary Clinton, where recordings of the interviews allowed for accurate measurements of silence durations. They suggested that if Clark and Fox Tree (2002) are correct then, being a professional speaker, Clinton should be an expert at using filled pauses as signals. Such expertise should mean that she would be well able to use different forms of filled pauses to differentiate between minor and major delays. Their study found that there tended not to be delays following filled pauses, and that those delays that were present did not differ in duration as a function of whether the filled pause was an uh or um. We may, though, question the logic of O’Connell and Kowal’s argument. While a professional speaker *could* be proficient at using filled pauses, it is perhaps more likely that they are better able to avoid the sorts of disruptions that filled pauses are argued to signal.

Schnadt (2009) similarly failed to find any evidence for a consistent difference in silence durations after filled pauses. However, his analysis was based on only 169 filled pauses, while Clark and Fox Tree (2002) analysed over 4000. Moreover, the pattern observed by Clark and Fox Tree has also been found elsewhere (Barr, 2001; Fox Tree, 2001).

It is still unclear whether the presence of the nasal coda in a filled pause is the product of a choice, intended to signal the length of the delay which will follow. One could conceive of an alternative account where a speaker who anticipates that they are about to produce a long delay may simply prefer to close their mouth while they wait to be able to resume. Regardless of the explanation, however, the weight of evidence does suggest a relationship between the phonetic form that a filled pause will take and the duration of the silence that will follow.

2.1.2 Prolongations

Prolongations are speech segments whose duration is stretched beyond what might be expected in normal speech. Despite occurring relatively frequently (Eklund, 1999; Schnadt, 2009), prolongations have so far been relatively neglected by those interested in typical disfluency, with the lack of the attention they have received leading Eklund (2001) to describe them as the "dark horse" of disfluencies. One explanation for their neglect may be that while it is clear when a speaker has produced a filled pause or a repetition, it may not always be as easy to determine when a segment has been prolonged. Establishing a baseline for segment duration may in itself be difficult, given speaker and situation variability, providing little to compare a possible prolongation with. Those studies which have explored prolongations have found that the segments which are prolonged can appear at any point of any word; however, there appears to be a tendency for prolongations to occur at word-final positions, and in function words more frequently than content words (Eklund, 2001; Eklund & Shriberg, 1998).

In addition to being prolonged, there may be cases where a normally reduced vowel is fully realised (such as producing the normally reduced *the* as "thee" rather than "thuh"). While segments containing these vowels need not be of any greater duration than those containing their reduced forms, these cases are frequently considered alongside prolongations. Fox Tree and Clark (1997) explored cases of non-reductions of *the* in the LLC. They extracted an equal number of instances of "thee" and "thuh" from the corpus, matching them so that each speaker produced an equal number of each realisation. They found that both filled pauses and silent pauses were more likely to occur immediately following thee than thuh. Clark (1996) suggests that such non-reductions are a choice made by speaker, with the intention of signalling that they are experiencing difficulty.

Bell et al. (2003) extend Fox Tree and Clark's study to include the function words *I*, *and*, *that*, *a*, *you*, *to*, *of*, *it* and *in*, while also considering their durations. Using a regression analysis of the Switchboard Corpus (Godfrey, Holliman, & McDaniel, 1992), which controlled for possible confounds such as age and gender of speaker and rate of speech, they investigated whether vowels were less likely to be reduced when words appear in a disfluent context. They found that words appearing in a disfluent context were almost one and three quarter times as likely to contain a non-reduced vowel as those in a fluent context; however, there was no evidence to suggest that the location of the disfluency (preceding or following the word) influenced the odds of a vowel being non-reduced. In considering the duration of the words, Bell et al. found that words in a disfluent context were almost one and a half times as long as those in a fluent context, with a greater effect for those words which precede a disfluency than those which follow (similar results were observed by Shriberg, 1999).

For the sake of clarity, in the remainder of the thesis we will use *prolongation* to refer to the subjectively judged stretching in duration of speech segments, and not to the non-reduction of vowels.

2.1.3 Silent pauses

During spontaneous speech, speakers may occasionally cease to produce vocalisations, and instead pause silently. As we saw when discussing filled pauses and prolongations, silent pauses may co-occur with other forms of disfluency; however, they have long been a subject of interest in their own right (e.g. Goldman-Eisler, 1958)

Ferreira (1993, 2007) draws a distinction between *planning-based pauses* and *timing-based pauses*. While the former are forward-looking, associated with what will follow, the latter are backward-looking, determined by the linguistic material which precedes them. Timing-based pauses represent the time remaining after "subtracting" the duration of vocalising a word from the time allocated to each phrase by the prosodic structure. As such, we would argue that timing-based pauses should not be considered disfluent as their production is unlikely to be associated with difficulty. Perhaps more critically, pauses of this sort do not interrupt the flow of speech, rather they seem a constituent of the normal flow of speech. They would therefore fail to meet Fox Tree's (1995) definition of disfluency quoted at the beginning of this section. As Fox Tree, herself, points out: “not all pauses are disfluencies” (p. 709).

While timing-based pauses tend to appear at the ends of phrases, planning-based pauses may appear anywhere in an utterance where a speaker encounters difficulty. This variability in location can make it difficult to differentiate between those between-utterance pauses which are disfluent and those which are merely the product of the speaker’s natural prosody. Consequently, this imposes a methodological concern for those interested in the production of silent pauses.

If moments of silence can naturally arise in fluent speech then it raises the question of how we can identify those silences which are disfluent? For the practical purposes of research, duration is frequently used as a criterion for identifying silent pauses. Durations that have been used as cut-offs vary from 80ms to 2sec (Hieke, Kowal, & O’Connell, 1983); however, Goldman-Eisler’s (1958) cut-off of 250ms has been widely adopted. Goldman-Eisler claimed that below this level many pauses would reflect the time required by the articulators to make necessary adjustments to move between sounds; although, Hieke et al. (1983) found that pauses as short as 130ms (the shortest their analyses considered) could be accounted for by psychological, rather than articulatory, explanations.

Silent pauses may also be identified perceptually, rather than on the basis of their objectively-measured duration. Listeners are not infallible however, and may miss relatively long pauses whilst reporting pauses that do not exist (Cowan & Bloch, 1948). A common cause of false positives has been found to be where a speaker slows down before increasing their tempo (Martin & Strange, 1968a). Martin (1970) suggests that this reflects a general tendency to wrongly report a pause when the lengths of syllables are stretched. While at times there may be a disconnect between the perception of a pause and its acoustic duration, Duez (1985) found that duration was a strong cue to whether or not a listener could correctly identify when a pause had occurred; however, its strength was sensitive to whether the pause occurred within or between constituents.

2.1.4 Repetitions

When an interruption occurs in spontaneous speech, speakers will frequently repeat sounds, words or phrases before resuming. In (1), the speaker repeats two whole words, “if you”, and one sound, “t-”, but there is no overt indication that they are making any revisions to what has already been said.

(1) [if you t-] if you take a line due south you're gonna hit it[1]

Two of the words repeated in the above example are function words and this is often the case. It has long been known that function words are repeated more often than content words (Maclay & Osgood, 1959). One possible explanation for this pattern may be that function words occur more frequently, providing more opportunities to be repeated. Clark and Wasow (1998) attempted to deconfound the effect of frequency differences between function and content words using the Switchboard Corpus (Godfrey et al., 1992). After collecting all function words and all content words in the corpus, they found that, per one thousand mentions, the former were more frequently repeated than the latter.

It has been claimed that repetitions may not be a homogeneous phenomenon. Hieke (1981) proposes a divide between repetitions which are *retrospective* and those which are *prospective*. Retrospective repetitions are a reaction to an ongoing interruption. A speaker pauses, and upon being ready to resume they repeat preceding parts of the utterance to reconnect whatever followed the pause to the constituent boundary of what preceded it. On the other hand, prospective repetitions may be used strategically: When a speaker is anticipating difficulty in production, they may repeat sounds or words in order to "buy time" while the difficulty is resolved. By producing this prospective repetition, the speaker is able to accommodate the delay in production without having to delay their speech.

Further evidence of the heterogeneity of repetitions has been provided by Plauché and Shriberg (1999). They applied a clustering approach to different prosodic features of repetitions, for example silent pause durations, and durations and f_0 patterns of repeated tokens, which identified different sets of repetitions. In one set, termed *canonical repetitions*, the token that is subsequently repeated is prolonged and frequently followed by a long pause, suggesting production difficulties. This is consistent with Hieke's (1981) retrospective repetitions.

Another set, *stalling repetitions*, match Hieke's prospective repetitions. There is no pause between the original token and the preceding material, however one often appears immediately afterwards. Inversely to canonical repetitions, it is the repeated token which is prolonged, and frequently followed by a pause, suggesting that the speaker is still planning what next to say.

[1] All examples in this chapter come from the HCRC Map Task Corpus (Anderson et al., 1991), unless otherwise stated.

While Hieke proposed two types of repetitions, a third set was identified by Plauché and Shriberg. In this third set, a pause frequently occurs prior to the first mention; however, there is no pause prior to, or following, the second mention. Both mentions in the repetition see some prolongation. They suggest that these are covert self-repairs, where a speaker realises that there is an error in an, as yet unarticulated, speech plan which is covertly repaired (covert repairs will be covered in greater depth in the following section); however, it is not clear how distinct this set is functionally from stalling repetitions. An alternative account of this set could be that upon realising they have produced a pause of "disfluent" duration, the speaker produces a repetition to try to buy time before they are ready to resume. In this account, covert self-repair repetitions become an alternative realisation of prospective repetitions.

2.1.5 Repairs

Thus far the disfluencies that have been discussed appear to arise as a consequence of upcoming material, either because it is not fully prepared or because it has been found to require repair. However, there are times when a speaker realises that something that they have already said is inappropriate or erroneous. On making this discovery, the speaker will frequently cease speaking and attempt to repair what has already been said. Before reviewing the forms that repairs may take, we will first note that the status of repairs as a type of disfluency is not entirely clear. The interruption that takes place before the speaker produces the repaired material could arguably be considered to be a form of hesitation, and in some cases it may be accompanied by a silent or filled pause. However, as we have already suggested there is a difference between repairs and the types of hesitations we have already discussed in that the hesitation during a repair is a response to what has already been said rather than a response to planning what will be said next. Given this difference, throughout this thesis we will consider repairs as being distinct from other types of hesitations.

Levelt (1983) provides a structure of repairs, depicted in Figure 2.1, which has influenced much of the subsequent work on this type of disfluency. In his structure, each repair contains three parts. Firstly, there is the *original utterance* which begins at the last sentence boundary before the speaker suspends speech.

Within the original utterance is the *reparandum*, the material which is subsequently edited. The reparandum can range from single sounds to entire phrases, and in some cases may span the entire length of the original utterance.[2]

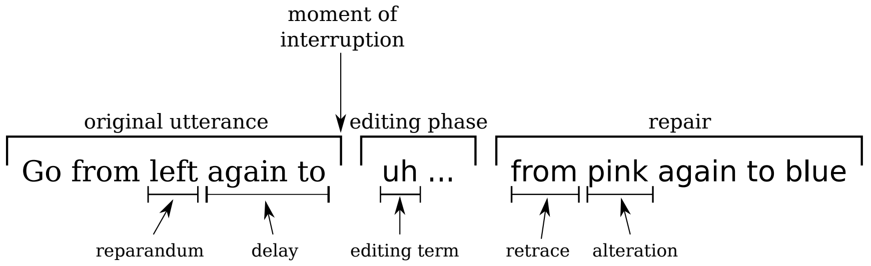


Figure 2.1: Levelt's structure of a repair (adapted from Levelt, 1983)

Following the *moment of interruption* is the editing phase. The content of the editing phase may include silent pauses as well as editing terms, such as filled pauses and interjections (e.g. *well*). Blackmer and Mitton's (1991) observation that reparanda may be immediately followed by repairs suggests that the editing phase need not always occur.

The final section of the repair is the *repair* proper, where the edited material is produced. The repair continues until the end of the current sentence boundary, and includes the alteration, the part of the original utterance which was edited. Prior to the alteration, the speaker may retrace to part of the original utterance which precedes the reparandum. Such retracing may occur to facilitate integration of the edited material with the original utterance, and could be viewed as being similar to Hieke's (1981) retrospective repetitions.

While Levelt (1983) suggests that most repairs share this structure, he identifies different categories of repair. His taxonomy takes a functional perspective, based on the different acts that each category of repair are performing. Examples of each category of repair, adapted from Levelt's corpus of Dutch, are given in (2).

(2a) [We go straight on or] we come in via red

(2b) We start [in the middle with] in the middle of the paper with a blue disc

(2c) Turn left [at node] to node blue

[2] Following the reparandum, there may also be a *delay*: material which is subsequently repeated unchanged. For compatibility with the annotation of the Map Task Corpus (Anderson et al., 1991; Lickley, 1998), any reference to the "reparandum" in this thesis is intended to include the delay in addition to Levelt's reparandum. See Chapter 3 for further details.

(2d) Then right uh grey

If, while speaking, a speaker decides that the message that they wish to convey with the utterance that they are currently producing should be replaced by a different message, and consequently an entirely different utterance, then they will perform a *D-Repair* (for example 2a). D-Repairs were rare in Levelt's corpus of Dutch participants describing coloured images, accounting for only 1% of all repairs. They were similarly rare (2.6%) in Blackmer and Mitton's (1991) analysis of callers to a Canadian radio show.

Sometimes the utterance that the speaker is producing is correct, yet it may not be suitable within the discourse or context. If a speaker produces an utterance which they subsequently realise is ambiguous, for example if they ask "can you pass the mug?" when two mugs are present, then they may repair the utterance to be more specific, "...the red mug?". In Levelt's taxonomy these are known as appropriateness repairs, *A-Repair* (for example 2b). A-Repairs may be used to reduce ambiguity, increase precision, or to maintain coherence with the preceding discourse; *AA-Repairs*, *AL-Repairs* and *AC-Repairs*, respectively. Taken together, they accounted for 25% of all the repairs in Levelt's corpus.

Sometimes utterances which are appropriate may still contain errors when they are articulated. *E-Repairs* (for example 2c) are the repairs speakers make to errors that they have made. Speakers may attempt to repair lexical errors (*EL-Repairs*), syntactic errors (*ES-Repairs*), or phonetic errors (*EF-Repairs*). Although Levelt suggests that many errors are not repaired, E-Repairs remain the most common type of repair in his corpus (42%).

While the three categories of repairs discussed up until this point are performed on already uttered speech, repairs may also be made to material that has not yet been vocalised. According to Levelt's (1989) model, between formulation (i.e. selecting the syntactic structures and lexical items) and articulation, speech plans are monitored for problems. Detecting an error in inner speech may allow for a repair to be made before any of the erroneous or inappropriate material is articulated. Such repairs are said to be covert, *C-Repairs* (for example 2d). As the repair is made before a reparandum is produced, and therefore altered, much of Levelt's repair structure does not appear. Instead, during a C-Repair we may see only an interruption and editing phase, or a repetition of immediately preceding material (the covert self-repair repetition of Plauché & Shriberg, 1999).

Given the relationship between hesitations and planning of upcoming material that we will review in 2.4, it is difficult to be sure that phenomena which resemble Levelt's C-Repairs are truly repairs. For example, while Levelt suggests that the *uh* in example 2d is an editing term appearing in a C-Repair, there may be alternative explanations, such as that the speaker has not yet decided what colour to refer to or that they are experiencing difficulty in lexical access. Levelt, himself, recognises this difficulty in identifying C-Repairs; however, he claims they represent 25% of repairs in his corpus.

Finally, Levelt uses *R-Repairs* to refer to the remaining repairs which cannot be fit into any of the other categories.

Classifying different types of repairs according to their different purposes may be valuable when investigating the psychology of repairs; however, a weakness of this approach is the requirement of the analyst to make subjective interpretations of speakers' intentions. If A-Repairs, for example, are made when a speaker's utterance is unsuitable given the context, then identifying such a repair would require an understanding of relative aspects of the context to at least the level of the speaker. Shriberg (1994) offers a taxonomy which allows repairs to be categorised on the basis of their structure, rather than requiring pragmatic knowledge. Examples of the four types of repair in this taxonomy are given in (3).

(3a) I don't suppose you've got [the balloons] the baboons
(3b) just above [a forest fire] site of a forest fire
(3c) [well the bottom of it] right just draw a straight line
(3d) [go north and] go [north] due north and proceed east

Firstly, *substitutions* (for example 3a) are when some or all of the words in the reparandum are replaced by new words. Secondly, *insertions* (for example 3b) are when a speaker adds one or more words to the repair which were not present in the reparandum. Often insertions may occur to increase the specificity of an utterance, sharing a similar function as Levelt's A-Repairs; however, as Shriberg is concerned only with the structure of the repair an insertion need not always be a repair made for appropriateness. Thirdly, *deletions* (for example 3c) are when any or all words in the reparandum are erased from the repair without substitution. Deletions differ from D-Repairs, by not requiring that all words be deleted. Shriberg also includes *repetitions*, however she makes no assumption that they reflect covert repairs. Finally, *complex* repairs (for example 3d) are those which contain multiple interruption points (i.e. Levelt's moment of interruption), and are composed of, often nested, combinations of other types of repairs. For example, in (3d) *go north and* is repeated, while, within the "repair" section of the repetition, *due* is added in an insertion.

2.2 Clark's account of dialogue

In this section we will introduce Clark's (1996) account of dialogue, in which he suggests that certain disfluencies are produced to serve communicative functions. Clark's account extends wider than disfluencies however, addressing how meaning is understood and successfully expressed by partners in dialogue. We would suggest that there are similarities surrounding Clark's ideas about disfluencies and his ideas about how meaning is established, not just that they both derive from a common idea about the nature of dialogue but importantly that establishing evidence for both is prone to similar problems. The approach that we adopt in this thesis in order to test Clark's ideas about disfluencies was born out of an attempt to solve some of these problems. As such, in this, and the following section, we begin by discussing Clark's ideas about meaning before addressing his ideas about disfluency.

In his 1996 work, *Using Language*, Clark sets out a comprehensive account of dialogue which has come to provide a framework for much of the work which has followed on possible communicative functions of certain disfluencies. The focus of this account lies not in the grammatical structures and speech sounds that occur in dialogue, but rather in how conversational partners come to use language. Clark suggests that dialogue is not merely two (or more) people taking turns to produce and comprehend language. Rather, it is cooperative and collaborative, a joint activity that language users participate in, similar to dancing or two people lifting a heavy object.

During a conversation, both parties have jobs to do. Speakers must communicate their intended message, while listeners have to understand what is being said. These jobs are not undertaken in isolation however, and both parties share responsibility for ensuring the process goes smoothly. Conversational partners work together to ensure that the utterances which are being produced are not only an accurate depiction of the state of affairs, but are also sufficient for being mutually understood (Grice, 1975).

One means by which partners are able to understand one another, Clark claims, comes from sharing common ground (Clark & Marshall, 1981). The common ground contains all of the knowledge which is shared by both partners. Each can then call upon this knowledge, as needed, for expressing ideas. Without explicit negotiation, it is difficult to establish that knowledge is shared. Instead, Clark suggests that common ground contains all of the knowledge that each partner may reasonably expect the other to share given physical, linguistic and community co-presence. If two Edinburgh natives are in conversation, then each may expect the other to have knowledge of the "One O'Clock Gun"; therefore, we would consider it to be in common ground. This would not be the case if the native of Edinburgh was talking to a native of Glasgow. However, if they learn that the Glaswegian was previously a student in Edinburgh then it may come to enter into the common ground.

By knowing what is in the common ground, speakers are able to design their utterances to be understandable by listeners without violating Grice's (1975) maxim of quantity, by giving more information than is required. Evidence that speakers' knowledge of their audiences guides their utterances is provided by Isaacs and Clark's (1987) study of conversations between experts and novices. Pairs of participants were given sets of images of New York City landmarks. One member of each pair, the *director*, was given the set in an order and their task was to help their partner, the *matcher*, arrange their images in the same order. Half of participants were New York natives (*experts*), while the other half were not (*novices*). Pairs were selected so that either both partners were New Yorkers, neither partner was a New Yorker, or only one was a New Yorker (the director, in half of the pairings). Pairs of experts were the most efficient (quantified by the number of words used per landmark) of all possible types of pairs. This is unsurprising as pairs of experts will share a common ground that likely contains knowledge of New York landmarks. Participants were found to quickly establish their partner's "expertise" and design their descriptions accordingly: With an expert matcher, an expert director may refer to landmarks by name; however, with a novice matcher, they may describe physical properties of the landmark (often alongside the name).

As we may not always share relevant information with those we talk to, it is vital that common ground can be developed during a conversation (Brennan & Clark, 1996). Brennan and Clark had participants perform a task similar to that of Isaacs and Clark, which instead used images of everyday objects. In one of the sets of images, one item appeared alongside a pair of similar items, for example a *pennyloafer* appeared alongside *trainers* and *high heels.* This manipulation rendered the use of a referring expression based on the general category (for example *shoe*) insufficient. Below is an example (4) of one trial where the pennyloafer was encountered (p. 1487).

(4) Director: a docksider
Matcher: a what?
Director: um
Matcher: is that a kind of dog?
Director: no, it's a kind of um leather shoe, kinda preppy pennyloafer
Matcher: okay, okay, got it

In this trial the director first refers to the depicted shoe as a "docksider". When the matcher responds that this is not helpful the director suggests an alternative: "pennyloafer". The item is then referred to by this name for the remainder of the experiment, as it becomes part of the common ground. The process by which knowledge which was not originally mutual becomes shared is an example of *grounding* (Clark, 1996).

Considering the importance for successful communication that expressions be mutually understandable, Clark suggests that there is a *principle of closure*: "Agents performing an action require evidence, sufficient for current purposes, that they had succeeded in performing it" (p. 222). At its most simple, sufficient evidence may come in the form of the addressee responding to the addresser's utterance in an expected fashion. If you invite a guest into your living room and invite them to "sit down on the couch", then evidence of "closure" comes when they subsequently sit down on the couch, rather than remaining standing or sitting down on another piece of furniture. As shown by the pennyloafer example earlier, and the following example (5), grounding can also be achieved verbally.

(5) Roger: now, -um do you and your husband have a j- car
Nina: -have a car?
Roger: yeah
Nina: no-

In this example (Clark, 1996, p. 254), Nina appears unsure as to whether Roger is asking about a car (perhaps because he spoke disfluently). Nina responds by repeating what she understood him to have meant, allowing him the opportunity to clarify. When Nina asks "have a car?", she is commenting on Roger's linguistic performance. Clark suggests that such comments take place on what he calls the *collateral track*. While the *primary track* carries signals which are relevant for the official business of the dialogue, collateral signals provide a commentary on the primary track and facilitate successful communication. When Roger asks if Nina and her husband have a car, and when Nina responds that they do not, they are communicating on the primary track, discussing the official business. When Nina asks Roger for clarification, and when Roger provides it, they are communicating along the collateral track.

Clark's idea of a collateral track, where conversational partners are able to produce metacommunicative signals is an important part of the claim that speakers are designing certain disfluencies to be signals. In viewing conversation as a joint act, Clark claims there is a responsibility for partners to account for their use of time. Often, this accounting is done by speaking itself. As long as one partner is speaking then it is obvious to all that they are using the time to speak. Consistent with this idea, it has been observed that when speakers detect an error in their speech they prefer to continue speaking until they are ready to repair it, than to immediately stop talking and wait until they can produce the repair (Seyfeddinipur, Kita, & Indefrey, 2008). Producing speech, even speech that may be erroneous, accounts for their use of time better than silence does.

When they are forced to cease speaking, for example because they are having difficulty in planning what they intend to say next, the speaker must find alternative means to account for what is happening to the time that is being used. It is when time is being used to plan rather than to speak, that Clark (1996, 2002) claims that speakers rely on hesitation disfluencies (such as filled pauses, prolongations and repetitions, but not silent pauses, as they do not account for the use of time) in order to account for how time is being used, and to manage the conversation. When a speaker is forced to stop speaking this often coincides with the production of hesitations. Clark suggests these are purposeful, and to support this claim he invokes his *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).

As an aside, it is not clear how broadly Clark considers his principle of choice to apply. A speaker may choose to produce a filled paused or a prolongation to account for the time spent silent. Alternatively, they may choose to remain silent, and it is not clear what a speaker may be trying to signal by failing to produce a perceptible signal, unless the signal is to be seen as an invitation for their partner to intervene. More generally, when inviting a guest in my living room to sit down I could refer to the "sofa" or the "couch" and be referring to the same piece of furniture, but, while in some situations I may be trying to signal my social class with my choice (see Ross, 1954), it is not clear that this choice must *always* be a signal.

Clark and Wasow (1998) further elaborate on the function of one particular type of hesitation, repetitions, in their Commit-and-Restore model. In the model, it is suggested that following a disruption, words may be repeated for one of three reasons. The first is because of difficulty caused by the syntactic complexity of the utterance that is being produced. Consistent with this, Clark and Wasow found in Switchboard and the LLC that repetitions of *the* occurred more frequently as part of a complex NP (e.g. "the dog down the street") than as part of simple NP (e.g. "the dog"). In the Commit-and-Restore model, syntactic complexity causing a speaker to produce a repetition is a pure process. Such processes are defined by the authors as those which are uncontrollable outcomes of another process (such as syntactic planning during speech).

The second reason that a person may produce a repetition is that they may wish to achieve continuous delivery of a syntactic constituent, rather than having the constituent disrupted by a filled or silent pause. This preference for producing syntactically complete constituents without disruption is known as *the continuity hypothesis*. Clark and Wasow suggest that if the continuity hypothesis is correct then the more severe a disruption is, the more likely speakers will be to produce repetitions. Consistent with this, they found that more severe disruptions of syntactic constituents were associated with an increased frequency of repetitions. Clark and Wasow suggest that one explanation for the continuity hypothesis is that "complete" syntactic constituents may be easier for listeners to parse. As such, using repetitions to achieve continuous delivery is, they suggest, a controllable strategy that is used by speakers for cooperative purposes (i.e. to help their audience). This proposal is similar to Hieke's (1981) retrospective repetitions and Shriberg's (1999) canonical repetitions.

The final reason for producing a repetition, as suggested by Clark and Wasow, is also strategic. They suggest that speakers may sometimes want to make a "preliminary commitment" to the utterance that they intend to produce. Such a commitment may allow the speaker to justify their use of time by producing part of an utterance, and then ceasing speech as they continue to plan its remainder. If the speaker did not make this preliminary commitment then the delay that they would produce as they plan their utterance could be interpreted by listeners as them having reached the end of their turn, giving listeners the opportunity to take a new turn for themselves (a similar function for filled pauses is discussed in the following section). A similar function is suggested by Hieke for prospective repetitions, and by Plauché and Shriberg for stalling repetitions.

From this point onward we will refer to Clark's claim that hesitations are designed by speakers to perform communicative functions as the *hesitation-as-signal hypothesis*. Before continuing to discuss evidence that may be consistent with hesitations serving a communicative function, we will first briefly make clear which types of disfluencies we do and do not consider as being represented by the hesitation-as-signal hypothesis. In Clark's writings on the functions of certain disfluencies (e.g. 1994, 1996, 2002; Clark & Fox Tree, 2002; Clark & Wasow, 1998; Fox Tree & Clark, 1997; Smith & Clark, 1993) he has consistently argued for a communicative role of filled pauses, repetitions and prolongations. Therefore, it is these on which we will focus. For reasons alluded to above, we would argue that silent pauses could not be signals as we do not see what function the absence of a perceptible signal could serve (at least not in Clark's account, where hesitations are suggested to often allow the speaker to make the listener aware that such a pause is about to take place). We would also argue that repairs are not being designed to perform a communicative function. If a speaker detects an error of some sort and interrupts their utterance to make an edit neither the act of interrupting nor the act of editing is in itself communicative in a sense that Clark may intend in the hesitation-as-signal hypothesis (although a speaker could produce an editing phrase, such as a filled pause, in order to alert the audience to the delay that takes place while they prepare their repair). We would suggest that when speakers make repairs it is in order to fix their mistakes, not to signal that they have made a mistake.

2.2.1 Filled pauses and turn-taking

Almost three decades before Clark set out, what we have termed, the hesitation-as-signal hypothesis, a few psychologists were already beginning to explore communicative aspects of one particular type of hesitation, filled pauses. Fundamental to the hypothesis that hesitations, such as filled pauses, are being designed by speakers is that hesitations should be subject to volition. If a speaker could not control the hesitations that they produce, then it would be difficult to imagine how they could be choosing them as Clark suggests. Siegel, Lenske, and Broen (1969) found that speakers could produce fewer filled pauses and repetitions when they received a cash reward for being fluent. Five participants were tested in a series of sessions (between 10-17 in total). Participants were allowed to speak spontaneously about a topic of their choosing, with cue cards suggesting topics if necessary. In some sessions, a counter displayed the number 200, with the number occasionally dropping. Participants were told that the number on the counter at the end of the task would be the number of cents they would earn, in addition to their payment for attending the session. What they were not told, however, was that the number on the counter was decreasing each time they produced a filled pause or repetition.

Debriefing of the participants revealed that they noticed that the decrease coincided with moments of hesitation, and four of the five participants became less hesitant in sessions where their hesitations were being punished. What is perhaps most striking about this study is that participants reported in the debriefing that they chose to pause when they were uncertain (silent pauses were not punished), rather than produce costly filled pauses or repetitions. Siegel et al. had a relatively small number of participants, and there is little evidence of strict control of the few participants they had; however, if speakers can choose not to fill a pause then it would be compatible with Clark and Fox Tree's (2002) claim that speakers choose filled pauses to signal delays.

One reason that a speaker may wish to signal that they are delaying, rather than stopping speaking is to ensure that they retain their conversational turn. If a speaker is forced to stop speaking due to encountering problems then, it is suggested, they are at risk of their conversational partner assuming they have finished and taking over the floor. Instead of simply producing a silent pause, a speaker could produce a filled pause to signal that they are not finished, and that a partner should not attempt to take the turn. Duncan (1972) has labelled signals which a speaker can produce to avoid losing their turn attempt-suppressing signals (although he does not appear to consider filled pauses to be one of those signals). Before discussing a pair of studies which have investigated the possible use of filled pauses as attempt-suppressing signals, we will pause briefly to address some common misunderstandings about this claim. Frequently it is attributed to Maclay and Osgood (1959), particularly in the psychological literature on hesitations. While it is true that they were at least among the first (if not the first) to make this claim in print, it often tends to be presented as the main thesis of the work. Rather, their study was one of the relationship between disfluencies and "both individual differences and linguistic distribution" (p. 19), and this claim seems intended as little more than light speculation as they discuss their conclusions. It is also important to point out that their suggestion was not based on any data (at least not any that they report). This is not always the impression given by those who cite the claim (e.g. Lallgee & Cook, 1969).

Several studies have tested the claim that filled pauses are produced by speakers as attempt-suppressing signals (Ball, 1975; Beattie, 1977; Cook & Lallgee, 1970; Lallgee & Cook, 1969). Beattie (1977) recorded five conversations (three meetings between a supervisor and student, two conversations between attendees at a seminar) and analysed them for filled and silent pauses, and interruptions. Interruptions were more likely to occur during a silent pause (it is not specified whether silent pauses were mid-utterance or whether they were at potential turn exchange points) than during fluent speech, and less likely during a filled pause than a silent pause. Beattie concludes that his results support Maclay and Osgood's claim; however, this is not the case. What is shown in this study is that listeners are less likely to interrupt when a speaker is vocalising (whether fluently or disfluently) than when they are not. What is not shown, however, is that speakers are producing filled pauses with this function in mind. Similarly, both Ball (1975) and Cook and Lallgee (1970) look for support for Maclay and Osgood's claim in the responses of listeners (with mixed results), rather than in the motivations of speakers.

Lallgee and Cook (1969) investigated the issue of speakers' motivations by manipulating a conversational partner's tendency to interrupt. The authors predicted that if speakers felt under pressure to keep hold of their turn then they should produce more filled pauses. Participants took part in ten-minute conversations with a confederate of the experimenter on a political or social topic. Participants were divided into two groups (pre-testing allowed both groups to be matched for tendency to produce filled pauses and general verbosity). In one group, participants were told that the person they were talking to had a tendency to interrupt. The confederate then proceeded to do this three times in the first five minutes, and subsequently began speaking whenever the participant paused. In the other group, participants did not receive this instruction and the confederate was instructed to avoid interrupting as much as possible. Participants appeared sensitive to the pressure manipulation, with those under high pressure more likely to interrupt the confederate than those under low pressure. Critically, however, participants who were under pressure were not any more likely to produce filled pauses or repetitions than those who had no pressure.

Beattie (1977) raises two concerns about Lallgee and Cook's experiment: Firstly, that it relies on the assumption that experimental participants are concerned about being interrupted (all that Lallgee and Cook's results show is that speakers who expect to be interrupted are more willing to interrupt others themselves). Secondly, that instructing participants that they are likely to be interrupted focuses their attention towards aspects of social interaction which they may not consider in normal conversation. While both points are valid, this experiment remains the only evidence that speaks to whether speakers use filled pauses as attempt-suppressing signals.

2.3 Testing for design

In the previous section we introduced the hesitation-as-signal hypothesis: the claim that speakers design hesitations, such as filled pauses and repetitions, in order to provide an account of their use of time to conversational partners. For example, a speaker who experiences difficulty while planning and producing an utterance may produce a filled pause to alert their audience that they will produce a delay. An alternative account to the hesitation-as-signal hypothesis may suggest that while hesitations are a *sign* of difficulty, they are a *symptom* of difficulty, rather than a signal. In other words, hesitations are merely the sound of a speech production system breaking down.

The distinction between a symptom and a signal introduces a difficulty for those who wish to test the hesitation-as-signal hypothesis. It is not enough to show that speakers produce the behaviours that are argued to be being designed, nor that listeners' interpretations of those behaviours are concordant with the function for which they argued to be being designed, rather, they have to produce them in a manner which is consistent with them being designed. To illustrate this point, consider our review of studies investigating the proposed attempt-suppressing function of filled pauses. It was not enough that listeners were less likely to interrupt a speaker when they produced filled pauses than when they did not (Beattie, 1977), rather, speakers had to be more likely to produce filled pauses when they were under pressure to retain their turn (which they were not; Lallgee & Cook, 1969).

The difficulties that we have just discussed are not limited only to the hesitation-as-signal hypothesis. Rather, they are faced by any theory that claims that a particular aspect of language is designed. The theory of audience design suggests that speakers design the content and style of their linguistic behaviour for the benefit of listeners (Bell, 1984). Studies using referential communication tasks, such as those described in 2.2, suggest that speakers may sometimes design their utterances to take advantage of common ground, helping to make them understandable for listeners.

Determining whether or not a linguistic act is designed to be readily understood by the audience is not straightforward. That speakers come to produce utterances which are easily understood by listeners should not be surprising given the large amount of knowledge and context, as well as psychological architecture and mechanisms, which conversational partners are likely to share. In their *interactive alignment account*, Pickering and Garrod (2004) suggest that, generally, successful dialogue is accomplished when interlocutors' situation models (Zwaan & Radvansky, 1998) become aligned. Such alignment is first achieved at lower levels of representation, for example at the levels of syntax and lexical items. Alignment at these lower levels tends to occur through priming, where hearing a particular word or syntactic structure leads speakers to be more likely to produce that word or structure themselves (e.g. Bock, 1986; Branigan, Pickering, & Cleland, 2000; Levelt & Kelter, 1982). Alignment can percolate between levels (for example, the "lexical boost"; Cleland & Pickering, 2003) allowing situation models to become indirectly aligned.

By reusing linguistic material that their interlocutor has already produced (e.g. Garrod & Anderson, 1987; Pickering & Garrod, 2004), speakers are more likely to produce utterances that are mutually understandable without having to design them especially for their audience. Interlocutors need not rely on common ground. Rather, they have an implicit common ground which grows more extensive as their situation models become greater aligned. If, after inviting a guest in my flat to "sit down on the couch", they later refer to the piece of furniture as a *couch* then they could be designing the referring expression on the basis of common ground; however, Pickering and Garrod's (2004) parsimonious alternative account would suggest that my having used the word increased its activation for my guest, causing it to be more easily available when they attempt to select a word to refer to the piece of furniture. As a result, what may appear to be designed by the guest may not in fact have been so.

Kraljic and Brennan (2005, pp. 196–197), formalising the approach followed by Brennan and Williams (1995), provide three criteria against which features of speech and language (for example a referring expression or a hesitation) can be evaluated in order to determine whether they are being designed by the speaker for the benefit of their audience (rather than a word being produced because it currently has the strongest activation, or a hesitation being produced as a symptom of difficulty). Firstly, they must be "produced reliably and spontaneously in dialog". Secondly, they must be "interpretable by addressees". Finally, they must "vary depending on speakers' intentions in the situation or toward addressees". Simply showing that a feature of speech is produced reliably, and that the audience appears sensitive to this reliability, is not sufficient for establishing that the feature is designed. Rather, we must show evidence that the speaker is producing the feature in a manner consistent with it being designed.

For Kraljic and Brennan, a feature of speech is being designed to be beneficial to the audience if the speaker intends it to play this beneficial role. Many utterances are produced to achieve a particular goal or to perform a particular purpose. For example, a speaker may wish to inform a partner of something, inquire into something, or prompt them to perform a specific act. This goal or purpose is known as a communicative intention. Utterances which have communicative intentions are known as speech acts (Austin, 1962). The communicative intention of a speech act may not simply be to inform an interlocutor. If I am in my office and say "it is hot in here" I may not just simply want to comment on the temperature. Instead I may intend that by hearing the comment the office-mate will interpret that I wish them to open a window. If I produced such a comment and an office-mate simply nodded in agreement then the speech act would be unsuccessful. What is necessary for a speech act to be effective is that the office-mate recognises its communicative intention, for example by opening a window.

The importance of recognising the intention behind a communicative act was emphasised by Grice (1957, 1969). In his work on meaning, he draws a distinction between that which is natural and that which is non-natural. An example of natural meaning used by Grice is the case of the spots which result from measles. Upon seeing that a person has spots, you could take that to mean that they have measles, and as such spots become a sign of measles. One would not suggest that the spots were intending to signal the presence of measles, rather the spots are a symptom of measles. If the patient said "I have measles", it would likely be because they intend to signal their condition. Saying "I have measles" conveys a non-natural meaning. The critical difference between these two types of meaning, Grice suggests, is that in the non-natural case "A must intend to induce by x a belief in an audience, and he must also intend his utterance to be recognized as so intended" (1957, p. 383). While the patient who tells you that he has measles may be intending that you respond by, for example, staying away from him, the measles spots have no such communicative intention.

The distinction between natural and non-natural meaning has consequences for the hesitation-as-signal hypothesis. An alternative account of why speakers produce hesitations would be that while hesitations are indeed a sign of difficulty, they are a symptom rather than a signal. Speakers do not produce hesitations because they wish to account for their use of time whilst "holding the conversational ball". Rather, hesitations may merely be the sounds that the language production system makes as it grinds to a halt. In such an account, hesitations are not being designed to signal difficulty. Rather, difficulty causes hesitations in much the same way as measles causes the appearance of spots.

Brown and Dell (1987) provide a demonstration that an appearance of design may be misleading, by adopting an approach that is in keeping with Kraljic and Brennan's (2005) third criterion. When retelling a story, speakers are more likely to refer to atypical instruments than typical instruments. For example if retelling the story of a robber stabbing a man, speakers are more likely to refer to the instrument used when it was an *ice pick* than a *knife*. Atypical instruments were also found to make stories harder to comprehend (with ease of comprehension quantified as the time spent reading the sentence introducing the instrument). As atypical items are less likely to be inferred by listeners, one explanation for speakers explicitly referring to them could be that doing so makes the story easier to understand. In other words, speakers could be designing their stories in such a way that helps ensure that they are better understood by their audience.

Brown and Dell tested this explanation by having participants retell stories to a listener. Half of participants told stories to a listener who could see a picture depicting the event; in half of those pictures the instrument could be seen. If speakers explicitly refer to the atypical instrument to help listeners then we would expect fewer mentions when the image showed the instrument (as the listener had an additional source of information). In Kraljic and Brennan's terms, Brown and Dell manipulated aspects of the situation (the accessibility of information to the listener) and investigated its effect on the production of the feature of interest (explicit referring to instruments). While storytellers explicitly referred to the instrument more often when the listener could view the picture, it did not matter whether the instrument was visible or whether it was typical or atypical. This suggests that speakers were not designing at least one aspect of their language (whether or not they referred to the instrument) in order to help the listener to better understand what they are saying.

Kraljic and Brennan have used their criteria to examine whether speakers employ prosodic cues to help listeners cope with syntactically ambiguous sentences. In the sentence shown in (6), the first prepositional phrase (PP), *in the basket*, could be referring to a particular dog (e.g. if there are two dogs then it is the one in the basket that should be put on the star), or it could be the place that the dog should be put (i.e. in the basket which is sitting on a star). These would correspond to a modifier and a goal interpretation of the first PP, respectively. In spoken language, speakers could use prosodic cues (e.g. lengthening) to disambiguate interpretation of the first PP. Signalling of the modifier interpretation could be achieved by prosodically marking this PP; whilst signalling of the goal interpretation could be achieved by marking the first noun (*dog*).

(6) Put the dog in the basket on the star

In Kraljic and Brennan's study, pairs of participants gave each other instructions similar to (6). The director in each trial was given three objects that they should mention (from a display of four items), and were instructed to use structures like that used in (6). The matcher in each trial was shown the same four items, and followed the instructions while their eye movements were recorded.

Each display that matchers saw contained four images, including the target to be moved (e.g. an image of a dog in a basket). In a 2×2 design, the experimenters varied whether the displays led to a modifier or a goal interpretation (e.g. whether, when a modifier interpretation was intended, a second dog was present, which would make *the dog* ambiguous without a modifier), and whether they were ambiguous or unambiguous (e.g. whether, when a modifier interpretation was intended, a goal interpretation was possible) by the addition of a second basket already on top of a star. These manipulations allow for each of the three criteria to be tested. If directors' prosodic marking was consistent with the intended interpretation then it would suggest that it was being produced reliably. As marking would occur earlier in the goal interpretation than in the modifier interpretation, if matchers looked at the target earlier in the goal condition than the modifier condition then it would suggest they were interpreting what the marking could mean. Finally, if prosodic marking was stronger when the display was ambiguous than when it was unambiguous then it would suggest that the director was sensitive to the needs of the matcher. If participants exhibited marking more frequently when the matcher needed it, then this would suggest that the director was designing their marking with the matcher in mind, consistent with the theory of audience design.

In the goal condition, the director produced a first noun of longer duration than the second noun, while this pattern was reversed in the modifier condition (consistent with the first criterion). Matchers were also quicker to look at the target in the goal condition than in the modifier condition (consistent with the second criterion). Critically, however, directors' prosodic marking appeared insensitive to the ambiguity manipulation: Participants' marking did not vary depending on whether or not the matcher needed it (failing to meet the third criterion). Taken together, these results suggest that while listeners may readily interpret prosodic cues, that are reliably produced by speakers, to successfully parse ambiguous sentences, these cues are not being designed by the speaker for the benefit of the matcher.

2.3.1 Evaluating hesitations against Kraljic and Brennan's criteria

When we consider the hesitation-as-signal hypothesis, we see parallels with the problem posed by the audience design theory: Speakers may produce hesitations with the intent of signalling the difficulty they are experiencing to their audience (cf., for example, using a referring expression because its recent use adds it to the common ground), or hesitations may be an automatic by-product of the difficulty that the speaker is experiencing (cf. producing a referring expression which is highly activated because it has recently been used). Following Kraljic and Brennan (2005), we argue that the hesitation-as-signal hypothesis makes three predictions about the production of hesitations during conversation.

According to the first criterion, hesitations should be produced reliably by the speaker. If hesitations are being designed to signal that a speaker is experiencing difficulty then they should reliably occur when the speaker experiences difficulty. If a speaker is regularly hesitant when production is effortless, if a speaker could routinely maintain fluency when they were burdened, or if different speakers produced different types of hesitations when they face a similar burden, then it could be argued that hesitations do not mean (in either the natural or non-natural Gricean senses) that a speaker is experiencing difficulty. Evidence that there are reliabilities in the production of hesitations, consistent with the first criterion, will be reviewed in 2.4.

According to the second criterion, listeners should be sensitive to any reliability in the production of hesitations. Non-natural signals are produced with the intention that their recognition provokes an effect in an audience. Similarly, according to Clark's (1996) principle of closure, listeners should acknowledge that they understand hesitations. Acknowledgement could take a variety of forms, for example inferring the cause of hesitant speech or not interrupting a speaker who is pausing. Evidence supporting this prediction will be reviewed in 2.5.

Finally, in Chapter 5, we present an experiment and a set of analyses of a corpus of task-orientated dialogue to test the third prediction: that the production of hesitations should vary depending on whether they are more or less necessary for the listener in a given situation.

2.4 Hesitations are reliably produced

If hesitations are being designed by speakers to help manage disruptions that occur during conversation then Kraljic and Brennan's (2005) first criterion predicts that the production of hesitations should be reliably associated with difficulties experienced by the speaker. Much of the early empirical interest in disfluencies, and particularly hesitations, was concerned with what disfluent speech could tell us about the process of language production. As a result, research has extensively explored many factors which lead to the production of different types of disfluencies. This section will discuss factors which may be associated with being hesitant, in order to show that hesitations meet Kraljic and Brennan's first criterion. We begin with the differences between the words and phrases that speakers may produce.

2.4.1 Uncertainty and lexical access

Much of the work carried out by psycholinguists on disfluency has focused on the relationship between accessing the words that you intend to say and producing disfluent speech. Goldman-Eisler (1958) investigated the relationship between speaker uncertainty about upcoming words and the production of hesitations. She used a variant of the Shannon guessing technique (Shannon, 1951), where participants have to predict each successive word in a sentence using preceding context alone. Sentences were transcribed from spontaneous speech which contained silent pauses. Participants' predictions were found to be more accurate for words which preceded silent pauses than for those which followed, suggesting a relationship between hesitations and the subsequent production of unexpected words.

Using the Cloze test (Taylor, 1953), where participants must predict words omitted from complete sentences, Cook (1969) observed a similar relationship between predictability and the production of filled pauses. Analyses of several corpora have also extended this finding to repetitions (Shriberg & Stolcke, 1996).

One reason for a word being unpredictable may be that it has a low lexical frequency. Infrequent words are less likely to be said, by definition, leaving some ambiguity in the results of the above studies. Beattie and Butterworth (1979) attempted to deconfound possible effects of frequency by Cloze testing words and subsequently dividing them into high or low frequency groups. Significantly more low frequency words had low Cloze probabilities than high Cloze probabilities, suggesting that this confound was likely present in the previously mentioned studies. Examination of low frequency words alone did however reveal that less predictable words were more likely to be preceded by a hesitation than those which were more predictable.

While Beattie and Butterworth show that frequency and predictability covary, and that predictability has an effect on the production of hesitations independent of this relationship, we may still expect lexical frequency to have an independent effect on the fluency of speech. Using picture naming, Jescheniak and Levelt (1994) found that participants have a greater naming latency when naming low frequency items than high frequency items. Where the item appears mid-utterance, as is often the case in spontaneous speech, this latency might result in a delay (perhaps with a corresponding filled pause or prolongation to signal the delay). However, while Beattie and Butterworth found that low frequency words were less likely to be preceded by hesitations when the they examined all words, no effect of frequency was found when low Cloze probability words were investigated in isolation. This suggests that, in their data, there was no independent effect of frequency on the production of hesitations.

Levelt (1983) found a correlation between frequency and the production of covert repairs (which, as we argued in 2.1.5, may be difficult to distinguish from hesitations); however, his data comes from a limited set of Dutch colour names. Where more extensive sets of items have been used, findings have tended to be in line with those of Beattie and Butterworth (1979), with little support found for the claim that disfluencies result from the difficulty in lexical retrieval that frequency is thought to produce (see Schnadt, 2009). Schnadt used the Network Task (Oomen & Postma, 2001), where participants are shown networks, consisting of images of items connected by lines, and are asked to describe the path of an animated dot moving through the network. After manipulating the frequency of the items represented in the networks, he found that prolongations were more likely to occur before naming low-frequency images. In this experiment, lexical frequency was confounded with difficulties in pre-lexical processing of the images, demonstrated by a subsequent experiment where the effect was eliminated when participants were shown the names of the items prior to the network task. Again, therefore, we see an absence of strong unequivocal evidence of a link between frequency and hesitations.

In summarising the results of these studies, Schnadt suggests that hesitations do not occur as the result of difficulty retrieving words, but rather difficulty in choosing the words to say. Picture name agreement offers an avenue to explore such choices. For pictures of different items, there may be different numbers of names which could be used to describe them. Sometimes one name may be used by most or all people, other times different people may use different names. Using the network task, low name agreement (i.e. where there is no *dominant* name) has been shown to increase the probability of producing prolongations, filled pauses, silent pauses and repairs (Hartsuiker & Notebaert, 2010; Schnadt, 2009).

Further evidence of the effect of lexical choice on speakers' fluency comes from Schachter and colleagues' (Schachter, Christenfeld, Ravina, & Bilous, 1991; Schachter, Rauscher, Christenfeld, & Tyson Crone, 1994) study of the hesitations of university lecturers. They predicted that the more options a speaker has at any point in talking, the greater the likelihood that the speaker will produce a filled pause. In order to vary the number of options available to the speaker they chose to examine lecturers lecturing in different subjects. Their rationale was that as topics moved from the natural sciences, through the social sciences, to the humanities, the numbers of synonyms, and so the amount of choice, would increase (e.g. while there is no synonym for *atom*, there may be many for *beauty*). They observed 47 lecturers in disciplines from these three areas and found that, despite all being similarly disfluent in discussions of neutral topics in an interview setting, the rate of producing filled pauses increased as subjects moved from the natural sciences to the humanities.

2.4.2 Structural complexity

Up until this point, we have only considered the effects of immediately subsequent words on a speaker's fluency; however, hesitations are not only subject to such "local" influences. Hawkins (1971) examined the location of silent pauses in the spontaneous stories of children and found that two-thirds were at the beginning of clauses (although, almost 20% of these were between the first and second word of the clause). Similarly, Boomer (1965) found that both filled and silent pauses were more likely to occur between clauses than within them. Furthermore, Clark and Wasow (1998) found that function words that tended to appear at the beginning of a clause were more likely to appear in a disfluent repetition than those which tended to appear later in a clause.

Just as the likelihood of being hesitant is influenced by the difficulty of producing an upcoming word, it is also influenced by the difficulty of upcoming clauses and constituents. Ferreira (1991) examined the influence of two aspects of sentences, their length and syntactic complexity, on *initiation times* (the length of pauses which preceded production of the sentence). Participants read sentences and subsequently recited them from memory. Here, initiation times were found to be longer when the sentence was longer.

While Ferreira likens her paradigm to the experience of knowing what you want to say but being forced to wait for your conversational partner to stop before you can say it, it is not clear how well the results can be applied to everyday speech. Fortunately, similar relationships between utterance lengths and disfluency have been found elsewhere. Cook, Smith, and Lalljee (1974) found that in monologues speakers were more likely to produce filled pauses at the beginning of longer sentences than shorter sentences. Shriberg (1996) investigated the relationship between utterance lengths and a wider set of sentence-initial disfluencies (including repairs, repetitions and filled pauses, but not silent pauses) in the Switchboard corpus, finding that as the lengths of sentences increased, the likelihood of them beginning fluently decreased. It is not just at the beginning of a sentence that length may influence fluency. Longer sentences are more likely to contain disfluencies at any point (Oviatt, 1995; Shriberg, 1996); however, analyses of individual types of disfluencies suggest that this relationship may not hold for filled pauses (Shriberg, 1994, although cf. Bortfeld et al., 2001).

In her experiments, Ferreira (1991) also manipulated the syntactic complexity of the sentences that speakers produced (defined as the number of syntactic nodes). After controlling for length, she found that initiation times were shorter before low complexity sentences than high complexity sentences. While Cook et al. (1974) did not find a relationship between syntactic complexity (defined as the ratio of subordinate clauses to all clauses) and the production of filled pauses in English, a relationship between complexity (defined by the complexity of subordinate clauses) and the likelihood of producing a filled pause has been observed in Japanese (Watanabe, Den, Hirose, & Minematsu, 2004).

Initiation times are sentence-initial pauses, and Ferreira also investigated whether greater complexity would result in mid-sentence pauses. By orthogonally manipulating the complexity of the subjects and objects of sentences, she found that the probability of a mid-sentence pause increased with the complexity of the object, with the duration of the pauses increasing with the complexity of the objects. More syntactically complex utterances appear, therefore, to be associated with more, and longer, silent pauses. When taken together with studies of sentence-initial disfluencies, syntactic complexity may be associated with the production of certain hesitations; however, the nature of this association may vary across languages, and depend upon the measure of complexity used.

Filled pauses have been shown to be related to complexity in discourse structure (Fraundorf & Watson, 2008; Swerts, 1998). In an earlier study (Swerts, 1997), participants were instructed to label paragraph boundaries in transcriptions of disfluent Dutch monologues. While participants were given no definition of "paragraph" to use, it was assumed that boundaries would be inserted when a shift in the discourse occurred. The numbers of participants agreeing on each boundary could then be used to provide a measure of the strength of the boundary. As stronger boundaries may reflect greater shifts in topic, we may expect their occurrence to coincide with moments of increased difficulty of planning which may lead to the production of hesitations. Swerts (1998) examined the distribution of filled pauses around these boundaries and found that filled pauses were more likely to occur following stronger boundaries (where there was most agreement) than weaker boundaries.

Further evidence for the link between disfluencies and planning difficulty comes from Bortfeld et al.'s (2001) study of disfluencies in task-orientated dialogue. Pairs of participants performed a referential communication task similar to Brennan and Clark (1996). The pictures which participants had to match depicted either childrens' faces or geometric shapes formed by tangrams. As tangram shapes are less familiar and more abstract than faces it was expected that describing them would place greater demands on planning. Similar to earlier research showing an association between hesitations and the discussion of abstract topics (e.g. Lay & Paivio, 1969; Levin, Silverman, & Ford, 1967; Reynolds & Paivio, 1968), participants were found to be more likely to produce repetitions and repairs when describing tangrams; however, the reverse was found for filled pauses. Bortfeld et al. suggest their results show that filled pauses are strongly related to communication rather than planning; however, they make no attempt to reconcile this claim with other evidence appearing to show that filled pauses occur during moments of difficulty.

2.4.3 Metacognition

So far, we have seen associations between the production of certain disfluencies, primarily hesitations, and the difficulties of planning what to say; however, hesitations have also been found to be associated with how certain we are that what we are saying is true.

Smith and Clark (1993) investigated the relationship between the fluency with which people speak and their *feeling of knowing* (FOK) for what they say (i.e.

how confident they felt about what they were saying). Participants gave spontaneous spoken answers to general knowledge questions. They were subsequently shown the questions again and asked to rate how confident they felt that they could recognise the answer (their FOK). Participants were then shown the answer, along with three alternatives, in a multiple choice quiz which they gave answers to.

Participants' feeling of knowing was found to correlate with their ability to recognise the answer, even when they had answered "I don't know" in the first round of questions, suggesting that having a greater feeling of knowing tends to result from actually knowing. When examining filled pauses (which included interjections such as "oh"), Smith and Clark found they were more likely to be produced in the first quiz when a speaker was incorrect or did not have an answer. When a speaker answered (regardless of whether or not their answer was correct), filled pauses were associated with a lower FOK, suggesting that filled pauses were more likely to be produced when a participant was uncertain of their answer.

We began this section by making the prediction that if speakers design hesitations to help make listeners aware that they are experiencing difficulty, in order to manage the disruption that the difficulties cause, then the production of hesitations should regularly coincide with moments when the speaker faces difficulty. To summarise, there is a clear association between the production of certain disfluencies, particularly hesitations, and difficulties encountered by speakers. Such difficulties may arise from choosing an upcoming word, especially when the chosen word is unpredictable or where there are greater numbers of candidates. The difficulties that may lead to hesitations may also be found when planning larger units, particularly when they are longer or syntactically complex. Finally, hesitant speech may also occur when planning is made harder by the unfamiliar or abstract nature of what is being discussed, or when the speaker lacks confidence in what they are saying. The fact that hesitations occur predictably in spontaneous speech shows that they meet the first of Kraljic and Brennan's (2005) three criteria for being designed: they must be "produced reliably and spontaneously in dialog" (p. 196). In the next section we will review evidence showing that hesitations meet the second criterion by being readily interpreted by listeners.

2.5 Hesitations are readily interpretable

If speakers are producing hesitations in order to alert their audience that they are experiencing difficulty, then Kraljic and Brennan's (2005) second criterion predicts that listeners should be able to interpret that a speaker is experiencing difficulty after hearing a hesitation. While much of the early research on hesitations tended to focus on their production, recent years have seen a growing interest in the comprehension of speech that contains hesitations. In this section we will review evidence showing that hesitations have effects on listeners which are consistent with the claim that listeners can interpret hesitations.

2.5.1 Recognition of disfluent speech

Disfluent speech would seem to pose a grave problem for listeners. The disjointed, abandoned and truncated words that appear in repetitions and repairs often render ungrammatical the surface form of what is produced, while filled pauses systematically pepper the speech signal with "words" which may lack meaning. Yet, despite the disruptions to the structure and flow of spoken language that disfluencies cause, listeners appear to cope, clearly demonstrated by the fact that we readily understand each other during conversations. If disfluencies are problematic for language comprehension then this may lead us to question whether some of them are being designed as signals. After all, Clark (1996) suggests that hesitations should facilitate successful communication. One suggestion for how we manage to weather the storm of disfluent speech which has received some empirical support is that we somehow filter it out, processing only what is correct and not what is erroneous. Such filtering could be in response to some cue that lies in disfluent speech (e.g. Hindle, 1983), or may arise from disfluent speech having properties which reduce its recognisability.

In a series of studies, Lickley (1994, 1995; Bard & Lickley, 1998; Lickley & Bard, 1996, 1998) examined the abilities of listeners to recognise and recall disfluencies during spontaneous speech. If listeners excise disfluent speech during parsing then we may ask what it is that allows them to recognise that speech is disfluent. Lickley and Bard (1998, Experiment 1) explored at what point listeners became aware that a disfluency was impending when they heard repairs and repetitions. They used a word-level gating paradigm, where participants repeatedly heard recordings that built up incrementally, word by word (e.g. "It's", "It's just", "It's just a", etc.) Participants heard recordings of fluent sentences, and sentences containing either a repair or a repetition. At the end of each presentation (e.g. between “It’s” and “It’s a”) they were asked to judge how likely they thought it was that the utterance would continue disfluently. Prior to the interruption point, there was no difference in judgements between fluent and disfluent sentences, suggesting that if there is a cue to disfluent speech then it does not emerge before the disruption itself.

Similar word-level gating was used in a later experiment, except participants were this time instructed to judge how sure they were that the sentence had become disfluent (Experiment 2). Immediately following the interruption point, judgements of disfluency were higher for disfluent sentences than fluent sentences, suggesting that after a disruption listeners can rapidly detect that the speech is disfluent. A final experiment, using much shorter “gates”, found that accurate judgements of disfluency could take place before the first post-interruption word could be identified. As the identification of a disfluency does not appear to require lexical processing this would suggest that any cues to disfluency may be prosodic, rather than semantic or syntactic.

While word-level gating provides some insight into on-line processes, listening to the same sentence repeatedly as it increases may not provide an entirely accurate representation of speech perception in spontaneous dialogue. In a more naturalistic study, Lickley (1995) investigated disfluency recognition by having participants listen to recordings of a speaker disfluently describing the process of building a paper house. While listening, they were provided with a transcript with all of the disfluencies removed. Participants were instructed to mark on the transcript any point at which the speech did not match the transcript, and also follow the instructions by building their own paper house. Of interest was the accuracy with which participants detected when a filled pause, repair or repetition had been removed. Participants accurately detected the removal of a filled pause 55.2% of the time, with participants particularly sensitive to those occurring between-sentences (with a similar bias observed by Martin & Strange, 1968b). Such insensitivity to within-sentence filled pauses could result from listeners tending to represent the semantic form of the sentence and discarding its surface form (as found by Jarvella, 1971). Accuracy with single word repetitions and repairs was worse, 27% and 39.3% respectively; although accuracy for both improved when more words were affected.

By instructing participants to build houses whilst detecting mismatches, Lickley hoped to stop participants from focusing on what they were hearing more than they might in everyday conversation. Christenfeld (1995) investigated the effect that participants' focus has on their ability to recognise filled pauses. Rather than identifying individual instances of pauses, participants heard a recording of a particularly disfluent caller to a radio talk show and subsequently, as part of a larger questionnaire, were asked to estimate the number of filled pauses produced by the speaker. Two-thirds of participants heard a version of the recording that had either all filled pauses replaced by silent pauses or all pauses eliminated entirely. Before hearing the recording participants either received no instructions on what they should focus on, or were instructed to focus either on the content of what was said (e.g. "What is his position?") or the style with which it was delivered (e.g. "Is he eloquent?"). Participants whose instructions emphasised the style of delivery estimated a higher number of filled pauses when they heard the filled pause version of the recording; while the estimates of those participants who received the instructions emphasising content were lower, and appeared insensitive to the editing of the recordings. Taken together, Lickley's and Christenfeld's studies suggest that listeners who are focused on the content of what is said, which we might assume to be the default position in dialogue (as listeners appear unable to focus on both the meaning and sound of what is heard; Martin & Strange, 1968b), show a tendency to miss filled pauses.

Bard and Lickley (1998, Experiment 1) found that the content of disfluent speech may also be less likely to be recognised than that of fluent speech. Participants heard samples of spontaneous speech from the Map Task Corpus (Anderson et al., 1991) which were either fluent, or contained a repair or repetition. Samples were presented in word-level gating, and at the end of each presentation participants wrote down the words they had heard and were able to make corrections to words they were unsure about. Bard, Shillcock, and Altmann (1988) have shown that not only is word recognition helped by preceding context, but also that words are not always recognised as soon as they are heard. Instead, *late recognition* often occurs where a word is only recognised following subsequent context (21% of words in their word-level gating experiments). In Bard and Lickley's study, words appearing in reparanda (the token that is later repeated, in the case of repetitions) were less likely to be subject to late recognition and, likely as a result, more likely to be missed. As these words immediately precede the interruption point, subsequent context may not be relevant (particularly for repairs, where, by definition, the reparandum may not fit the context), as a result they were less likely to be subject to late recognition. Words following the interruption point are also less likely to be recognised immediately and more likely to be recognised late. We may view this as the interruption damaging preceding context, forcing the listener to rely more heavily on subsequent context. Bard and Lickley characterise the recognition of disfluent speech as exhibiting graceful failure. The disruption that disfluencies cause to context deprives us of a valuable aid to recognition which, in effect, leaves us "deaf" to the reparanda of repairs and repetitions.

2.5.2 Comprehension of disfluent speech

While Bard and Lickley (1998) show that sometimes we may not always fully recognise the content of repairs, there is growing evidence to suggest that they may still influence the process of comprehension.[3] Lau and Ferreira (2005) showed that words appearing in the reparanda of repairs, which Bard and Lickley (1998) found were particularly likely to be missed, may help disambiguate garden path sentences. Participants made grammaticality judgements about sentences which contained main verb/reduced relative ambiguities, for example "the little girl selected for the role celebrated with her family and friends". In a sentence such as this, when the verb *selected* is encountered the preferred interpretation is that it is a main verb, such as in the sentence "the little girl selected one piece of candy"; however, the verb is intended to be part of a reduced relative clause. Despite being grammatical, when participants make grammaticality judgements of garden path sentences they tend to be less likely to rate them as grammatical as matched unambiguous sentences (Ferreira & Henderson, 1991). Participants heard fluent sentences that were either ambiguous (e.g. containing *selected*) or unambiguous (e.g. *chosen*), and sentences containing repairs which, although always ambiguous in their repaired form, included a reparandum supporting either an unambiguous reduced relative (e.g. *chosen*), or ambiguous main verb (e.g. *picked*), interpretation. While disfluent sentences were less likely to be rated grammatical than fluent sentences, where reparanda were unambiguous, the likelihood of being judged as grammatical was found to be higher than when they were ambiguous; suggesting that, despite being repaired, the contents of the reparanda exert a "lingering" effect on how listeners parse sentences.

[3] It remains an open question whether the repairs that influence comprehension are those that are not missed, or whether they can influence comprehension without being consciously recognised (cf., e.g., Sereno & Rayner, 1992; Trueswell & Kim, 1998). An additional, and potentially related, open question is whether the effects observed in the studies presented in this section are artefacts of the use of scripted repairs as experimental stimuli. Both of these questions are outside of the scope of this thesis, but we would suggest that they are worthy of future investigation.

Bailey and Ferreira (2003) found that filled pauses can also be used to resolve garden path ambiguities. In a sentence beginning "while the man hunted the deer...", the *deer* may be the object of the verb *hunted* or the subject of a subsequent clause, for example "while the man hunted, the deer ran into the woods". They proposed that as filled pauses are frequently found at clause boundaries (e.g. Boomer, 1965), listeners could exploit this regularity to determine whether *deer* was an object or subject. Participants heard sentences where the ambiguous noun was the subject of a subsequent clause, and was accompanied by an adjacent filled pause. As filled pauses are frequently found at clause boundaries, when it preceded the ambiguous noun this would suggest a subject interpretation, while when it followed the noun it would suggest an object interpretation. Sentences were more likely to be rated as grammatical where the use of filled pauses encouraged the correct interpretation (i.e. where the filled pause preceded the noun).

One interpretation of Bailey and Ferreira's finding is that listeners are sensitive to their own experiences of being disfluent (or are at least sensitive to regularities in others' disfluencies), and use these to anticipate the cause of someone else's disfluencies. A listener may become aware that filled pauses tend to precede clauses, and so upon hearing a filled pause they assume a new clause is about to begin and are able to avoid "being led down the garden path". The possibility that listeners make inferences about filled pauses which are guided by knowledge of their distribution has guided several recent studies interested in their effects on comprehension.

In section 2.4.1 we discussed the well-established relationship between the predictability of an upcoming word and the likelihood that it would be preceded by a hesitation. In a study of Event-Related Potentials (ERP), Corley, MacGregor, and Donaldson (2007) investigated the effect of hearing a filler on the N400—an ERP component thought to be associated with the integration of words into the unfolding linguistic context (Kutas & Federmeier, 2011; Kutas & Hillyard, 1980) and closely related to contextual probability and predictability (DeLong, Urbach, & Kutas, 2005; Kutas & Hillyard, 1984). Upon encountering "tongue" in *Everyone's got bad habits and mine is biting my tongue*, the amplitude of N400 is expected to be larger than it would be upon hearing the more predictable "nails". Corley et al. found that when the critical word was preceded by "uh" the difference in amplitudes between the predictable and unpredictable conditions was reduced. This reduction suggests that the presence of the filler made "tongue" less unpredictable. When participants were subsequently surprised with a recall test for target words that had appeared in sentences, it was found that recall was better for those words which had been initially preceded by filled pauses. This suggests that filled pauses may have relatively longer term effects. Similar reduction and recall effects were also observed for silent pauses (MacGregor, Corley, & Donaldson, 2010), but not for repetitions (MacGregor, Corley, & Donaldson, 2009).

Another source of evidence suggesting that listeners are able to use regularities in the production of filled pauses to predict information about upcoming words is the visual world paradigm (VWP; e.g. Arnold, Fagnano, & Tanenhaus, 2003; Arnold, Hudson Kam, & Tanenhaus, 2007). Cooper (1974) showed that participants who heard a sentence while viewing a scene containing items referred to showed a tendency to gaze towards those items in the scene. Such eye-movements tend to occur rapidly (Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995), and can commence before the item has been mentioned if preceding context allows it to be predicted (Altmann & Kamide, 1999; Kamide, Altmann, & Haywood, 2003; Kamide, Scheepers, & Altmann, 2003).

The introduction of new entities into the discourse is more likely to be preceded by filled pauses than referring to an existing entity (Arnold, Wasow, Ginstrom, & Losongco, 2000). Using the VWP, Arnold et al. (2003) found that listeners may be sensitive to this pattern. Participants viewed grids containing images of four objects (e.g. a candle, a camel, grapes, and a salt shaker). While viewing each of them, they heard two sentences. The first sentence instructed them to put either the candle or camel below the grapes. In the second sentence they were instructed “now put the candle below the salt shaker”. Up until *now put the ca-*, it is ambiguous which item will be named and in the absence of context we would expect an equal likelihood of fixations on the candle or camel (Allopenna, Magnuson, & Tanenhaus, 1998). With the context provided by the first sentence, participants showed a preference for the previously named item. However, when *candle* was preceded by a hesitation, *thee uh*, there were more looks to the item which had not previously been named. Developmental evidence demonstrates that listeners first exhibit this association between filled pauses and the mention of new items at around two years of age (Kidd, White, & Aslin, 2011).

Arnold et al. (2007) found listeners to be similarly sensitive to the relationship between filled pauses and the ease of describing an item (e.g. Barr, 2001). They reasoned that abstract images would be harder to describe than images depicting more familiar objects, as the abstract images would be less likely to have conventional names. Participants viewed sets of images of everyday objects and unfamiliar abstract shapes. They then heard sentences instructing them to click on one of the images. While objects were described by name (e.g. *red ice cream cone*), abstract shapes were described by their form (e.g. *red funny squiggly shape*). When participants heard *thee uh* before the object was referred to they were more likely to look at the abstract shape than they were when the sentence was fluent.

The temporal delay hypothesis

While listeners appear sensitive to filled pauses, we have so far remained agnostic as to whether the effects of filled pauses that we have discussed are due to the sound-form (e.g. the *uh* or *um*) or whether the delay they provide offers respite to listeners to process speech. Corley and Hartsuiker (2011) have termed this latter explanation the *temporal delay hypothesis*. While from some perspectives this distinction may be trivial, for example if our interest is in how people come to understand disfluent speech then it may make little difference if a pause is filled or "unfilled", it may be a significant issue for those arguing that disfluencies are designed to alert an audience that a speaker is experiencing difficulty. If the consequences for listeners of producing a signal are equivalent to those of not producing a signal, then this would lead to the question of why speakers bother to produce the signals when they do.

Several studies offer support for the temporal delay hypothesis. The reduction of N400 amplitude for unpredictable words following pauses has been observed both when they were filled (Corley et al., 2007) and unfilled (MacGregor et al., 2010);[4] while Bailey and Ferreira (2003, Experiments 1 & 2) showed that pauses help to disambiguate garden path sentences whether they were filled by an *uh* or a background noise of similar duration (e.g. dog barks). In their own experiments, Corley and Hartsuiker (2011) had participants view pairs of images depicting objects while hearing sentences instructing them to press the button which corresponded to the object mentioned in the sentence (e.g. "now press the button for the bed, please"). Each sentence included a delay, either before *button*

[4] It is not clear why the same effect was not observed for repetitions (MacGregor et al., 2009), as the disruption they cause should also provide a delay. However, one might conjecture that a delay is less helpful when it is filled by linguistically meaning, albeit repetitive, speech (see MacGregor, 2008).

or before the name of the target, ensuring that the duration of sentences was the same between conditions. Across three experiments the delays included a filled pause, a tone or a silent pause. In each experiment, participants were faster to press the button when the delay preceded the target, but, consistent with the temporal delay hypothesis, a comparison across experiments showed there to be no differences between the three different types of delay.

Despite some empirical support for the temporal delay hypothesis, there is also evidence that the sound of a filled pause could be driving the observed filled pause effects (Fox Tree, 2001).[5] In separate experiments in both English and Dutch, participants were instructed to listen out for particular words in recordings of spontaneous speech and press a button as soon as they had detected them. All target words were preceded by filled pauses (50% *uh*, 50% *um*), although in half of trials the filled pause was excised. For *uh*, participants were faster to detect the target when it had been preceded by a filled pause than when it had not. For *um*, however, no difference was found between the fluent and disfluent conditions. A difference between how listeners respond to *uh* and *um* would seem to suggest that the effect they have may not be wholly attributable to the additional time that they provide for processing.

Fox Tree suggests that these differences may be explained by the differences in meanings of *uh* and *um* proposed by Clark and Fox Tree (2002). As *uh* is supposed to signal a short delay, the listener may heighten their attention in anticipation of the end of the delay. *Um*, on the other hand, is thought to signal a longer delay; and it may be that heightening your attention for a resumption that will take an indeterminately longer time to occur is impractical. Corley and Stewart (2008) suggest that an alternative explanation may lie in Fox Tree's decision to retain silence surrounding the filled pauses when she excised them. As a consequence, the pauses left behind when removing *um* were greater than those left behind *uh*; so, in effect, even the "fluent" *um* condition may have been relatively disfluent.

Stronger evidence to support the idea that the effects attributed to filled pauses are not being driven by the delay alone comes from studies which compare filled pauses to coughs (Barr & Seyfeddinipur, 2010; Fraundorf & Watson, 2011). Fraundorf and Watson had participants listen to spontaneous retellings of stories

[5] See also Fox Tree (2002), who found that hearing silent pauses and filled pauses had different effects on participants' impressions of a speaker.

from *Alice's Adventures in Wonderland*. The recordings were edited so that either filled pauses or coughs (with silence added to match filled pause durations) appeared before some plot-points. Participants were subsequently asked to recall the story that they had heard, and were found to be more accurate in the filled pause condition than in either the cough or the fluent control condition. Furthermore, the improved recall was not just restricted to those plot-points which were immediately preceded by the filled pause. Taken together, these results suggest that *uh* and *um* were providing a general benefit to memory for plot-points which was not offered by delays of similar length.

What types of knowledge guide listeners' interpretations?

So far, we have assumed that listeners have a knowledge of factors that cause certain hesitations and use that knowledge to help predict the cause when they hear a hesitation occur. A question which follows from this assumption is whether other sources of information guide this prediction. For example, if a person hears a hesitation, do they predict that the cause will be whatever is most likely to have caused the speaker to be hesitant, or do they predict that the cause will be whatever would be most likely to cause his or herself to be hesitant? Compatible with the idea that listeners may take a speaker-centric approach to predicting the cause of hesitations, Arnold et al. (2007, Experiment 2) found that the tendency to look at abstract shapes following a filled pause was eliminated when participants believed the speaker was anomic, and might therefore have difficulty naming all items.

Barr and Seyfeddinipur (2010) discriminate between these two accounts using a modified version of Arnold et al.'s (2003) VWP experiment. Participants heard speakers describe abstract shapes, and were able to see the shapes that the speakers saw. Shapes were presented in pairs, and in experimental trials the shape that was described was presented alongside a shape that had previously been seen, but critically had not been described. According to an account suggesting that people do not take a speaker-centric approach, if the speaker was disfluent in this trial then we should expect the listener to predict that they were about to refer to the previously unmentioned item. However, in half of trials, the speaker changed before the final trial so the participant would not know how familiar they were with the shapes. The question here is whether the participant would rely on general distributional knowledge alone and infer that the speaker was about to refer to the previously unmentioned shape, or whether they would avoid making predictions because they were unable to take the speaker's perspective. Using mouse-tracking, Barr and Seyfeddinipur found that participants appeared only to make predictions when they had previously heard the speaker, and knew what they had and had not already described. Their results suggest that when listeners hear hesitations, such as filled pauses, they may attempt to take the perspective of the speaker and infer what may be the cause of their difficulty.

This section began with the prediction that if speakers are designing their hesitations in order to alert their audience that they are experiencing difficulties then listeners hearing hesitations should exhibit behaviour that is consistent with them recognising this. To summarise the evidence reviewed, when listeners hear a filled pause they appear aware that the speaker has encountered difficulty, and attempt to take the speaker's perspective to infer its cause. Such inference likely draws upon knowledge of the sorts of situations in which hesitations tend to occur (summarised in 2.4); however, inference is mediated by knowledge of the speaker (e.g. whether they have previously mentioned an item). Given these findings, and mixed evidence on whether the benefits that are observed to come from filled pauses result from the uhs and ums which fill the pauses or the delays themselves, we would be inclined to reject the temporal delay hypothesis which suggests that such benefits arise from delays which provide listeners with longer time to process speech. In sum, there is strong evidence that, at least in the case of filled pauses, certain hesitations are interpretable by listeners, consistent with Kraljic and Brennan's (2005) second criterion for design.

2.6 Conclusion

Hesitations are commonplace in spontaneous speech, and some have suggested that hesitations (a class of disfluencies including filled pauses, repetitions and prolongations) are designed by speakers to be signals in order to manage conversation. In particular, speakers may produce hesitations as a signal to their audience that they have not yet finished their turn, therefore reducing the likelihood of an interlocutor attempting to start a turn of their own. While we may know what disfluencies *could* signal, and that listeners appear responsive to the signal, this alone is not sufficient to accept that speakers are designing them as signals for listeners. Rather, what is needed is evidence that the production of hesitations varies in manners consistent with them being designed with this purpose. Examining this possibility will be the focus of Chapter 5.

In the following chapter we will introduce the Map Task Corpus (Anderson et al., 1991); analyses of which will be included in Chapters 5 & 7. After discussing the collection and preparation of the corpus, we will present some descriptive statistics of disfluencies and other aspects of speech, before introducing the statistical framework which will be used in all following empirical chapters.

Chapter 3

The Map Task Corpus and approaches to statistical analysis

This chapter comprises three sections. In 3.1 we introduce the corpus of task-oriented dialogue which is analysed in Chapters 5 & 7. We will discuss its collection and annotation, as well as the preparations we have made for our own analyses. Analysing a corpus of spontaneously elicited speech may be problematic for many of the commonly-used statistical tools of psycholinguistics, so in 3.2 we introduce a statistical framework that is better suited to dealing with problematic data of this sort. We also discuss the steps taken to ensure that our data meets the assumptions of the framework that we adopt. Finally, in 3.3 we describe the process adopted to construct the models used to make statistical inferences in the three empirical chapters in this thesis.

3.1 The Map Task Corpus

The Map Task Corpus (MTC; Anderson et al., 1991) was designed with the intention of providing those interested in investigating linguistic phenomena in a dialogue context with data that avoids two of the significant methodological difficulties that may be encountered when using naturally elicited corpora, or experiments: Firstly, that the sorts of phenomena that they may be interested in may not occur with sufficient frequency in corpora of naturally occurring dialogue, and secondly, that it may be impossible to control for, or perhaps even know, the aspects of context which influence the phenomena. One solution to these problems, frequently employed by psycholinguists, is the use of research paradigms which elicit relevant phenomena in a controlled experimental context. Anderson et al. (1991) liken the outcome of this approach to the story of the blind men and the elephant,[1] as the speech that is elicited in that strict context is informative only for those interested in the same narrow field of linguistic phenomena. The MTC was intended to supplement this approach by providing a corpus of dialogue within which a larger array of phenomena could be investigated without having to sacrifice such control that robust conclusions could not be drawn.

True to the aims of its creators as being a source of data for speech and language researchers with a wide range of interests, a variety of aspects of the MTC dialogues have been analysed. Research has covered a broad range of topics, including timing and turn-taking (Bull, 1996; Bull & Aylett, 1998; Forsyth, Clarke, & Lam, 2008), the effects of context on intelligibility of speech (Anderson, Bard, Sotillo, Newlands, & Doherty-Sneddon, 1997; Bard et al., 2000) and disfluencies (Bard, Lickley, & Aylett, 2001; Branigan, Lickley, & McKelvie, 1999), syntactic priming (Reitter, Moore, & Keller, 2006), and factors relevant for achieving communicative success (Boyle, Anderson, & Newlands, 1994; Carletta & Mellish, 1996).

The corpus was generated by having participants take part in a collaborative task, the Map Task, which allowed them the freedom to produce natural linguistic phenomena within a controlled situation. Each member of pairs of participants in the map task had a map of landmarks which their partner could not see. Through collaboration, one partner came to draw a route on their map which was only present on the other's map. Labelling of the landmarks allowed the experimenter to control some of the words that participants used, while manipulation of the context of the dialogue (e.g. whether participants were friends or strangers, whether partners could see each other) allowed for exploration of the effects of context on speech. Finally, as the task was goal-oriented (i.e. to cooperatively recreate a route) it was possible to quantify communicative success using several objective metrics (e.g. how closely routes matched, how long it took partners to complete the task).

[1] In this story a group of blind men attempt to learn what an elephant is like using touch alone. Each man touches only one part of the elephant. When they later compare their experiences they find they are in disagreement about what an elephant is like. For example, the man who touched the ears thinks the elephant is like a fan; while the man who touched the tail thinks the elephant is like a rope.

3.1.1 Design

The MTC is composed of annotated transcriptions of dialogues recorded between 64 University of Glasgow students (32 male, 32 female) performing a cooperative task. Each participant was recruited with a friend, with each pair of friends randomly matched up with another pair of friends to form quads, and members of each pairing being strangers to the members of the other. Pairs of participants took turns to direct each other through one of twelve maps. Each map consisted of labelled images of objects which formed landmarks to which partners could refer. Each participant had their own version of the map which their partner was unable to see. In each dialogue, one of the maps included a path visible to only one participant, the *giver*, whose job it was to describe the path to their partner, the *follower*, so that they could draw it on their own map. Each participant performed the task four times, twice each with two members of their quad (their friend and one of the strangers, with the order of friend and stranger counterbalanced). Participants performed both the giver and follower roles twice, using the same map for each dialogue that they were a giver. Finally, half of all quads (and therefore half of all participants) performed the task with a screen separating participants, preventing them from seeing each other.

3.1.2 Annotation

All transcription and subsequent annotation of the corpus was carried out by human coders, unless otherwise noted. The annotation is strictly hierarchical, with the smallest units representing each individual word, non-linguistic noise or period of silence produced by each participant. A unit, or set of units, may be referred to by tags at various levels, representing various layers of annotation (e.g. prosodic and syntactic information: see Isard, 2001, for details). All annotation has been converted to XML and can be queried using the NITE XML Toolkit (Carletta, Evert, Heid, & Kilgour, 2006). In this section we will discuss only those levels of annotation which are relevant for the analyses presented in this thesis.

In addition to the word form, each token is annotated for the start and end of the utterance and the conversational turn in which it occurs. Conversational turns represent paragraphs, as identified by the Spoken Dialogue Parser (McKelvie, 1998) used to tag parts of speech and parse the corpus. Turns alternate between

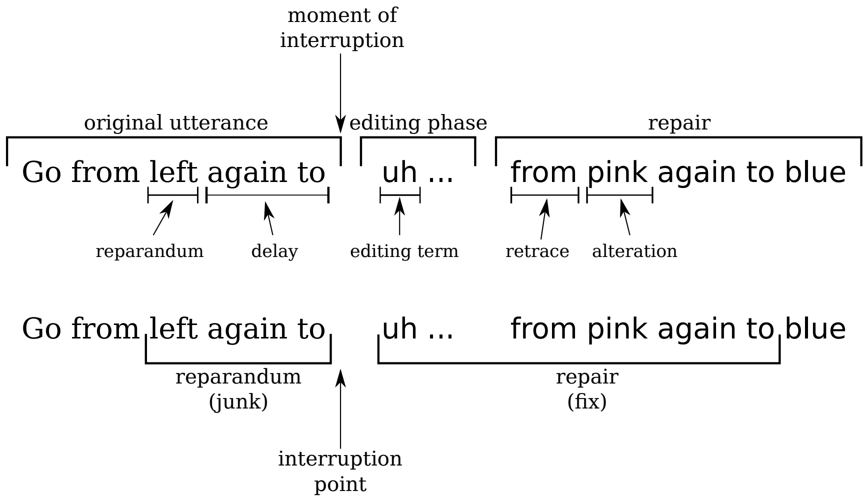


Figure 3.1: Comparison of the structure of repairs used in the annotation of the MTC and Levelt's (1983) structure. Junk and fix are labels used in the XML.

conversational partners, and one turn can begin before the previous turn has ended if partners interrupt one another.

Coding of repairs in the MTC follows Lickley's (1998) taxonomy. Substitutions, insertions, deletions, repetitions and complex repairs are coded at the disfluency level. Figure 3.1 illustrates how the annotation of repairs in the MTC maps onto Levelt's (1983) repair structure, introduced in the previous chapter. Reparanda and repair segments are annotated for each disfluent event, separated by an interruption point. For complex repairs, disfluent events have more than one interruption point, and reparanda and repair segments may be embedded. In the XML, *junk* and *fix* are labels used to identify different parts of a repair. Junk tokens are all of those tokens which precede the interruption point from the beginning of the reparandum, while fix tokens are those tokens which "fix" the reparandum.

Filled pauses are not annotated at the disfluency layer. Within the annotation of the MTC, filled pauses are treated as "fluent" tokens with a part-of-speech tag that identifies them as being filled pauses.

3.1.3 Preparations for analyses

We extracted 152,690 tokens from the corpus, representing all whole words and word fragments, as well as information from the part-of-speech and disfluency

Table 3.1: Total numbers of each token marked as a filled pause in the MTC.

Eh	Ehm	Er	Erm	Uh	Uhm	Hmm	Huh	Mm	Nah
689	640	162	139	107	77	13	1	155	1

layers of annotation. For each token the number of syllables was counted, using the MRC Psycholinguistic Database (Coltheart, 1981) where possible. For those words that did not appear in the database (including word fragments), the number of syllables was counted by the author. The number of tokens (fluent or disfluent) appearing in each turn was counted to provide a measure of turn length. Finally, for each turn we calculated the participant's articulation rate (measured in syllables per second, excluding pauses) by dividing the total number of syllables spoken in the turn by the summed duration of all tokens in the turn.

Ten different words were coded as being a filled pause: *eh*, *ehm*, *er*, *erm*, *hmm*, *huh*, *mm*, *nah*, *uh* and *um*. Counts for each word are given in Table 3.1. While some of these may be considered back-channel responses, we know of no example in the literature where hmm, huh, mm, and nah have been included as a form of filled pause which may be being designed as a signal. To ensure that we were fairly assessing others' claims about hesitations being designed, we did not consider them as disfluent for the purposes of our analyses and instead coded them as fluent tokens.

Summary statistics for the MTC are given in Table 3.2.

Table 3.2: Summary statistics for the MTC. For turn and conversation length, ranges and means are given for each conversation. For rates, ranges and means are given for each participant.

	Range	Mean (SD)
Turn length (in tokens)	1–133	7.30 (7.80)
Conversation length (in turns)	32–478	163.86 (83.77)
Articulation rate (in syllables/second)	4.30–6.84	5.38 (0.47)
Disfluency rate (per 100 words)	3.23–14.51	8.59 (2.50)
Filled pause rate (per 100 words)	0.17–5.95	1.24 (0.97)

3.2 Statistical analyses

In many of the analyses of the corpus appearing in this thesis we are interested in not only categorical predictors (e.g. speaker's role, familiarity of partners) but also continuous predictors (e.g. partner's articulation rate, conversational turn). Statistical tools designed for analysis of factorial designs (e.g. ANOVA) are unsuitable for analysing categorical and continuous predictors together. It is also unlikely that corpus data (not just from the MTC) meets the assumption that data be balanced, which is made by many of these tools. Instead, we will use linear regression for all primary analyses appearing in this thesis.

For the sake of clarity, before discussing linear regression further we first define some of the terminology that we will be using. The linear model, which lies at the heart of linear regression, can be expressed as

$$y = \beta_0 + \beta_1 x + \epsilon. \tag{3.1}$$

While this model is much less complex than many which will appear in this thesis, all of the constituent parts of these larger models are contained within. We will refer to y as the *outcome*. Just like the dependent variable in an ANOVA, it is the variable that we are interested in modelling (e.g. how fast a speaker speaks, or how long a listener spends fixating on a word). We will refer to x as a *predictor*. In our analyses, it is a variable which has either been manipulated or measured (e.g. how fast a speaker's partner is speaking, or whether or not the sentence a listener hears is disfluent) and, like the independent variable in ANOVA, our analyses investigate whether it shares a relationship with the outcome (e.g. Do people speak faster with faster partners? Do people spend longer fixating words which they have earlier heard preceded by a disfluency?). β_0 and β_1 are *coefficients*, which describe the relationship between predictors and the outcome. It is the value of these which are estimated when we construct a model. ϵ is the *error* that is associated with each observation of the outcome in the data (i.e. the noise that a model cannot account for). Finally, there is the *intercept* which is the value of y when x is zero. In 3.1, the intercept is represented by β_0 (this can be considered as a coefficient multiplying a variable with a constant value of 1).

Using linear regression gives us more freedom in the types of variables which we can test. For example, we are able to build models which control for an array of continuous and categorical confounds which may be commonplace in a corpus of spontaneous conversation. However, there are still assumptions which our data must meet if we are to use linear regression. In the following section we will discuss what these assumptions are, and the steps taken to ensure that they are met.

3.2.1 Meeting regression assumptions

By using linear regression our analyses took place within the framework of the general linear model. However, we did not assume that the outcome variables that we analysed would meet all of its assumptions. There are four assumptions which we would reasonably expect that our data would not meet: 1) that errors were normally distributed (i.e. across all the data analysed, the differences between the actual outcome variable and the model's predictions of the outcome variable, the *residuals*, should follow a normal distribution); 2) that there was a linear relationship between predictors and outcomes (i.e. the relationship between each variable follows a straight line, rather than following a curve or sharing another non-linear relationship); 3) that there would not be any multicollinearity (i.e. all predictors in a model should be orthogonal); and 4) that there was independence (i.e. no correlations) between the error for each observation.

In this section we will discuss the steps we took to ensure that our data met each of these assumptions. For the purposes of our analyses we applied one of two approaches depending on the nature of the outcome variable (i.e. whether it was continuous or categorical). We begin by discussing the approach taken with continuous outcome variables (e.g. how fast a participant is speaking) before discussing the approach that was taken with discrete outcome variables (e.g. whether or not a speaker is disfluent).

Continuous outcomes: Box-Cox transformation

When an outcome was continuous, we wanted to use linear regression to regress predictors onto outcome variables. For linear regression it is required that residuals be normally distributed. One step that can be taken to help ensure that this is the case is for the outcome itself to follow a normal distribution. As we adopted an incremental approach to model construction (see 3.3) in our corpus analyses (rather than testing a single pre-specified model), we tested whether our outcomes were normally distributed prior to the model construction process.

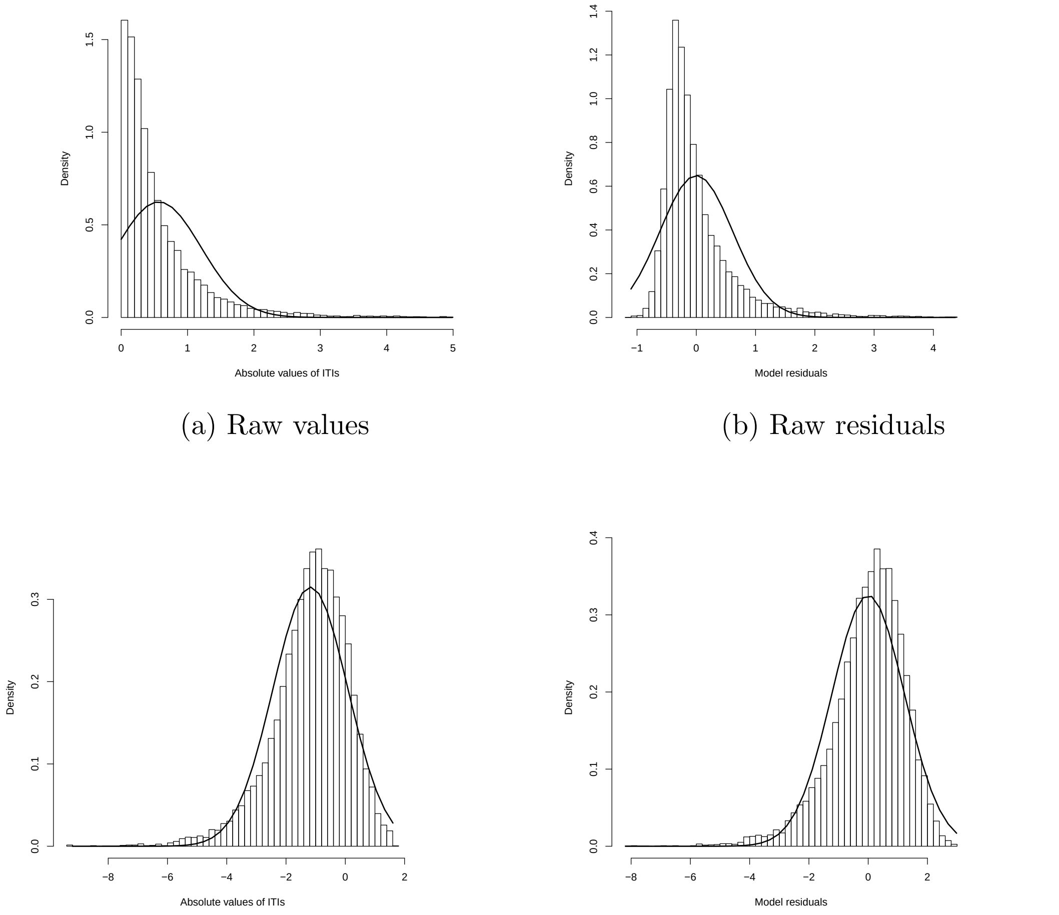
We did not assume that all continuous outcome variables came from a normal distribution. Rather, we used a goodness of fit test to assess the normality of each variable. Two tests are commonly used to assess normality: the Shapiro-Wilk test and the Kolmogorov-Smirnov test. However, neither of these tests were deemed appropriate for use in the work presented in this thesis: The Shapiro-Wilk test is known to be overly-sensitive when testing variables consisting of large numbers of observations, while it is not possible to accurately estimate p values for the Kolmogorov-Smirnov test when ties are present in the variable that is being tested (i.e. when there are two or more observations of the same value). As some of the corpus analyses presented in this thesis considered data with over ten-thousand observations, and we had no reason to believe that there would not be ties, we instead assessed normality using the Cramér-von Mises test, which is better suited to larger numbers of observations and allows p values to be calculated accurately regardless of whether or not there are ties. The test is implemented in the nortest (Gross & Ligges, 2012) package for R (R Core Team, 2013).

Where outcome variables were found not to be normally distributed we used a power transformation designed to make the variable as close to being normally distributed as possible. For each outcome, y, a Box-Cox transformation (Box & Cox, 1964) was applied:

$$y(\lambda_1, \lambda_2) = \begin{cases} \frac{(y+\lambda_2)^{\lambda_1}-1}{\lambda_1} & \text{if } \lambda_1 \neq 0 \\ \log(y+\lambda_2) & \text{if } \lambda_1 = 0 \end{cases} \tag{3.2}$$

where λ_1 is the *power* parameter and λ_2 is the *shift* parameter. Values of λ for each variable can be estimated using a maximum likelihood method, implemented in the geoR package (Ribeiro Jr. & Diggle, 2001) of R. When all values of y are positive, λ_2 is taken to be 0, otherwise a value is estimated which ensures all values of y are positive (i.e. it is greater than the absolute difference between zero and the minimum value of y).

A demonstration of the benefit of Box-Cox transformations is shown in Figure 3.2. In Chapter 7 we will investigate factors which influence the precision of the timing of turn-taking in conversation. Precision will be operationalised by taking the absolute value of inter-turn intervals (ITIs; the time between the end of one turn and the beginning of the next). We would reasonably expect that this variable would not be normally distributed (as it is bound at zero). A histogram showing the raw values of the precision is given in Figure 3.2a. Differences are strongly positively skewed and the data does not appear to be normally distributed (visual inspection suggested that the original "signed" ITIs came from a normal distribution with a mean close to zero). When the data was Box-Cox transformed, shown in Figure 3.2e, the data more closely resembles the normal distribution. Comparison with a log transformation (often used for variables with a positive skew) of the same variable, shown in Figure 3.2c, provides a demonstration of the advantage of such bespoke transformations as are offered by the Box-Cox transformation. Furthermore, we tested the final model of precision (constructed in 7.2.3), with each "version" of the variable as the outcome. Comparison between Figures 3.2c & 3.2e and Figure 3.2b provide a

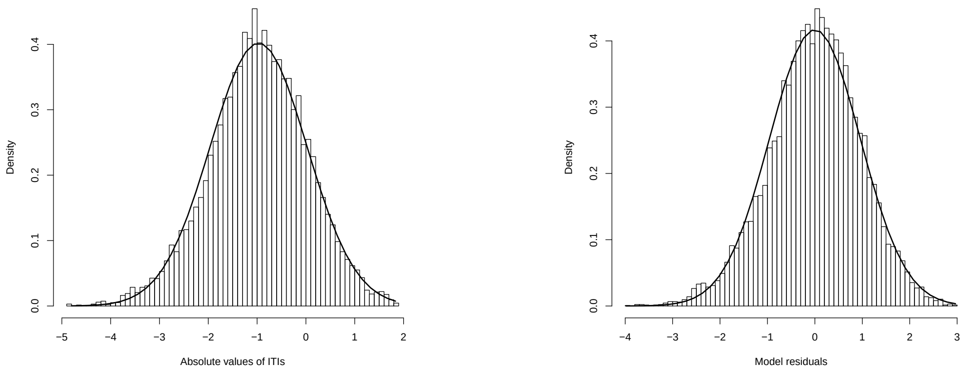


(a) Raw values

(b) Raw residuals

(c) Log transformed values

(d) Post-transformation residuals



(e) Box-Cox transformed values

(f) Post-transformation residuals

Figure 3.2: Histograms of the absolute value of ITIs (a, c and e) and the residuals of the final model of this outcome constructed in Corpus Analysis 2c (b, d and f), with un-transformed, log transformed and Box-Cox transformed values ($\lambda = 0.161$). For further details of the variable, and the analyses, see Chapter 7.

demonstration that transforming outcome variables can help to ensure residuals are normally distributed.

A further strength of the Box-Cox transformation is that at different values of λ_1, commonly-used transformations will be applied (e.g. where $\lambda_1 = 0.5$, it is equivalent to a square root transformation; where $\lambda_1 = -1$, it is equivalent to an inverse transformation; and where $\lambda_1 = -1$, it is equivalent to a log transformation). Where $\lambda_1 = 1$, an identity transformation is performed. Therefore, for each y, the transformation that will be applied will be that which brings y as close as possible to being normally distributed, even if the best transformation is no transformation (i.e. if y comes from the normal distribution).

Discrete outcomes: Generalized linear regression

For the discrete outcomes that we are interested in (e.g. the probability of a speaker being disfluent) the assumption of linearity between outcomes and predictors can not be met. To highlight the incompatibility between discrete outcomes and the assumption of linearity in linear models, imagine that we are interested in the relationship between the length of an utterance and the probability that it will begin with a filled pause. We collect samples of speakers producing utterances of various lengths and find that the probability of producing an *uh* at the beginning of a one word utterance is .4. For two word utterances we find that the probability is .8. If we assume a linear relationship between length and disfluency then we would conclude that producing each word is associated with a .4 increase in the chance of being disfluent. When we come to consider three word utterances we find ourselves predicting a 1.2 probability that the speaker will be disfluent! As this example shows, because probabilities are bounded (i.e. they must lie on the interval between 0 and 1) the increase in probability associated with each unit increase of the predictor (e.g. the number of words) cannot be constant and therefore the relationship is not linear.

One solution to this problem is the use of the generalized linear model.[2] This allows the general linear model to be used with data from other distributions in the exponential family (e.g. binomial, beta, gamma, Poisson, etc.). In practice, this means that the outcome is allowed to come from one of these non-normal distributions, while a link function provides a relationship between the linear component of the model (the equation shown in 3.1) and the non-normal outcome (McCullagh & Nelder, 1989). The link function helps ensure that the errors come from the same distribution as the outcome, whilst also preventing coefficient estimates which would lead to "impossible" predicted values (e.g. expecting a probability greater than 1 or less than 0). For the analyses in this thesis, we used logistic regression, with a logit (log-odds) link function, for binomial outcomes (e.g. whether or not a speaker was disfluent); and Poisson regression, with a log link function, for count outcomes (e.g. how many times a listener fixated on a word).

3.2.2 Analysing unbalanced data

The corpus analyses presented in Chapters 5 & 7 took individual tokens or turns, respectively, as units-of-analysis. While in both cases these choices brought advantages over alternative approaches (which are discussed in the respective chapters), they also brought the potentially harmful consequence of losing the ability to ensure that our data was balanced, thereby increasing the possibility of multicollinearity in our models (and therefore violating one of the assumptions of linear regression). As participants were free to produce as many words, in as many turns, as was required to complete the map task, for each participant, in each cell of the map task design, there could be an unequal number of observations. This was particularly concerning in the MTC, where we expected some consistency

[2] An approach sometimes used is to apply an arcsine square root transformation to the probabilities before using a linear regression (cf. Jaeger, 2008, for a critique of this approach).

Table 3.3: Total number of tokens produced by participants in the MTC by role and ability to make eye contact

	Followers	Givers
Unable to make eye contact	26,218	56,608
Able to make eye contact	21,493	48,386

in the pattern of imbalance. Givers of instructions have been shown to produce more words per turn than followers (Boyle et al., 1994), similarly—perhaps to compensate for the lack of non-verbal communication allowed—partners who are unable to see each other produce more words. We therefore expected that givers who are unable to see their partners would produce longer utterances, thereby contributing more observations to our data. In addition to correlations between utterance length and role, and between utterance length and eye-contact, the consequence of having significantly more units in one cell than in another (as Table 3.3 shows to be the case) is that there would be a correlation between the variables used to code role and eye-contact themselves (and in the MTC this appeared to be the case, Spearman's $\rho = 0.01$, $p < .001$). Such correlations between fixed effects should, of course, not have occurred given the orthogonal design of the MTC.

In order to lessen the possibility of multicollinearity, which can lead to inflated variance estimates (and, consequently, inflated standard errors, bringing a greater likelihood of type II errors; Marquardt & Snee, 1975), several precautionary measures were taken before our data was analysed. To help reduce the previously suggested correlations between fixed effects, discrete predictors were sum coded (using values of $-.5$ and $.5$ to aid interpretation of coefficients). In a balanced data set, the use of sum coding gives a mean of zero for each variable. As the data is unbalanced, however, the mean of each variable would not necessarily be zero. Therefore, the values used for sum coding were themselves subsequently centered.

All continuous parameters were centered before being tested, as should be common practice in order to ensure their mean is zero—helping to avoid ill-conditioned models (e.g., Bradley & Srivastava, 1979). Additionally, all continuous predictors were standardised. While this has no effect on multicollinearity, it was found to help ensure that models converged.

3.2.3 Dealing with random effects

For many of our analyses, the data we are concerned with includes random effects, such as participant and item effects. While there are no widely-agreed definitions of fixed and random effects (see Gelman, 2005; Gelman & Hill, 2007, for discussions of several incompatible definitions), for the purposes of the analyses in this thesis, we will follow Clark (1973) in defining an effect as random if the levels we have observed are samples of a larger population that we intend to generalise to. So, for example, speakers and their partners in the MTC are random effects, because they are a sample of a wider population of people, and the passages heard by participants in Experiment 1 are also random effects, because they are a small selection of possible sentences a person could hear.

An effect is considered fixed if we are interested in only the levels that are observed in the data, and if we do not intend to generalise to other possible levels. So, for example, in Experiment 1 we investigated whether listeners were differently sensitive to speech containing disfluent repetitions and pauses than they were to fluent speech. In this case we would not expect our statistical analyses to generalise to other types of disfluency (e.g. filled pauses); however, we would expect that our results should generalise beyond the participants in our study (i.e. beyond the observed levels of random effects). In the regressions that we have discussed up to this point all parameters are fixed effects.

Random effects are a challenge for the fourth assumption of linear regression: that errors be independent. As we would expect observations within each level of a random effect to be similar (e.g. we would expect a faster speaker to be generally consistent in speaking faster than slower speakers in all their turns), this clustering of observations should also manifest in the errors (as the model should consistently underestimate, or overestimate, their rate of speech)—violating the assumption of independence of errors, as a result—unless steps are taken to take into account the possibility of these clusters.

Random effects cannot be readily accommodated by classical linear regression. One solution would be to use separate by-participants and by-items (and by-partners, and by-maps, etc.) ANOVA (Clark, 1973). However, for reasons that should be obvious by this point, a return to the factorial analyses, continuous dependent variables and balanced data of ANOVA would be far from desirable. Mixed-effects regression (Breslow & Clayton, 1993; DebRoy & Bates, 2004) provides an alternative by allowing us to account for as many random effects as may be required within a single model, subject to computational tractability, in addition to the fixed effects that we are interested in. In effect, this "breaks down" the clusters of errors that are associated with the levels of the random effects: What makes any two errors similar, that they are, for example, produced by the same person, is built into the model in much of the same way as we treat observations from within the same experimental condition.

By including a random intercept for a random effect, such as participants, we allow for the possibility that each participant will have their own intercept. For example, in the context of an analysis appearing in this thesis, this would allow participants to have their own baseline tendency to be disfluent.

It is not just the baseline, however, which may vary between participants (or items, conversations, etc.). Different participants may be differently sensitive to the manipulations which give rise to our fixed effects: If we are interested in whether givers of instructions in the MTC are more likely to be disfluent than followers of instructions, for example, then it may be the case that certain participants find the giver role more difficult and would be more likely to be disfluent as a consequence. Random slopes can be included for any fixed effect which varies *within* the levels of a random effect (i.e. within each participant or item). They are, however, inappropriate for fixed effects varying *between* the levels of a random effect (including interactions containing at least one between-level fixed effect), as data will not be available for every level of the fixed effect.

As an example of why between-level fixed effects should not be included as part of random slopes, consider the ability to make eye contact, which is manipulated between-participants in the MTC. Each participant will either always be able to make eye contact or they will always be unable to make eye contact. It is not unreasonable to imagine that some people will be more disfluent than others when they are unable to make eye contact (perhaps because they rely on visual cues more than others); however, we have no record of their speech when they are able to make eye contact. Because of this absence of data, we cannot be sure whether their disfluency rate is indicative of their baseline disfluency rate or the effect that being deprived of the ability to make eye contact has on them. In contrast, if we imagine that some people are more likely to be disfluent with strangers than other people are (perhaps because they are less concerned with making a good impression than other people) then this possibility can readily be controlled for because we have a record of their speech with strangers *and* with friends.

When including both a random intercept and a random slope in a model we also have the choice to allow for a correlation between the intercepts and slopes for each level of the random effect. An example of such a correlation would be if participants who were generally more likely to be disfluent were particularly strongly affected by the cognitive burden that arises from filling the giver role.

3.3 Model construction

One strength of using regressions is that researchers have options in constructing the model that they will use to make inferences from their data: They may choose to construct a single model containing all the fixed and random effects that they expect to potentially vary, or they may build and compare a set of different models, containing different parameters, in order to determine which is best justified by their data. While we do not argue that either approach is better than the other, we would suggest that different approaches to model construction are better suited to different situations. Empirical work appearing in this thesis includes both experimental and corpus-based methods, and we would argue that the data emerging from each of these methods should not be treated in the same fashion. In this section we will discuss the approaches that we took to analysing data from each of these sources. Regardless of the approach taken to model construction, all analyses were performed in R, using the lme4 package (Bates, Maechler, & Bolker, 2013).

3.3.1 Experimental analyses

When designing an experiment, a researcher begins with (an often small) number of hypotheses and makes as few manipulations as are necessary to test them. Consequently, in designing an experiment the researcher is implicitly constructing a statistical model of the relationship between predictors (the manipulations that they make) and an outcome (the measures they are interested in). When it comes time to analyse the results of the experiment, we would argue that it is *this* model that the researcher should be concerned with. Throughout this thesis we therefore analysed all experimental data using models which included all fixed effects, without testing whether or not the presence of any particular fixed effect significantly improved the fit of a model.

Our use of full models extended to using the maximal random effects structure. In practice, this meant that in addition to including random intercepts for all random effects, our models included random slopes for all fixed effects which licensed them (i.e. those which varied within the levels of the random effects). Our decision to use maximal random effects structures did not merely arise from beliefs about the relationship between a statistical model and the experiment which generated the data the model is applied to: Recent simulations have shown that the use of both random intercepts and random slopes reduces the probability of Type I & II errors (Barr, Levy, Scheepers, & Tily, 2013).

While the testing of models with maximal random effects structures is preferable to testing "simpler" models, there may be cases (particularly in experimental data, which typically contains fewer observations than corpus data) where the maximal model does not converge. In order to avoid this possibility, we did not impose correlations between random intercepts and random slopes when we analysed experimental data. Simulations reported by Barr et al. suggest that a failure to include these correlation terms does not have harmful consequences for significance testing. In cases where this "fuller" model still did not converge, we followed the approach adopted by Gann and Barr (2012) of identifying the highest order random slope (for our purposes, this was always an interaction; however, in a study where predictors were polynomials this would be the highest degree of the polynomial) with the least variance in the partially-converged model and eliminating it. This was repeated, as necessary, until the model converged.

3.3.2 Corpus analyses

Unlike the experiments appearing in this thesis, which are intended to directly test hypotheses about the effects of manipulations on outcomes, our corpus analyses were predominantly exploratory. When conducting an exploratory investigation of a corpus there may be variables which could be taken into account which we have no hypotheses about (and, in the case of control parameters, we may not even be interested in them at all). Unlike experimental research, where an, at least implicit, statistical model will precede data collection, when working with corpora the data often precedes the development of hypotheses and models. We would argue that model construction for exploratory corpus analyses should adopt a more exploratory incremental approach.

All generalized linear mixed effects models were fit by Laplace approximation. For linear mixed effects models, the models that were compared were fitted using a Maximum Likelihood (ML) approach; however, the final models that we report were fitted by Restricted Maximum Likelihood (REML) approach. While REML is preferable to ML, as it produces more reliable standard errors for estimates of coefficients (Patterson & Thompson, 1971), it is inappropriate for comparing models which differ in their fixed effects (Pinheiro & Bates, 2000). By using ML for model comparison we could be sure that our models are constructed in a manner which is appropriate, while fitting the final model with REML ensured that in the model from which we intended to draw inferences, standard errors for the estimated coefficients were more accurate.

The approach to model construction taken in the corpus analyses in this thesis consisted of a two-step process that was applied to each of the different types of parameters in the regression (i.e. random intercepts, random slopes and fixed effects).

Firstly, sets of candidate models, which differed in only one parameter, were compared and ordered according to their absolute log-likelihood. For example, if the parameters were A, B and C (parameters were either all fixed effects or all random effects), then a model containing A was compared with a model that was identical except A was replaced by B, and a model that was identical except A was replaced by C. Secondly, the model with the smallest absolute log-likelihood (the test model) was compared to a base model using a log-likelihood ratio test. As model construction was an iterative process, the base model at each iteration was the model constructed in the previous iteration. The log-likelihood ratio was calculated as $-2(l_1 - l_0)$ (where l_0 and l_1 are, respectively, the log-likelihoods for the model before and after the addition of each parameter). As this statistic has a null distribution which follows that of χ^2, improvement could be assessed with a χ^2 test, with the number of additional parameters taken as degrees of freedom. If the test model significantly improved fit (if $p < .05$) then that model was accepted as the base model for the next iteration. For example, if the model containing A was found to be significantly better than the models containing either B or C then a model containing A and B would be compared to a model containing A and C, and the better model of these two would then be compared to the base model containing only A. If the test model did not significantly improve fit then that parameter was not considered again. These two steps were repeated for each of the remaining candidate models, until all parameters had been tested.

The process was used to construct each aspect of the model in a series of stages. In the first stage random intercepts were tested, with a fixed intercept automatically included.[3] As the lme4 package did not allow us to compare models with and without random effects, we took the candidate model with the smallest absolute log-likelihood as our initial base model. The two-step process was then applied to the remaining random effects.

In the second stage random slopes were tested for those fixed effects which we intended to interpret, *predictors-of-interest*, but not for those which were intended to eliminate noise and confounds in the data, *control predictors* (simulations reported by Barr et al., 2013, suggest that excluding random slopes for control predictors does not increase Type I error rates for predictors-of-interest). Random slopes were only tested for those random effects which were accepted as random intercepts in the previous stage. Using the two step process, random slopes were tested for one random effect at a time, with the order following that in which they were entered into the random intercept model (i.e. in order of their log-likelihood). Only random slopes for fixed effects which varied within the levels of random effects (e.g. within-participants) were tested, and as with random intercepts the order in which random slopes are tested was guided by their log-likelihoods (from smallest to largest). As we made no assumptions about the relationship between random intercepts and random slopes, each random slope was tested without imposing a correlation between intercepts and slopes for each group.

In the third stage the random slopes in the model were tested with and without correlations between intercepts and slopes. Log-likelihood ratio tests were used to determine if a correlation was justified. In any case where the model containing a correlation failed to converge it was automatically rejected, as this suggested the data could not support the correlation. At the end of this stage we had a model containing the fullest random effects structure justified by the data.

Finally, taking the previous model as a base model, the two-step process was used to test fixed effects: firstly control predictors and then predictors-of-interest. For each set of predictors, the order in which predictors was tested again followed the size of their log-likelihoods.

[3] While there may be cases where a fixed intercept is not necessary (e.g. in a linear regression, where the baseline is zero, or in a logistic regression, where a baseline likelihood is 50%) it is our experience that exclusion of an intercept produces model coefficients which are less readily interpreted.

3.3.3 Obtaining p values

In all of our analyses, we were interested in whether or not estimated coefficients for each predictor differed significantly from zero. For the generalized linear mixed-effects models provided by the glmer function in lme4, which we used to analyse, for example, the likelihood of a speaker being disfluent in the analyses reported in Chapter 5, *p* values were calculated using the Wald statistic (see Agresti, 2003). Currently, it is a matter of controversy as to how *p* values should be calculated for linear mixed-effects models (see Bates, 2006), such as those we use to analyse fixation durations in Chapter 4 or speech rate in Chapter 7. Whilst Baayen, Davidson, and Bates (2008) recommend the use of Monte Carlo Markov Chain simulations to estimate *p* values, this is yet to be implemented for mixed-effects models containing random slopes. As our models were likely to contain random slopes, we decided *a priori* that *p* values would be estimated from the *t* distribution, subtracting the number of fixed effect parameters from the number of observations to provide degrees of freedom (Baayen, 2008).

3.4 Conclusion

In this chapter we introduced the MTC which will be a source of data for analyses in Chapters 5 & 7. The MTC was designed with the intention of providing a body of spontaneously elicited spoken dialogue which would be of empirical value to a range of researchers, with a range of interests, while retaining sufficient control of context to allow generalisable conclusions to be drawn. However, the fact that the MTC is a corpus of spontaneous speech results in data which poses problems for many statistical techniques. We subsequently introduced mixed-effects regression, Box-Cox transformations, the generalized linear model, and approaches for dealing with multicollinearity, which help us to solve many of these problems.

In Chapters 5 & 7 we apply these tools to testing the hesitation-as-signal hypothesis, using corpus and experimental data, and Wilson and Wilson's (2005) theory of turn-taking, using data from the MTC, respectively. Prior to that, in Chapter 4, we will experimentally investigate the effects on disfluent repetitions on listeners. While it has been suggested that, from a production perspective, filled pauses and repetitions are functionally similar (Clark & Wasow, 1998) there has been little evidence to show that repetitions have similar effects on listeners' linguistic processing as filled pauses. In the following chapter we will investigate whether listeners' attention is modulated by hearing repetitions, as has been found to occur when hearing a filled pause (Collard, Corley, MacGregor, & Donaldson, 2008), and whether any such effect has a consequence for linguistic processing.

Chapter 4

Experiment 1: Do repetitions heighten attention?

4.1 Introduction

Kraljic and Brennan (2005) suggest that if a feature of speech is being designed by speakers for their audience then we would expect the audience to show some response to the feature. In Chapter 2 we presented evidence which suggests that the production of hesitations is associated with a speaker experiencing difficulty, and that listeners appear to show a sensitivity to this association (e.g. by predicting that an upcoming word will be difficult to name). Much of the research demonstrating that listeners are sensitive to hesitations has examined filled pauses; however, Clark and Wasow (1998) suggest that repetitions are similar to filled pauses in performing a communicative role. We might therefore expect listeners to show similar sensitivities to repetitions as they do to filled pauses. The current chapter presents an experiment investigating whether listeners' attention is affected by hearing a disfluent repetition, and whether this could have consequences for the ways in which they represent subsequent words. Before discussing the evidence suggesting that filled pauses heighten listeners' attention, we will first discuss two accounts of why it is that speakers come to produce repetitions and then introduce an experimental paradigm which has previously been used to investigate the relationship between language and attention, and which will be used in the experiment presented in this chapter.

4.1.1 Production of repetitions

In 2.4 we reviewed the findings of a considerable number of studies which have investigated the factors that may cause speakers to produce hesitations. Relatively little attention, however, has been given to why it is that we see a variety of different types of hesitations being produced when speakers encounter difficulty. One exception to this has been the case of repetitions, where several researchers have offered explanations for why people come to repeat parts of speech. Blackmer and Mitton (1991) have proposed that the articulator may possess an autonomous restart capability. If, during speech production, the articulator finishes producing material before earlier stages of planning have been able to finish preparation of subsequent material then the articulator may reproduce the just-produced material. In other words, if a person runs out of words to say aloud before the next word is ready then they may repeat the last word (or words) they said while they wait for the next word to be prepared (a similar idea is expressed in the EXPLAN theory, e.g. Howell & Au-Yeung, 2001; Howell & Au-Yeung, 2002). Of course, if a speaker is forced to delay continuing to speak because upcoming parts of an utterance are not yet ready then they could simply produce a silent pause. Blackmer and Mitton give relatively little attention to the reason that speakers might produce a repetition rather than a pause.

In Clark and Wasow's Commit-and-Restore model, the intention is to explain why speakers "fill" a pause with a repeated word. They offer two strategic reasons for a speaker to repeat a word when they encounter difficulty. Both of these reasons are broadly listener-oriented. In the first, a speaker may wish to produce syntactically complete constituents, perhaps because these are easier for the listener to parse than disrupted constituents. In the second, the speaker may wish to make a preliminary commitment to the utterance that they intend to produce, perhaps as a form of attempt-suppressing signal to stop the listener from interpreting the disruption as the end of the speaker's turn and an invitation to begin a new turn or just to account for their use of time (e.g. Clark, 1996).

Clark and Wasow also offer a third reason for repeating a word, where the repetition results from difficulty arising from planning syntactically complex utterances. In this explanation, the repetition would appear to be symptomatic of the difficulty that the complexity induces; however, it is still not clear why difficulty would result in a repetition.[1] Increased syntactic complexity has also been shown previously to be related to the production of other types of hesitations (e.g. Ferreira, 1991; Watanabe et al., 2004). If speakers who are experiencing difficulty

[1] Although Clark and Wasow (1998) do not make the link themselves, it is possible that this is akin to the autonomous restart capability. If the burden of preparing a syntactically complex utterance leads to a slow down of planning then this may cause the articulator to run out of material prematurely. Of course, this still leaves us unable to explain *why* the speaker produces a repetition, as opposed to, for example, producing a filled pause or remaining silent.

due to syntactic complexity have options for the types of hesitations that they will produce, and they produce a repetition then, according to Clark's (1996) principle of choice, that choice should be a signal. It is not clear that Clark and Wasow intend their reasons to be mutually exclusive, and it may be the case that syntactic complexity is one of the causes of disruptions which speakers produce repetitions in order to strategically "manage". For example, a speaker who realises that they will have to delay due to syntactic complexity may make a preliminary commitment to the utterance, as is suggested is the second reason for producing a repetition, in order to buy time while they plan the utterance.

It is in the second reason where we see the suggestion of a functional similarity between repetitions and filled pauses. In either case, a speaker realises that they are unable to proceed fluently and instead either produces a repetition or a filled pause in order to justify the time that they are using. If speakers are producing repetitions for the same reasons as they produce filled pauses then we might reasonably expect that some of the effects that filled pauses have been observed to have on listeners (such as those effects on comprehension reviewed in 2.5.2) should also be observed when a listener hears a repetition. While, as we will see, several studies allow us to make comparisons between the effects of filled pauses and repetitions, one question that has received relatively little attention is whether the heightening of attention, and its subsequent consequences for linguistic processing, that has been observed to occur when listeners hear a filled pause can also occur when they hear a repetition. Before going on to discuss several studies investigating the effects of hesitations on listeners, we will first introduce an experimental paradigm which has previously been used to investigate the relationship between language and attention, and which will be used in the experiment presented in this chapter.

4.1.2 Change detection paradigm

Although predominantly used by researchers interested in visual cognition, there has recently been a growth in the use of the change detection paradigm in psycholinguistics. In studies of visual cognition, the paradigm frequently involves participants inspecting a visual scene in which changes occur. Participants see the visual scene twice, with, for example, a blank screen between each presentation. They are instructed to report when they have observed that a change has occurred (i.e. when something has changed in the visual scene between the first and second presentation). By manipulating the circumstances in which a change occurs, or the type of change that occurs, researchers are able to investigate factors which facilitate or impair the ability to detect changes. A theme that has emerged from research in this area is that the ability to detect a change is greatest when attention is drawn to the item that changes, and, inversely, that changes are more frequently missed when the item is not receiving attention (e.g. O'Regan, Deubel, Clark, & Rensink, 2000; Rensink, O'Regan, & Clark, 1997, 2000).

In this section we will review the use of the paradigm in investigations of linguistic processing (see Collard, 2009, for a fuller review of the insights into visual processing gained from the paradigm). In a typical linguistic version of the change detection paradigm a participant reads or hears two passages (occasionally—as in the experiment presented in this chapter—reading one and hearing the other) which are identical in all but one respect: One word that appears in the first presentation is replaced by a new word in the second presentation.

In an early example of the use of a linguistic version of the paradigm, Sachs (1967) had participants listen to target sentences which appeared within a larger discourse. The target sentence was then repeated either immediately after the target sentence or following some intervening material. The repeated version of the sentence contained either a semantic change (e.g. an individual mentioned in the discourse changing from the sender to the receiver of a letter) or a syntactic change (e.g. switching between active and passive constructions), or it appeared unchanged. When the second presentation immediately followed the first presentation, participants were able to accurately detect whether or not a change had occurred for all conditions in over 80% of trials. As the length of the intervening material increased, accuracy decreased to near chance levels for all conditions except for the semantic change (which decreased to just below 80%). On the basis of these findings, Sachs suggested that while the meaning of a sentence is retained in memory its surface form (including syntactic structure) is quickly discarded (a claim later supported by Jarvella, 1971).

Ferreira and colleagues (Ferreira, Bailey, & Ferraro, 2002; Ferreira & Patson, 2007) have suggested that the depth to which linguistic input is represented varies according to what is sufficient for a person's current purposes. Using the change detection paradigm, Sturt, Sanford, Stewart, and Dawydiak (2004) investigated the effect of linguistic focus on the specificity of linguistic representations. Focusing has been shown to reduce the "Moses illusion" (Erickson & Mattson, 1981), where participants incorrectly report that the statement *Moses put two of* *each sort of animal on the Ark* is true. If Moses is focused in an *it*-cleft, *It was Moses that put two of each sort of animal on the Ark*, then participants are more likely to recognise the anomaly and report that it is false (Bredart & Modolo, 1988).

Sturt et al. (2004) had participants read passages such as (7), which include a change to the second presentation where one word (e.g. *cider*) is replaced with a word that was either semantically related (*beer*) or unrelated (*music*). Where the change is semantically greater (e.g. *cider* → *music*) we would expect detection rates to be higher than when the change is between two semantically related words (e.g. *cider* → *beer*); however, the size of this difference would be expected to decrease if the target word had been represented in such detail as to include the differences between cider and beer. Passages contained either (7a) a *wh*-cleft (also known as a pseudo-cleft) or (7b) an *it*-cleft, which would place the focus on either the changed word or a word that was not changed, respectively.

(7a) Everyone had a good time at the pub. A group of friends had met up there for a stag night. What Jamie really liked was the cider, apparently.

(7b) Everyone had a good time at the pub. A group of friends had met up there for a stag night. It was Jamie who really liked the cider, apparently.

As expected, the type of cleft (*wh* or *it*) had no effect on detection of changes to unrelated words (approximately 95% for both clefts). Critically, however, participants were more accurate with the harder-to-detect changes to related words when the passage contained a *wh*-cleft, which focused the target word.

An alternative explanation for these results is that rather than having a direct effect on the granularity of semantic representations, focus improves detection of close changes by leading readers to spend more time looking at the critical word. Birch and Rayner (1997) observed that participants were slower to read a word that was focused in a sentence than the same word unfocused in a semantically similar sentence (although see Morris & Folk, 1998, who report the reverse pattern).

In order to determine whether the results reported by Sturt et al. (2004) were due to focus changing the depth of semantic representation or merely leading readers to spend more time looking at the critical word, Ward and Sturt (2007) recorded participants' eye movements whilst reading each presentation of the passages in a change detection experiment where the critical word could only change to a synonym. In contrast to Birch and Rayner (1997) and Morris and Folk (1998), for the first presentation there was no evidence that focus had any effect on eye movements during reading. When a change had occurred, participants fixated for longer (with significant differences for first fixation, first pass and total gaze duration), and more often, on the critical word when it had changed. When the target region was expanded to include preceding function words, significant interactions for first pass and total gaze durations were found between whether a change had occurred and whether the critical word was focused. In summary, while there was no evidence of an effect of focus on eye movements during the first presentation, in the second presentation focus was found to modulate the effects of change. Taken together, this suggests that the beneficial effects of focus on change detection (also observed in Ward & Sturt, 2007) were not due to focus leading readers to spend longer reading the critical word.

Sanford, Sanford, Molle, and Emmott (2006) suggest that the clefting which directs linguistic focus is a type of attention-orienting device. They suggest other examples of such devices, including the italicisation of text and prosodic stress in spoken language. For both of these examples they find the same benefit for semantically close changes observed with clefting by Sturt et al. (2004).

Sanford, Sanford, Filik, and Molle (2005) used the change-detection paradigm to investigate whether the representations of words are shallower when sentential load is increased (e.g. by increasing syntactic or referential complexity). In one experiment, participants read passages such as (8) where an anaphor in the second sentence was either a noun-phrase (*the student*) or a pronoun (*I*), while the underlined verb was changed to one that was either semantically close (*seen*) or semantically distant (*missed*).

(8) The college frequently held social functions for visiting academics.
The professor who {the student/I} had recently <u>met</u> at the party
was famous, but no one could figure out why.

As a first-person pronoun should be highly accessible to the reader (e.g. Gundel, Hedberg, & Zacharski, 1993) it would be expected to result in lower sentential load than a noun phrase anaphor appearing in a similar sentence (Warren & Gibson, 2002). Sanford et al. (2005) replicated Sturt et al.'s (2004) semantic distance effect: participants were less able to detect a change when it was semantically close. When a passage contained a noun phrase anaphor, rather than a pronoun, detection rates were found to decrease further, suggesting that a consequence of greater sentential load is that words in the sentence can not be represented in as fine a detail as they might if the load was reduced. Both of these effects were also observed with auditory presentations of the passage.

4.1.3 Hesitations and attention

Fox Tree (2001) was among the first to suggest that certain hesitations may modulate listeners' attention. She found that when participants were instructed to listen to recordings of speech and press a button when they heard a specific word, they were faster to respond when the target word was immediately preceded by the filled pause "uh". This facilitory effect of filled pauses was explained by suggesting that they heighten attention to upcoming speech, which leads participants to be faster to recognise the target word.

Fox Tree's claim was tested by Collard (2009; Collard et al., 2008) using Event Related Potentials (ERP), in addition to behavioural methods. Attention is frequently investigated in studies of ERP using the oddball paradigm. In the oddball paradigm participants are presented with a series of stimuli which are identical in some respect (e.g. beeps of the same pitch). Occasionally, the stimuli deviate from this series (e.g. by being of a higher pitch). When EEG is recorded during presentation the deviation leads to the observation of the mismatch negativity (MMN) and P300 components. The MMN is an early occurring (100–250ms following the deviant stimulus) component which is associated with detecting acoustic change. The P300 occurs slightly later than the MMN (with the amplitude peaking around 300ms following the deviant stimulus). While the MMN is found only for auditory stimuli, the P300 reflects a more general process of reorientating of attention towards a deviant stimulus.

Collard et al. (2008) had participants listen to the high-cloze versions of sentences taken from Corley et al.'s (2007) study (an example is given in 9). In half of the sentences that participants heard, the final word had been compressed, resulting in a poor, telephone-like quality. Deviant endings, such as these, should result in the MMN and P300. Final words either appeared in a fluent context or were preceded by a filled pause. In the fluent conditions, the deviant word was associated with an MMN and a P300, as would be expected for an acoustic oddball. However, in the filled pause condition, while the MMN remained, the P300 was eliminated (or reduced, at posterior sites on the scalp). Collard and colleagues used these findings to claim that upon hearing a filled pause attention is heightened to upcoming linguistic material. Such heightening subsequently eliminates the need to reorient attention when deviant linguistic stimuli is encountered.

(9) Everyone's got bad habits and mine is biting my nails

If the attention given to words that follow a filled pause is heightened, compared to words appearing in a fluent context, then—as is the case with linguistic focus—we might expect listeners to find it easier to detect when the word has changed. Evidence supporting this hypothesis comes from a series of experiments using a change detection paradigm (Collard, 2009). In Collard's version of the paradigm the first presentation of a passage was auditory, while the second presentation consisted of participants reading a transcript of what was said. Using Ward and Sturt's (2007) non-focused passages, Collard manipulated whether or not participants heard "uh" before the target word. While no effect of disfluency was found for semantically distant changes, participants were more accurate at detecting a semantically close change when the target word had been preceded by a filled pause. This pattern of results matches those found with linguistic focus (Sturt et al., 2004; Ward & Sturt, 2007).

4.1.4 Repetitions, attention and language

Recent years have seen a growth in the number of studies investigating the effects of hearing hesitation on cognitive processes, particularly those related to language comprehension. In Chapter 2 and the previous section, we reviewed evidence suggesting that the presence of a hesitation may affect the ways in which linguistic material is represented and understood. While studies of filled and silent pauses have been relatively common, few studies have investigated the effects of repetitions on listeners, with those that have finding mixed results.

Fox Tree (1995) used a word monitoring task (similar to that later used in Fox Tree, 2001) to investigate whether hearing a repetition had any effect on participants' abilities to recognise target words. Examples of repetitions were taken from a corpus of spontaneously produced Dutch speech. In half of the utterances that participants heard the second mention of the repetition was excised leaving a silent pause. Participants were quicker to identify the word when it had been recently preceded by a repetition than with a pause. In a second experiment, where the silent pause was eliminated, the benefit of repetitions remained. On the basis of these two experiments we might conclude that repetitions facilitated monitoring; however, an alternative explanation is that the disruptions produced by editing were instead hampering monitoring. Evidence in support of this second account comes from a further two experiments where repetitions were created (by repeating an existing part of the utterance). No benefits were found for monitoring when participants heard these artificial repetitions, suggesting that the monitoring differences observed in the first two experiments were due to editing rather than the repetitions that were being edited out.

Recall from Chapter 2 that Corley and colleagues (2007; MacGregor et al., 2010) found that the N400 component associated with processing an unpredictable word was reduced when that word had been preceded by a filled or silent pause. MacGregor et al. (2009) investigated whether a similar effect occurred when the unpredictable word was preceded by a repetition (for example 10).

(10) Everyone's got bad habits and mine is biting [my] my tongue

When participants heard the unpredictable *tongue* an N400 was observed, regardless of whether or not it had been preceded by a repetition. The presence of a repetition appeared to have no effect on participants' processing of the word (and a subsequent test found participants were no more likely to recall words that had been preceded by a repetition – in contrast to the recall benefits found for both filled and silent pauses); however, an effect of the repetition was found in an earlier time window (100-400ms), and an observed P600 suggested that the repetition was having a disruptive influence on syntactic processing. In sum, MacGregor et al.'s findings suggest that while listeners are sensitive to hearing a repetition it may not have an effect on the integration of a subsequent word.

While it has been suggested that, from the standpoint of production, filled pauses and repetitions share a similar function of helping speakers account for disruptions to linguistic acts (Clark, 1996; Clark & Wasow, 1998; Hieke, 1981) we have seen two examples where the beneficial consequences of hearing a filled pause on language comprehension do not appear to occur for repetitions. Instead, evidence from MacGregor et al. (the P600) suggests that repetitions may have a detrimental effect on listeners' syntactic processing. In their Commit-and-Restore model, Clark and Wasow suggest that there are two strategic motivations for producing a repetition, to maintain the continuity of a syntactic constituent in order to facilitate parsing and to make a preliminary commitment to a constituent in order to keep hold of a conversational turn. The P600 observed by MacGregor et al. would seem to be inconsistent with the first strategic function: Repetitions do not facilitate syntactic processing; rather, they would seem to disrupt it.

One would assume that if speakers are designing repetitions for their audience, as Clark (1996, Clark & Wasow, 1998) argues, then there ought to be some benefit for the listener of hearing a repetition. In Experiment 1, we used the change detection paradigm to investigate whether repetitions had any effect on the depth of semantic representations, similar to the effect found by Collard (2009) for filled pauses. If the functional similarity between filled pauses and repetitions extends to comprehension then we would expect that encountering a repetition would lead to a heightening of attention. Within the paradigm, this would manifest as an improvement in detection of semantically close changes. If the predicted effect was observed then an alternative explanation could be that it is not the phonological form of the repetition driving the effect but rather the delay that the repetition provides (for example, by allowing time to finish processing of previous material before encountering the subsequent word). To allow us to rule out this account we also considered silent pauses in addition to fluent passages and repetition passages. By matching the duration of these pauses to the duration of the delay provided by the repetition we could be sure whether any observed results were due to the phonological form of the repetition or to the delay it provides.

Collard (2009) had participants perform the change detection task while their eye movements were being recorded in order to investigate whether filled pauses functioned similarly to linguistic focus. Comparison between his results and those obtained by Ward and Sturt (2007) could have provided evidence that similar processes underlie both effects. While Collard did not observe any effects of filled pauses on any of the four reading measures investigated by Ward & Sturt (first fixation and pass duration, total gaze duration and number of fixations), he did find that participants were more likely to regress back to a word which had been preceded by a filled pause. In the present study we also recorded participant's eye-movements, not only for comparison with Ward & Sturt's results but also to explore whether the patterns observed with repetitions were similar to those that Collard observed for filled pauses.

4.2 Methodology

4.2.1 Participants

Thirty-six native British English speaking undergraduates volunteered to participate in this experiment. All reported having no known speech, language or hearing disorders; and normal, or corrected-to-normal, vision.

4.2.2 Materials

Forty-five three-sentence passages were used in the experiment (see 11 for an example, and Appendix A for the full list). Thirty-six passages were taken from Collard (2009, experiment 2). To meet the requirements of our design an additional nine passages were adapted from Sanford et al. (2005). As the Sanford et al. passages consisted of only two sentences, a third sentence was added for each which did not introduce new referents or change topic.

(11) The doctor checked to see how much longer he had to work. He saw that the patient with the **virus / infection / tissue** was at the front of the queue. A kind but strict-looking nurse brought the boy in.

For each passage, two factors were manipulated. Firstly, whether or not a target word had been changed in the second presentation and the semantic distance of the new word (no change, e.g. *virus*; close change, *infection*; distant change, *tissue*), where a change had occurred. Secondly, whether or not the target word was immediately preceded by a disfluency (fluent, preceded by a pause, preceded by a repetition of a function word).

In the two change conditions a noun in the second sentence of each passage was changed to one of two similarly plausible nouns. The noun was always immediately preceded by a function word. The new word was one that was either closely semantically related to the old word, for example a synonym (a close change), or that was less semantically related (a distant change). Frequency information for each target word was taken from the British National Corpus (1995). Log transformed frequencies were found not to significantly differ between conditions ($F(2, 44) < 1$). A comparison of the number of characters in each word also found no significant differences for word length between conditions ($F(2, 44) < 1$).

An additional twenty-seven filler passages were used. For one third of the fillers no change occurred, in the second third the change occurred in the first sentence, and in the final third the change occurred in the third sentence.

A native speaker of British English was recorded producing each of these passages. The speaker was instructed to produce the passage as naturally as possible and any accidental hesitations appearing in the first or third sentences were allowed so long as they did not lead to a mismatch between recording and transcription (for example, silent pauses were accepted, but repetitions or repairs were not). For each passage, the speaker produced one fluent recital, one containing a silent pause, and one containing a repetition. When producing the silent pauses and repetitions, the speaker was instructed to attempt to make the disfluency sound as natural as possible. For pause recitals, the pause appeared between the function word and the target word. For repetition recitals, the function word immediately preceding the critical word (typically a determiner) was repeated once. The speaker was instructed to produce a pause of comfortable duration. These were subsequently edited to match the interval between the offset of the first mention of a repetition and the onset of the target word in the repetition version of each passage. The mean duration of pauses for all items was 253 ms ($SD = 77$).[2]

One third of each type of filler passage (i.e. a third of each of the no change, first sentence change and third sentence change fillers) was fluent. In another third, the speaker produced a prolongation or filled pause in the first sentence. In the final third, this hesitation appeared in the third sentence. Where possible, we used incidences where the speaker was genuinely disfluent during the recording of these filler passages. All recordings were stored as mono 48kHz .wav files.

Participants heard fifteen fluent passages, fifteen pause passages, and fifteen repetition passages. Transcriptions were divided using a Latin square method, with the participant seeing five no change, five close change and five distant change transcriptions for each level of fluency.

[2] This may appear rather short for silent pauses; however, pause durations were not normally distributed, and 86.67% were greater than or equal to the 180 ms cut-off for silent pauses suggested by Hieke et al. (1983).

4.2.3 Procedure

Eye movement recordings were made with a SR Research EyeLink 1000 eye-tracker, sampling at 500Hz. Participants were informed that they would hear a passage, and then read a transcription which, in some cases, would include the addition of a new word which would replace a word said by the speaker. They were then instructed that, when prompted, they should decide whether such a change had taken place.

At the beginning of each trial a square appeared on the screen, located where the first character of the transcription would subsequently appear. Participants were instructed that they should gaze at the square to trigger playback of the recording, and should continue to look at it while the recording was playing.

A transcription appeared on the screen 500ms after the ending of the recording. The transcription was given in a 22 pt sans-serif typeface (Arial), and was presented as black text on a white screen. Participants were instructed to press a button when they had finished reading the passage. They were then asked if they had detected a change. Upon pressing a button to indicate that they had detected a change, nine words appeared on the screen as candidates for the replacement word. For each of the three sentences in the passage three words were selected as candidates, and were randomly assigned to one of nine regions of the screen. Participants were instructed to look at the word they believed had been added and press a button to select it.

The experiment consisted of three practice trials, where participants had to the chance to familiarise themselves with the procedure while also receiving feedback on their performance, followed by seventy-two experimental trials in six blocks of twelve. At the beginning of each block an SR Research nine point calibration routine was followed. The experiment lasted approximately 45 minutes.

4.2.4 Data analysis

Full mixed-effects models were used for analysing both behavioural responses and eye-tracking measures. Models included fixed effects for disfluency (fluent, pause and repetition) and change condition (close and distant), with random effects for participants and items.

We deviated from the approach to analysing experimental data that was described in Chapter 3 in only one respect: Treatment coding was used for both factors tested. This was done to allow us to make the comparisons critical to our hypotheses. The intercept in our analyses corresponded to the fluent, close change, condition. A simple effect of distance suggests an effect of semantic distance with fluent utterances, while a simple effect of either pause or repetition suggests effects of hesitation where a close change occurred. Interactions between these two sets of factors would suggest that there are effects of disfluency when a distant change occurred.

In order to deconfound any possible effects of repetitions from the delays that they provide, in any analysis where effects of both pauses and repetitions were found we intended to carry out a second analysis which excluded fluent utterances. Doing so allowed for a direct comparison between pauses and repetitions.

In all trials, the critical word appeared on the same line as a short, immediately preceding, function word. If a participant initially fixated on the function word then they may be have been able to detect that a change had occurred before fixating on the critical word itself (see Rayner, Well, & Pollatsek, 1980; Rayner, Well, Pollatsek, & Bertera, 1982). To take into account the possibility that participants may process the critical word before fixating on it we included the function word in our target region. As earlier noted, Ward and Sturt (2007) similarly expanded the size of their target region to include preceding function words. Trials were excluded where the participant did not fixate on either word in this region. This led to the exclusion of 28 trials, 0.03% of observations.

To allow for direct comparison with his results we calculated the five reading measures used by Collard (2009) in his investigation of the effects of filled pauses on attention. *First fixation duration* is the time spent by the participant on their first fixation within the target region. *First pass duration* is the sum of the duration of all fixations occurring in the target region before the participant leaves the region for the first time. *Total time* is the sum of the duration of all fixations occurring in the target region during the entirety of the trial. This includes fixations that have occurred when the participant fixates back on the region after looking elsewhere. *Number of fixations* are the total number of fixations occurring in the target region throughout the trial. Finally, *probability of regression back into region* is the probability that the participant will return to the target region after previously fixation on it and subsequently fixating elsewhere (i.e. the probability of making a second pass).

In line with Collard (2009) and Ward and Sturt (2007), our analyses of eye-tracking measures included those trials where a participant failed to correctly recognise that a change had occurred. In both linguistic (Ward & Sturt, 2007) and non-linguistic (Hollingworth & Henderson, 2002) forms of the change-detection paradigm participants' eye-movements have been found to be sensitive to change manipulations regardless of whether they subsequently reported detecting the change. For each measure we eliminated any observations that were over 2.5 SD from the respective per-condition mean.

4.3 Results

We first present the results of the change detection task itself (i.e. how accurate were participants at detecting when a change had occurred), before presenting analyses of participants' eye movements while performing the task. Unless otherwise noted, each analysis tested semantic distance of the change (close change, distant change) by fluency (fluent, pause, repetition).

4.3.1 Behavioural results

Participants correctly recognised when no change had occurred in 91.3% of trials. A logistic mixed-effects regression, with only pause and repetition as fixed effects, revealed that accuracy in the "no change" condition was not influenced by the fluency of the passage (for pause, p = both *ps* < 1). Full results of participants' accuracy by condition are shown in Figure 4.1.

The model of participants' accuracy at detecting when a change had occurred is given in Table 4.1. In trials where a change had occurred, participants correctly detected the change in 63.1% of trials. As changes occurred in two thirds of trials, and successfully registering that a change had occurred required not only detecting a change but also subsequently identifying which of the nine candidate words was the replacement, participants had a $\frac{1}{27}$ $(\frac{2}{3}\times\frac{1}{2}\times\frac{1}{9})$ chance of successfully detecting a change due to chance alone.

In the fluent condition, participants were no more likely to correctly detect when a change had occurred when it was a near change than a distant change ($p = .64$). A marginal effect of pause was observed ($p < .1$), suggesting that when a close change had occurred participants may have been more likely to correctly detect the change when it was preceded by a pause than when it appeared in a fluent context. Furthermore, pause was found to interact with distance ($p < .05$), suggesting that participants were almost two and a half times as likely to detect when a distant change had occurred when it had been preceded by a pause as when it occurred in a fluent context ($e^{0.894} = 2.44$). No effects of repetition were found, suggesting that repetitions did not have an effect on accuracy in either the close or distant change conditions ($p = .86$ and $p = .15$, respectively).

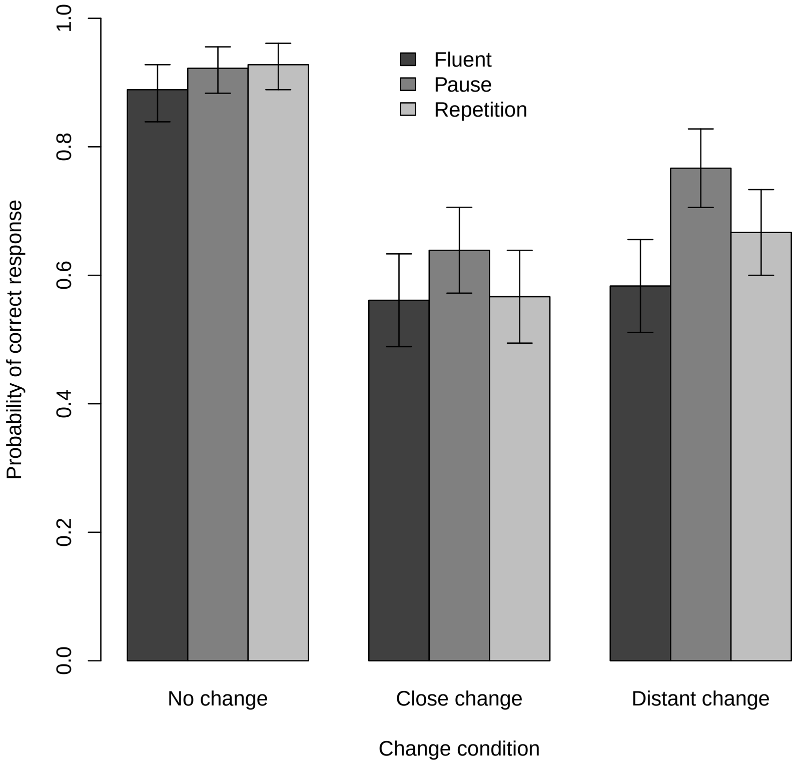


Figure 4.1: Experiment 1: Mean probability of correctly recognising whether or not a change had occurred by condition. Dotted line represents the probability of registering a correct response when a change had occurred due to chance alone (when no change had occurred the probability of responding correctly due to chance was 50%). Bars represent 95% confidence intervals estimated using bootstrap resampling (999 runs).

4.3.2 First fixation duration

First fixation duration for each condition are shown in Figure 4.2. In the fluent condition, no effect of the semantic distance of the change was found ($p = .58$). No effects were found of either pauses or repetitions in the close change condition

Table 4.1: Experiment 1: Logistic mixed effects model of the probability of participants correctly detecting when a change had occurred.

					Random effects		
Fixed effect	β	SE	z	$p(\beta = 0)$	Group	Predictor	Variance
Intercept	0.324	0.274	1.179	.24	Item	*Intercept*	1.650
Distant	0.131	0.284	0.461	.64		Distant	0.597
Pause	0.458	0.251	1.821	.07		Pause	< 0.001
Repetition	0.046	0.263	0.174	.86		Repetition	< 0.001
Distance × Pause	0.894	0.373	2.398	$< .05$		Distance × Pause	< 0.001
Distant × Repetition	0.516	0.355	1.454	.15		Distant × Repetition	< 0.001
					Participant	*Intercept*	0.324
						Distant	0.239
						Pause	0.076
						Repetition	0.387
						Distance × Pause	< 0.001
						Distant × Repetition	< 0.001

($p = .14$ and $p = .80$, respectively), nor were any interactions observed between these conditions and change ($p = .42$ and $p = .42$, respectively). Results of the analysis are given in Table 4.2.

4.3.3 First pass duration

First pass duration for each condition are shown in Figure 4.3. In the fluent condition, no difference was observed between semantically near and distant changes ($p = .43$). No effects were found for either pauses or repetitions in the close change condition ($p = .53$ and $p = .75$, respectively), nor were any interactions observed between these conditions and change ($p = .81$ and $p = .48$, respectively). Results of the analysis are given in Table 4.3.

4.3.4 Total gaze duration

Total gaze durations for each condition are shown in Figure 4.4. None of the fixed effects tested in this model were found to reach the threshold for statistical significance. Results of the analysis are given in Table 4.4.

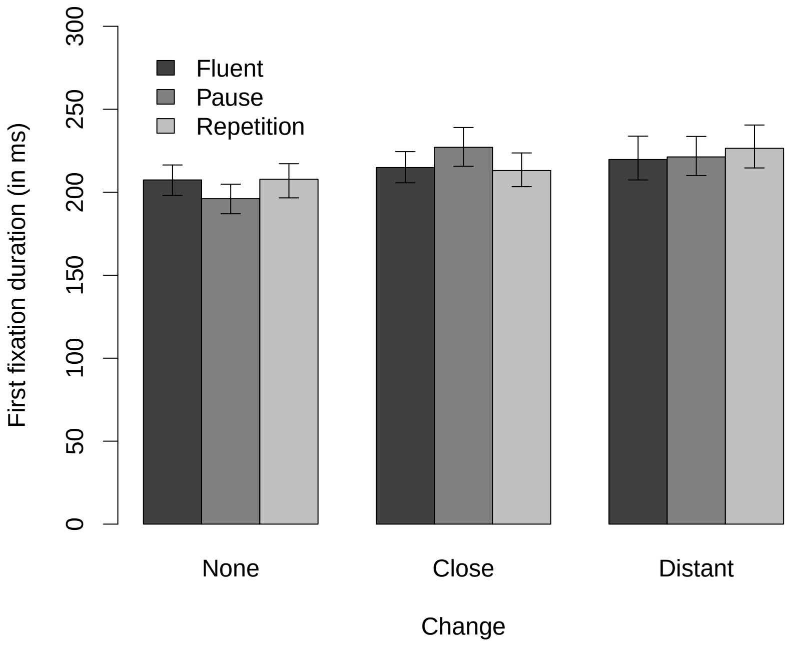


Figure 4.2: Experiment 1: Mean first fixation duration on target region by condition. Bars represent 95% confidence intervals estimated using bootstrap resampling (999 runs).

4.3.5 Number of fixations

Mean numbers of fixations for each condition are shown in Figure 4.5. Again, none of the fixed effects tested in this model were observed to reach the threshold for statistical significance. Results of the analysis are given in Table 4.5.

4.3.6 Probability of regression back into region

The probability of regression back into the target region for each condition are shown in Figure 4.6. We did not find that any of the fixed effects tested in this model reached the level of significance, however a marginally significant interaction was observed between distance and pause ($p < .1$). This may suggest that in the semantically distant change condition participants were less likely to regress back to the target word when it had been preceded by a pause than when it appeared in a fluent context. Results of the analysis are given in Table 4.6.

Table 4.2: Experiment 1: Linear mixed effects model of first fixation duration in the target region by condition.

					Random effects		
Fixed effect	β	SE	t	$p(\beta = 0)$	Group	Predictor	Variance
Intercept	214.863	6.981	30.779	$< .001$	Item	*Intercept*	< 1
Distant	4.574	8.364	0.547	.58		Distant	225
Pause	11.994	8.077	1.485	.14		Pause	< 1
Repetition	−2.149	8.321	−0.258	.80		Repetition	< 1
Distance × Pause	−9.570	11.811	−0.810	.42		Distance × Pause	42
Distant × Repetition	9.242	11.367	0.813	.42		Distant × Repetition	< 1
					Participant	*Intercept*	598
						Distant	< 1
						Pause	< 1
						Repetition	178
						Distance × Pause	307
						Distant × Repetition	< 1
					Residual		5520

Table 4.3: Experiment 1: Linear mixed effects model of first pass durations in the target region by condition.

					Random effects		
Fixed effect	β	SE	t	$p(\beta = 0)$	Group	Predictor	Variance
Intercept	296.728	13.046	22.745	$< .001$	Item	*Intercept*	331
Distant	14.170	17.787	0.797	.43		Distant	3310
Pause	−10.381	16.386	−0.634	.53		Pause	< 1
Repetition	−5.095	15.868	−0.321	.75		Repetition	< 1
Distance × Pause	5.263	22.116	0.238	.81		Distance × Pause	326
Distant × Repetition	15.960	22.796	0.700	.48		Distant × Repetition	< 1
					Participant	*Intercept*	1540
						Distant	< 1
						Pause	952
						Repetition	403
						Distance × Pause	< 1
						Distant × Repetition	1390
					Residual		20300

4.3.7 Non-change condition

In his own studies using the change detection paradigm, Collard (2009) raised the possibility that any effects of filled pauses that he observed could be general effects of hearing a hesitation on cognition, rather than a specific effect of filled pauses on attention. In order to eliminate this possibility, he examined eye

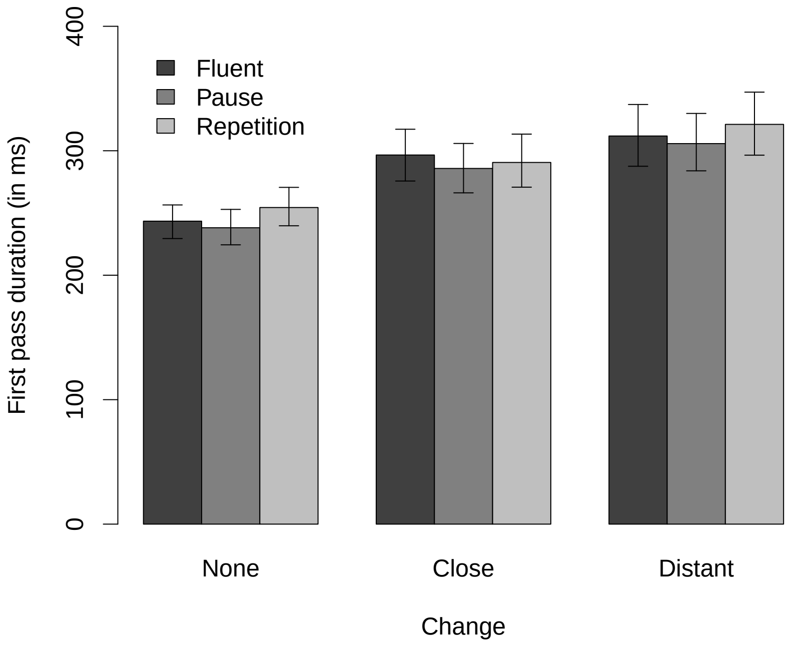


Figure 4.3: Experiment 1: Mean first pass duration in the target region by condition. Bars represent 95% confidence intervals estimated using bootstrap resampling (999 runs).

movement measures in the no-change condition, where the absence of a significant effect of filled pauses may suggest that there is no such general effect. As it was not clear in our experiment whether repetitions were having *any* effect on participants, we also examined all measures in the no-change condition for evidence of differences between repetitions, and fluent or pause conditions.

For total time, number of fixations and regressions back into the target region, no effects of fluency condition were found. For first fixations, a marginally significant effect of pause was found ($\beta = -11.509$, $t = -1.69$, $p < .1$), suggesting that when the target word was preceded by a pause participants may have spent less time initially fixating upon it. For the total time, a significant effect of repetition was found ($\beta = 83.090$, $t = 2.146$, $p < .05$). This would suggest that participants spent more time, overall, looking at the target word when it was preceded by a repetition that when it appeared in a fluent context.

4.4 Discussion

In the present experiment we did not find any evidence that the presence of a disfluent repetition had any effect on accuracy in the change detection task –

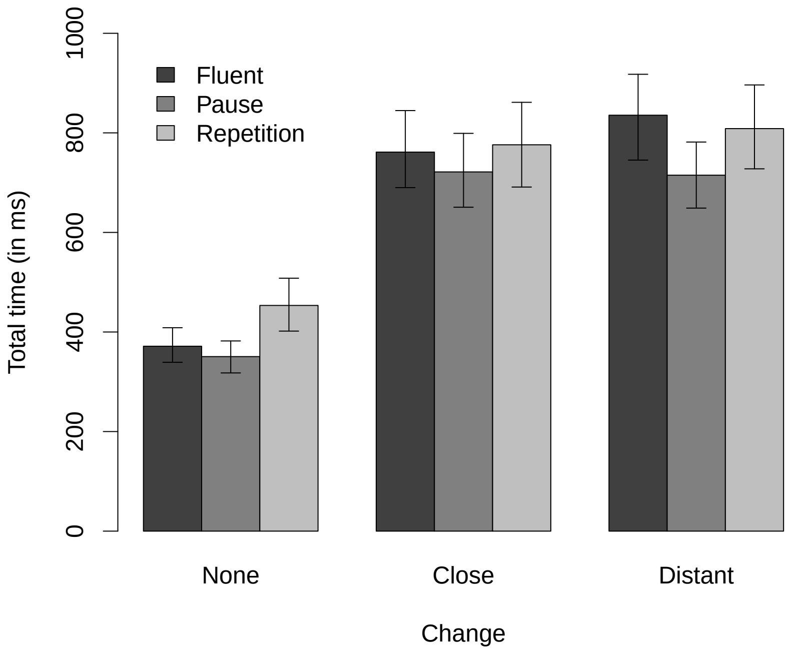


Figure 4.4: Experiment 1: Mean total gaze duration for the target region by condition. Bars represent 95% confidence intervals estimated using bootstrap resampling (999 runs).

either overall, or by increasing detection rates for semantically close changes. Our results would suggest that repetitions do not modulate the depth of semantic representations in the same way as filled pauses appear to do (Collard, 2009). Furthermore, we found very little evidence that having heard the disfluent repetition of a function word had any effect on the eye movements made while reading the content word that immediately followed it. While Collard found that participants were more likely to regress back to words that had been preceded by filled pauses, we found no such pattern for words that had been preceded by repetitions, although participants did spend more time looking at this word overall in the no-change condition. Prior to discussing repetitions, we will first discuss silent pauses, which do appear to have an effect on the depth of semantic representations.

4.4.1 Effects of pauses

Silent pauses were included in the present experiment with the intention that they would allow us to deconfound effects of repetitions from effects of the delays that repetitions provide. Our rationale was that while repetitions could be influencing linguistic processing directly, an alternative explanation could be that they simply offer respite for participants to wrap up linguistic processing of the

Table 4.4: Experiment 1: Linear mixed effects model of total fixation durations on the target region by condition.

					Random effects		
Fixed effect	β	SE	t	$p(\beta = 0)$	Group	Predictor	Variance
Intercept	762.724	51.325	14.861	< .001	Item	*Intercept*	6140
Distant	81.523	56.645	1.439	.15		Distant	6670
Pause	−45.619	51.407	−0.887	.38		Pause	< 1
Repetition	26.578	52.905	0.502	.62		Repetition	1
Distance × Pause	−79.944	72.742	−1.099	.27		Distant × Repetition	19300
Distant × Repetition	−59.214	75.729	−0.782	.43	Participant	*Intercept*	42500
						Distant	14500
						Pause	< 1
						Repetition	4910
						Distance × Pause	< 1
						Distant × Repetition	< 1
					Residual		223000

Table 4.5: Experiment 1: Poisson mixed effects model of number of fixations in the target region.

					Random effects		
Fixed effect	β	SE	z	$p(\beta = 0)$	Group	Predictor	Variance
Intercept	1.140	0.058	19.688	< .001	Item	*Intercept*	0.014
Distant	0.032	0.064	0.499	.62		Distant	0.011
Pause	−0.102	0.066	−1.555	.12		Pause	0.007
Repetition	0.037	0.067	0.547	.58		Repetition	0.009
Distance × Pause	−0.034	0.088	−0.384	.70		Distance × Pause	0.013
Distant × Repetition	−0.071	0.088	−0.801	.42		Distant × Repetition	0.038
					Participant	*Intercept*	0.046
						Distant	0.012
						Pause	0.014
						Repetition	0.029
						Distance × Pause	< 0.001
						Distant × Repetition	< 0.001

previous word in anticipation of the subsequent word (similar to the *temporal delay hypothesis* of Corley & Hartsuiker, 2011). While we observed no effects of repetitions on change detection we did find that a delay was beneficial when changes were semantically distant.

Collard (2009) previously investigated the effect that delays have on change detection rates. He predicted that if the filled pause effects he observed were due

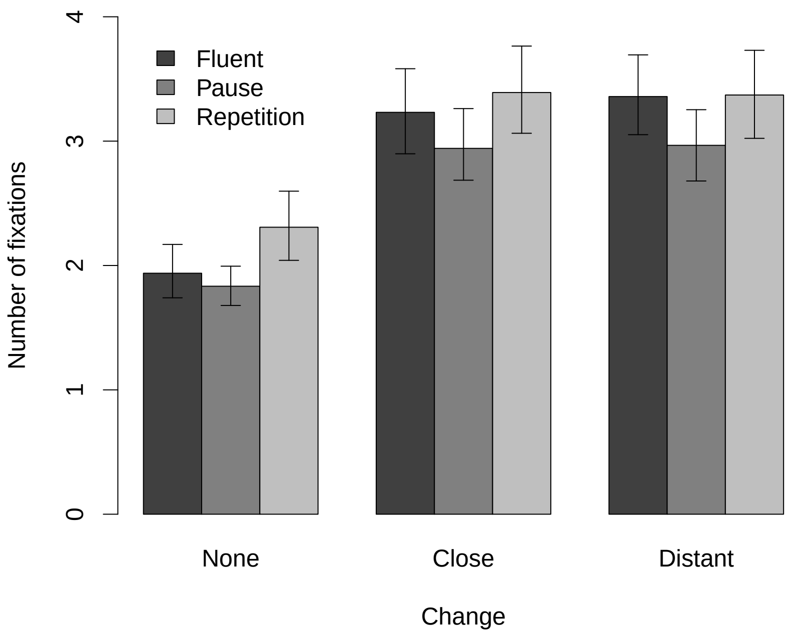


Figure 4.5: Experiment 1: Mean number of fixations in the target region by condition. Bars represent 95% confidence intervals estimated using bootstrap resampling (999 runs).

to the delays they provided then extending the duration of the delay may be even more beneficial. Where the lengths of delays between the filled pause and the target word were greater the benefit of the filled pause was actually found to disappear, with participants poorer at detecting semantically close changes compared to fluent passages. Collard used this to suggest that the phonological form of the filled pause has an important influence beyond the delay that the pause provides. The effects we observed with silent pauses may suggest that the delays do still provide a benefit which is somehow supplemented by the phonological form of the filler. Future research could compare silent pauses and filled pauses to determine if the benefit of a filled pause is greater than that of a duration matched silent pause.

In many linguistic change detection experiments the "attention capturer" (Sanford et al., 2006) is found to aid only the detection of semantically close changes. For semantically distant changes, the detection rates are typically very high and are insensitive to attentional manipulations (i.e. the presence or absence of an attention capturer). In the present experiment we found that when a distant change had occurred, detection accuracy was higher when the critical word was preceded by a silent pause. In the fluent condition, detection rates were particularly low relative to previous studies (both using spoken and written materials). A

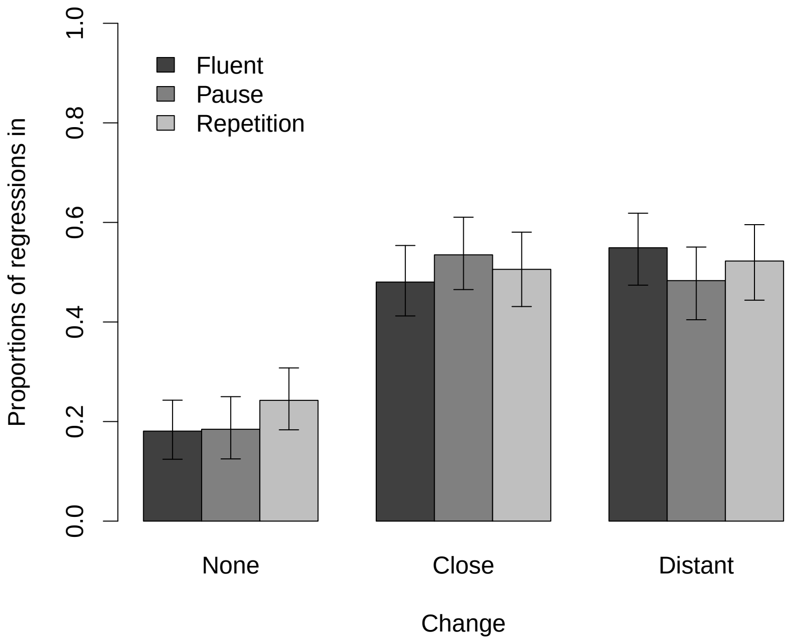


Figure 4.6: Experiment 1: Mean probability of returning to the target region after having earlier fixated and subsequently fixated elsewhere by condition. Bars represent 95% confidence intervals estimated using bootstrap resampling (999 runs).

marginally significant effect of silent pauses was found for close changes, similar to the standard finding in studies of this sort including the effect of filled pauses observed in several experiments by Collard (2009). That this effect narrowly failed to reach our criterion for significance ($\alpha = .05$) may suggest that our experiment lacked statistical power. Compared to Collard's studies, our design had an additional two cells, semantically close and distant changes in the pause condition. However, to compensate for this, we tested an increased number of participants (36 in the present study, compared to 24 in each of Collard's experiments). If the increase in participants in the present study was commensurate with the increase in the number of cells in the design, then an alternative explanation for a possible lack of statistical power would be that the effect size for silent pauses is smaller than that for filled pauses. If this is the case, then this would be consistent with Collard's suggestion that the phonological form of the filled pause has an influence beyond the delay that it provides.

4.4.2 Semantic distance effects

Examining the fluent conditions alone found no evidence that participants were better at detecting a semantically distant change than one that was semantically

Table 4.6: Experiment 1: Logistic mixed effects model of regressions back to the target region.

					Random effects		
Fixed effect	β	SE	z	$p(\beta = 0)$	Group	Predictor	Variance
Intercept	−0.092	0.196	−0.468	.64	Item	*Intercept*	0.209
Distant	0.328	0.224	1.467	.14		Distant	0.020
Pause	0.241	0.225	1.072	.28		Pause	0.048
Repetition	0.121	0.222	0.544	.59		Repetition	< 0.001
Distance × Pause	−0.554	0.315	−1.759	.08		Distance × Pause	< 0.001
Distant × Repetition	−0.254	0.314	−0.808	.42		Distant × Repetition	< 0.001
					Participant	*Intercept*	0.342
						Distant	< 0.001
						Pause	< 0.001
						Repetition	0.007
						Distance × Pause	< 0.001
						Distant × Repetition	< 0.001

close. This is in contrast to much of the previous literature (e.g. Collard, 2009; Sanford et al., 2005, 2006; Sturt et al., 2004). Although Collard did not make any direct statistical comparisons between close and distant changes, a visual survey of his means strongly suggests that distant change target words were fixated for longer and more often than close change target words. In the present experiment there was no evidence that semantic distance had an effect on any of our reading measures in fluent conditions.

It is not clear why our experiment did not replicate the semantic distance effect observed elsewhere. Our items were taken from two separate studies, which both found that semantically distant changes were more likely to be detected than semantically close changes. While there are differences in procedure between Collard's study and our own, for example the number of fillers (0 vs 24), whether sentences were presented on single lines or as part of a paragraph, and white text on a black background or black text on a white background, the wider literature offers no reason why these should have eliminated the semantic distance effect.[3] Detection rates for close and distant changes were approximately 57%. As this is well above the level expected by chance we do not believe that participants were merely guessing whether a change had occurred.

[3] Although some previous studies (e.g. Sanford et al., 2005, 2006; Ward & Sturt, 2007) do not describe their methodology in sufficient detail to be sure whether or not sentences were presented as paragraphs, or with each sentence presented on an individual line.

It may be that there is a certain level of attention required to detect any change which was only reached in our experiment when a silent pause was present. If this is the case, though, then it is not clear why our participants' attention was typically below this level when the participants in Collard's study (who were drawn from the student population of the same university, only a few years prior) were better able to detect semantically distant changes in the fluent condition. It is clear that further research is required to explain the absence of the semantic distance effect in our experiment.

4.4.3 Effects of repetitions

Our results add to a number of studies (e.g. Fox Tree, 1995; MacGregor et al., 2009) which did not find evidence to suggest that repetitions have an influence on linguistic processing that is similar to filled pauses. One possible explanation for why repetitions appear not to influence comprehension is that listeners may not always recognise when a repetition has occurred. Lickley (1995) found that participants who were asked to compare a disfluent recording of speech with a transcript which was lacking those disfluencies only recognised when a one word repetition had been removed approximately a third of the time. Participants were better able to detect when a filled pause had been removed (although still on only 55.2% of occasions). If filled pauses are more salient than repetitions then this could explain why the effects shown to occur with filled pauses are not observed with repetitions.

Analyses of eye movements in the no change condition did suggest that repetitions were not being missed by participants, even if they were not influencing the depth of semantic representations for words that follow. In the no change condition, participants spent longer gazing at the target region when the passage contained a repetition. This effect of repetitions was not present during the first pass, and so may represent later processes.

This result was unexpected given that no other effects of repetitions were observed; however, we can see at least three possible explanations for the effect. The first possibility is that there was in fact some form of attentional effect of repetitions; however, this would seem unlikely given the absence of effects in the change conditions where heightening of attention should have demonstrable consequences. Additionally, if participants' attention was being heightened then we might expect them to realise that the target word had not changed, and therefore not need to spend a greater amount of time in the region.

The second possibility is that, as repetitions appear to disrupt recognition of the words that are repeated (Bard & Lickley, 1998), participants might have missed the repeated function word and assumed that it had been added in the second presentation. This explanation would also seem unlikely however, given that the instructions stressed that a word would be *changed* rather than *added*. If participants were misunderstanding the instructions then we might expect an increase in false-positives (reporting a change that had not occurred) in the no-change condition; however, there was no evidence to suggest that accuracy was any poorer for repetitions in the no-change condition (numerically, participants were most accurate at detecting when a change had *not* occurred when the passage contained a repetition; although, this difference did not reach significance).

Registering a correct detection of a change took two steps in our experiment. First, the participant had to respond that they had detected a change and then they had to pick out the changed word from a set of nine options. If a participant responded that they had detected a change when one had not occurred then the options may have provided disconfirmatory evidence (if the word they wrongly believed had changed was not present in the options). This feature of the experiment may be obscuring the number of times participants falsely reported that they detected a change. With this in mind, we examined the responses to the initial detection question. We found no evidence here that participants were any more likely to incorrectly detect that a change had occurred in the repetition condition (92.78% accuracy) than in the fluent condition (88.89% accuracy). Numerically at least, participants actually produced more false positives in the fluent condition than in the repetition condition, suggesting that the repetitions were not causing them to misrecognise words which they later assumed had been changed.

The final possibility is that as the majority of repetitions occurred in change conditions, either close or distant, participants might have learnt that if they heard a repetition then the word following it would be more likely to change. Again we do not consider this to be the case as, if participants were sensitive to this pattern, they should have detected more changes when they did occur.

As none of the three possible explanations are compatible with all of the evidence, we are unable to account for the effect of repetitions on total gaze duration in the no change condition. However, this pattern would suggest that the absence of an effect of repetitions on change detection is not due to participants failing to notice that a repetition has occurred.

Clark and Wasow (1998) suggest that there is a functional similarity between filled pauses and repetitions. Our results provide no evidence to suggest that this is the case, at least not from the perspective of comprehension. Their suggestion appears to assume that filled pauses are largely homogeneous with regard to their function; however Fox Tree (2001) found that the beneficial effects for probe word recognition of "uh" did not occur with "um" (although, as discussed in Chapter 2, it is not clear that the null effect for ums was not confounded by pause durations left behind following the excising of the filled pause). It is currently unknown whether hearing an "um" has any effect on language comprehension, but if future research discovered the existence of effects then it would be sensible to determine if similar effects are present with repetitions to investigate whether any functional similarity in comprehension lies solely between repetitions and ums.

MacGregor et al. (2009) offer a suggestion of why effects obtained with filled pauses (e.g. Collard, 2009; Corley et al., 2007) are not observed with repetitions (e.g., the present study and MacGregor et al., 2009). They point out that unlike filled pauses, which have a disputable lexical status, the disfluent repetition of words may be less easily distinguished from the lexical context in which they occur. Furthermore, while they could be viewed as offering a delay to listeners, the delay is filled by linguistic content which forms a part of the discourse alongside the remainder of the utterance. Consistent with the idea that the delay that repetitions provide is not taken by listeners as a chance to further process preceding material, studies of silent pauses have observed effects that are not present for repetitions (such as in the present study, and MacGregor et al., 2010)

It may be the case that repetitions are not phonetically homogeneous (Hieke, 1981; Plauché & Shriberg, 1999), and therefore that only particular types of repetitions will elicit particular effects. Plauché and Shriberg identified three types of repetitions in the Switchboard corpus: canonical repetitions, covert self-repair repetitions and stalling repetitions. Each type, it is argued, serves a different function for the speaker and it is possible that effects on listeners are similarly varied. For their experiments, Fox Tree (1995) and MacGregor et al. (2009) created disfluent stimuli by taking fluent utterances and editing the recording to repeat a token with a pause inbetween. As identical tokens are used, these repetitions are most similar to Plauché and Shriberg's (1999) covert self-repairs;

however, the pause that is inserted between the two tokens is not compatible with repetitions of this sort. Moreover, as the original tokens were fluent we would expect them to be of a typical duration. Each token within all of Plauché and Shriberg's repetitions are prolonged (with the exception of the second token in a canonical repetition). In constructing stimuli for the present experiment the actor producing each utterance was naïve to the three types of repetitions and was only given the guidance to produce a repetition that they felt was natural. It remains a strong possibility that the repetitions used in all of these experiments are not, in fact, "natural" repetitions (at least not forms of repetitions identified by Plauché and Shriberg).[4] Any future research investigating the effects of listening to repetitions should construct stimuli informed by Plauché and Shriberg's sets of repetitions to eliminate the possibility that the lack of effects observed in those studies previously mentioned are not due to the use of "pseudo" repetitions.

4.5 Conclusions

The experiment presented in this chapter suggests that, unlike when it is preceded by a filled pause, a word preceded by a disfluent repetition is not represented in greater semantic depth than the same word appearing in a fluent context. While listeners may sometimes be sensitive to the presence of a repetition, demonstrated in the later reading effect observed in our experiment and the early EEG effects observed by MacGregor et al. (2009), there is no evidence that this sensitivity has any consequence for the linguistic processing of subsequent words. If, as Clark (1996) suggests, repetitions are being designed for the benefit of an audience then, unlike in the comprehension of filled pauses, there is little evidence that the audience is receiving any benefit.

In the following chapter we will finish testing the hesitation as signal hypothesis with an experiment and a set of corpus analyses focused on investigating whether hesitations (including repetitions) meet the third of Kraljic and Brennan's criteria by examining whether variations in the situation in which speech takes place has any effect on the types on the hesitations that speakers produce.

[4] While we have not conducted a comprehensive phonetic examination of the stimuli, listening to the recordings suggest that, in general, our repetitions do not fit neatly into any of Plauché and Shriberg's three categories. In large part, this is due to a general absence of perceived prolongations in either token.

Chapter 5

Testing the hesitation-as-signal hypothesis

In Chapter 2 we reviewed evidence suggesting that many of the hesitations that speakers produce are associated with difficulties in planning utterances and accessing the words which they will contain. Upon hearing a hesitation (although psycholinguists have tended to focus on filled pauses), listeners appear able to infer the cause of the difficulty which may have led to its production. Taking together studies on the production and comprehension of hesitations, it would appear that hesitations, or at least filled pauses, are *signs* of difficulty, which may be interpreted as such by audiences. What remains a matter of dispute is whether speakers are designing their hesitations so that their audience will interpret them as a signal that they are experiencing difficulty or whether they are natural symptoms of difficulty which listeners happen to interpret.

Evidence that hesitations are reliably produced (i.e. that they index specific types of difficulty), and that they are readily interpreted, suggest that they meet the first two of Kraljic and Brennan's (2005) criteria for a designed feature of speech. In order to further address the question of whether certain hesitations are being designed by speakers we turn our attention to Kraljic and Brennan's third criterion: that production of the feature "must vary depending on speakers' intentions in the situation or toward addressees" (p. 197). Invoking intentions introduces a difficulty for assessing whether or not hesitations are designed. Short of asking speakers what they intended with each hesitation that they produce, it is only possible to infer their intentions. One means of overcoming this difficulty is by investigating the production of the feature of interest in contexts which are manipulated to constrain the intentions that a speaker may have (Nicholson, 2007; Schober & Brennan, 2003). In their own study of prosodic marking, Kraljic and Brennan argued that if speakers use prosodic marking to help listeners correctly parse ambiguous sentences then they should be more likely to prosodically mark disambiguating words in utterances which are otherwise ambiguous than those which are not. In such a case, if speakers were more likely to prosodically mark disambiguating words when they knew the sentence was otherwise ambiguous for the listener then we may infer that they designed the prosodic marking to help disambiguate the sentence. We can draw a parallel with hesitations: If hesitations are more likely to be produced in situations where they could serve a communicative function then it may be because the speaker designed the hesitation to serve this function.

In this chapter we will investigate whether the production of hesitations varies according to manipulations of the two aspects that Kraljic and Brennan suggest should influence the production of designed features of speech: the audience and the situation. Firstly, we investigate whether having an audience for one's speech, thereby having someone to design hesitations for, increases the likelihood of producing a hesitation. Secondly, we investigate whether manipulating the situation in which a dialogue takes place, which may alter the strategies required for communicative success, has an influence on the types of disfluencies that speakers produce.

5.1 Experiment 2: The influence of an audience on hesitations

According to the hesitation-as-signal hypothesis, hesitations are produced for the benefit of an audience. In particular, it is argued that speakers produce hesitations in order to manage the flow of conversation (e.g. to account for their use of time when speech is disrupted, and to stop interlocutors from wrongly interpreting a disruption as the end of a turn). If the purpose of producing a hesitation is to manage conversation with an interlocutor, then we might expect that the elimination of the interlocutor—turning the dialogue into a monologue—should eliminate the production of hesitations (as there would no longer be a reason to produce them).

Philosophers have long recognised the importance of an interlocutor when producing signals in order to communicate. Grice (1957) defines a signal as being produced "with the intention of inducing a belief [in the audience] by means of the recognition of this intention" (p. 384). In the absence of an audience, there would be no one in whom to induce a belief nor anybody who could recognise that inducing the belief was the speaker's intention. If a person was performing an action (either verbal or non-verbal) when they knew that nobody would be able to recognise the intention of the action (because there was nobody to perceive the action), then this action could not be said to have a communicative intention.

The assumption that communicative behaviours are less likely to be produced without an audience to communicate to is reflected in methodologies used in comparative psychology (see Leavens, Russell, & Hopkins, 2005), where the presence, or attention, of an audience is used as independent variables in many studies of communicative behaviour in non-human primates (e.g. Call & Tomasello, 1994; Hostetter, Cantero, & Hopkins, 2001; Leavens, Hopkins, & Bard, 1996; Leavens, Hopkins, & Thomas, 2004). Following a similar logic, Bavelas, Gerwing, Sutton, and Prevost (2008) investigated the communicative function of hand gestures by manipulating whether participants were in a face-to-face dialogue or in a monologue (or in a dialogue over the telephone). Consistent with the idea that hand gestures are produced with a communicative intention, they were found to occur more often in dialogue than in monologue, particularly in dialogue where participants were able to see each other.

We know of three studies which allow for the comparison to be made between the production of hesitations in dialogue, with an audience, and in monologue, without an audience (Broen & Siegel, 1972; Finlayson & Corley, 2012; Oviatt, 1995). Oviatt compared the disfluency rates (considering repairs, in addition to hesitations) of participants describing how to build a water pump, in monologue and dialogue conditions, in order that a partner could build the pump out of its components (recordings were originally collected for Oviatt & Cohen, 1991). Oviatt found that participants were more disfluent in dialogue than in monologue.

While it has been suggested elsewhere that Oviatt's finding may support the hesitation-as-signal hypothesis (e.g. Corley & Stewart, 2008), it is not clear that this is the case. As Corley and Stewart highlight, in one of the studies reported by Oviatt, 77% of variance in disfluency rates was accounted for by utterance length. As this was not controlled for in the study which compared monologues and dialogues, it is not clear that the results of this study were not confounded. Furthermore, it is not clear that the manipulations they investigated are well suited to evaluating the hesitation-as-signal hypothesis. In the monologue condition, participants knew that their instructions would later be used by someone else to construct the pump. In Gricean terms, participants believed that they had an audience (albeit after-the-fact) by whom their communicative intentions could be recognised. While we do not dispute that there is a substantive difference between the monologue and dialogue conditions (the manipulation had an effect on disfluency rates; although we cannot be sure that this is not confounded by utterance length), the finding does not provide unequivocal evidence in support of the hesitation-as-signal hypothesis.

Further evidence to suggest that there are differences between monologue and dialogue in the likelihood of speakers producing hesitations comes from Broen and Siegel's (1972) investigation of the effect that a speaker's belief in the importance of being fluent has on the hesitations that they produce. Participants performed two monologues, initially alone, and then either in front of a TV camera or whilst imagining that an audience was physically present (no significant differences were found between the TV camera and the imagined audience for any dependent variables, and we will subsequently refer to them as the audience condition). Finally, they were recorded in conversation with the experimenter. Participants were allowed to talk about any subject they wished (they were provided with cards prompting particular topics if necessary), with these subjects returned to in the subsequent conversation. When asked to rate how important they thought it was to be fluent in each situation, participants were found to believe that fluency was most important in the audience condition, less important in the alone condition, and least important in the conversation condition. Their perception of the importance of fluency was reflected in the speech that they produced, with hesitations more frequent in conversation than in the audience or alone conditions.

In an earlier study (reported as Experiment 1 of Finlayson & Corley, 2012), we tested the hesitation-as-signal hypothesis in an experiment where participants had to perform a picture-naming task in each of a monologue and dialogue situation.[1] While such a task produces language which is less like conversational speech than that elicited by Broen and Siegel, the linguistic constraints imposed (i.e. that all that could be discussed were the names, and locations, of the pictures) allow us to be sure that the conditions being compared are similar except for the critical manipulation. Participants performed the task as a monologue and a dialogue within the same session, with the order of conditions counterbalanced. They were told that each situation was part of a different experiment, designed by different researchers with different purposes (although the same experimenter

[1] This experiment was conducted as part of the author's undergraduate studies. Experiment 2 of the present thesis was reported as Experiment 2 of Finlayson and Corley (2012)

collected data for both experiments). In the monologue condition, participants were told that the purpose of the experiment was to "record phonemes" to be used in the development of speech synthesizers. This was in order to stop them from treating the task as communicative. In the dialogue condition, participants were told that the purpose of the experiment was to investigate communicative strategies adopted during cooperative tasks. Both conditions required the participant to name images contained in grids; however, in the dialogue condition, the naming was part of a picture-sorting task that the participant undertook with an interlocutor (unknown to participants, the interlocutor was a confederate of the experimenter). In the dialogue condition, both the participant and confederate took turns to name the images in their respective grids, so that their partner could recreate the layout of the grid using individual images. Participants named two "types" of images: *disfluency* images, which were either easy or hard to name (where difficulty was defined as low frequency and low name agreement); and *alignment* images, which, unknown to the participant, the confederate had been scripted to name using either a commonly used (*preferred*) or an alternative (*dispreferred*) name. Images were ordered in such a way that the confederate would name the alignment images before the participant.

Participants were more likely to produce hesitations when naming hard-to-name images; however, while the language used by participants appeared to be influenced by the presence of an interlocutor (reflected by participants being more likely to use dispreferred names when their interlocutor had previous used them), they were no more likely to produce hesitations in the dialogue condition than in the monologue condition. Furthermore, the distribution of different types of hesitations did not change between conditions, so different types of hesitations were not trading-off against one another (for example, speakers producing more filled pauses but fewer repetitions in one condition compared to another).

One possible explanation for the null effect in this study is that the experiment may have lacked statistical power, due to the relatively small number of hesitations observed (less than 15% of trials contained a hesitation). Consistent with this possibility, the proportion of trials containing hesitations was numerically greater when the confederate was present. If the hesitation-as-signal hypothesis is correct, with speakers expected to be more likely to produce hesitations with an audience, then the difference between monologue and dialogue may reach statistical significance if participants produced a greater number of hesitations overall (numerically, the differences were consistent with the hesitation-as-signal hypothesis). One way of increasing the number of hesitations would be to make the items depicted in the pictures harder to recognise. In an experiment using the network task (Oomen & Postma, 2001), Schnadt (2009) found that participants were more likely to produce hesitations when the images in the network were blurred than when they were clear. To determine whether the null effect observed in the previous study was just due to the scarcity of hesitations, Experiment 2 uses the same methodology as the previous study, with the exception that all of the images to be named were blurred.

5.1.1 Methods

As earlier noted, the methodology of this experiment was largely identical to that used in our earlier study; however, we will repeat the details of that experiment in order that our reporting of the present study may be understood without the reader having to refer to Finlayson and Corley (2012).

Participants

Twenty-four native British English speaking undergraduates from the University of Edinburgh volunteered to participate in this experiment. All reported having no known speech, language or hearing disorders.

Materials

Images were selected from the International Picture Naming Project (IPNP; Szekely et al., 2004). The IPNP provides normed information for 520 black-and-white line drawings of common objects. Where images could not be freely obtained from the IPNP, suitable replacement images were selected from a commercially available clip art package.

Participants named two types of images: *disfluency* images and *alignment* images. The names of the images used in this experiment are provided in Appendix B. Thirty-two disfluency images were classified as being either easy-to-name or hard-to-name (sixteen of each). Following Schnadt and Corley (2006), we used two forms of difficulty: name agreement (how many names are used for the image) and frequency. Name agreement can be measured using the H-statistic (alternatively known as U; Snodgrass & Vanderwart, 1980), which has a value of 0 when the same name is always used for a picture and increases when more names are used (high values of H correspond to low name agreement).

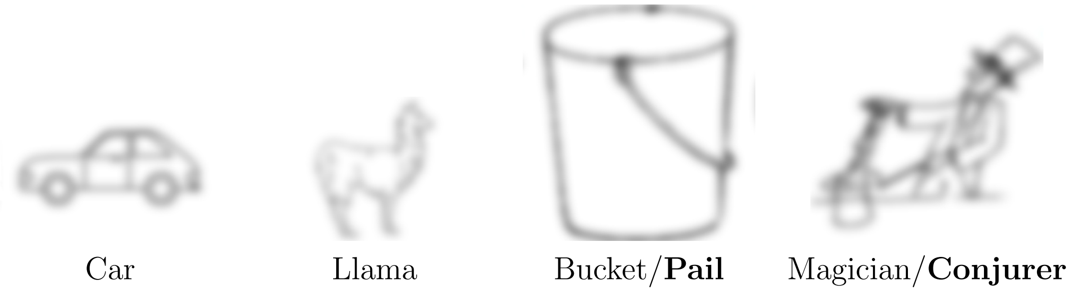
Values of H for each image were taken from the IPNP, whilst CELEX (Baayen, Piepenbrock, & van Rijn, 1993) was used to provide information on frequencies. Easy-to-name images were those with high name agreement, $H < .15$ ($M = 0.06$, $SD = 0.07$), and a high frequency dominant name, > 75 counts per million (cpm; $M = 255$, $SD = 167$). Hard-to-name images were those with low name agreement, $H > 0.85$ ($M = 1.60$, $SD = 0.39$), and a dominant name of low frequency, < 25cpm ($M = 4.00$, $SD = 4.75$).

Ten raters were shown an additional 40 images, and were instructed to name the images, as well as rating alternative names for appropriateness. Alternative names were infrequently used names for each image selected from the Beckman Spoken Picture Naming Norms (Griffin & Huitema, 1999). Eight images were discarded either because the most commonly-used name was used by fewer than 80% of the raters, or because the appropriateness rating of the alternative name was less than 2.5 out of 5. The thirty-two remaining images were used in the experiment as alignment images.

Finally, thirty-two filler images were selected which would be named by the confederate. These images depicted common objects, with no constraint placed on how difficult they were to name.

Four 4×4 grids were created for participants to name. Images were randomly assigned to one of the sixteen numbered squares in the grid (numbered from left-to-right, top-to-bottom). Each grid contained eight disfluency images (half of which were difficult to name), and eight alignment images. Four grids were created for the confederate. Instead of images, the names of the objects depicted in the images were printed, serving as scripts for the confederate. Each of these contained the names of eight alignment images and of eight filler images. For each grid, the confederate used five dispreferred names and three preferred names (to increase the opportunity for alignment). For the dialogue condition, both the participant and confederate were given a blank 4×4 grid, consisting of numbered squares, upon which each could arrange cards depicting the images named by their partner.

All images were digitally blurred using a Gaussian algorithm ($\sigma = 6$ pixels) with ImageMagick. Example images are shown in Figure 5.1.



Car Llama Bucket/**Pail** Magician/**Conjurer**

Figure 5.1: Experiment 2: Examples of easy-to-name (car) and a hard-to-name (llama) images, and two alignment images (for each image the preferred name is given, followed by the dispreferred name used by the confederate in bold).

Procedure

Participants were collected from a waiting area, along with a confederate who posed as a fellow naïve participant. Together, the participant and the confederate were provided with instructions and signed consent forms.

In order to prevent participants from realising that their performance in monologue and dialogue would be compared, participants were told that they would be performing two separate experiments for two separate researchers (only one of whom was present). They were also told that each researcher was based at a different institution (Queen Margaret University and the University of Edinburgh). To further reinforce the distinction between conditions, they were given two different instruction sheets and signed two different consent forms (each of which carried the letterhead of a different institution). In the monologue condition, participants were informed that a researcher at Queen Margaret University required recordings of phonemes occurring in semantically-arbitrary natural speech which would be used in the development of a speech synthesizer. These instructions were intended to minimise communicative aspects of the task. In the dialogue condition, participants were told that they would be performing an experiment for a researcher at the University of Edinburgh who was interested in the communicative strategies employed by speakers performing cooperative tasks. The order of conditions was counterbalanced across participants, and, following completion of both conditions, participants were debriefed as to the true nature of the experiment.

Each of the four grids were used equally often in both the monologue and dialogue conditions, with assignment of each grid to each trial following a Latin Square. In the monologue conditions, participants were provided with two grids and were instructed to name each image and the number of the cell in which it was located. It was suggested that participants produce a sentence containing the name and number, although no structure was suggested for this sentence. During naming, the experimenter left the room in order to eliminate the possibility that the participant felt that they were communicating with them.

In the dialogue condition, both the participant and the confederate were seated at a table separated by a partition which prevented them from seeing each other, and each other's grid, but did not prevent them from hearing each other. Both were given a grid of pictures, a blank grid, and a set of images printed on individual cards. They were instructed that they should take turns to name each picture, in the order in which they appear in their grid (e.g. one person would name the image in square one of their grid, then the other would name the image in square one of their grid, etc.). Both were given a suggestion of what they might say: "In box one I have a dog". They were instructed that, upon hearing their partner name one of their images, each should put the card containing that picture on the corresponding square of their blank grid (e.g. putting the picture of a dog on square one). The confederate always began the trial, ensuring that they named their alignment image before the participant named that same image. At no point did the participant name an alignment image immediately after hearing the confederate name that same image, ensuring that they could not simply "echo" the confederate. Instead, the number of turns between the confederate and the participant naming the same image varied between two and three. Once all images had been named, and both grids had been filled, the procedure was repeated with a second grid.

The confederate in the previous study was instructed only that they should read each name, and the square in which they were printed, as a sentence. To further convince participants that the confederate was also naming blurred images, rather than a script, the confederate in the present experiment was instructed to include some prolongations and some filled and silent pauses in their descriptions. While the confederate was given no guidance on when, or how often, they should be disfluent, they were coached by the author to produce natural sounding disfluencies.

Each participant's speech was recorded throughout the experiment, using a ZOOM H4n digital recorder. Whilst not switched on, another microphone was seated in front of the confederate to prevent participants from realising that only their speech was of interest.

Transcription and coding

The transcription and coding for fifteen participants was shared between two raters.

Each grid description was divided into 16 utterances, with each utterance corresponding to the description of one of the images. Descriptions consisted of two parts: a description of the numeric location of the image, followed by a description of the image. For the 768 utterances describing alignment images, they coded whether the participant used the preferred or dispreferred name. Where participants used more than one name, the first name mentioned was coded.

For the 768 disfluency images, raters were instructed to count the occurrences of each of the five categories of hesitation identified in the previous study: prolongations, the filled pauses *uh* and *um*, repetitions and silent pauses. The recordings of six participants were rated by both raters, with an 86.4% agreement on hesitations. For each of these six participants, one rater's coding was randomly selected for analysis. On the basis of these counts, a discrete outcome variable coded each utterance as either being fluent or as containing a hesitation.

5.1.2 Results

We conducted two independent analyses. The first focused on the utterances describing the alignment images, in order to establish that participants' linguistic behaviour was sensitive to the presence of an interlocutor. The second focused on the utterances describing the disfluency images. This second analysis investigated whether the presence of an interlocutor had any affect on the hesitations produced by participants.

As our dependent variables were binomial (whether or not participants used a dispreferred name; whether or not they produced a hesitation during the utterance), logistic mixed effects regression was used to model outcomes.

Influence of confederate on naming

Figure 5.2 shows the proportion of trials in which participants used the dispreferred name to refer to an alignment image. In the dialogue condition, participants heard the alignment images referred to by a dispreferred name for 63% of images. In the monologue condition, we still refer to the image as occurring in the dispreferred condition. As the experiment was fully counterbalanced we can compare cases where the dispreferred name was used (in the dialogue condition) with cases where it was not (in the monologue condition).

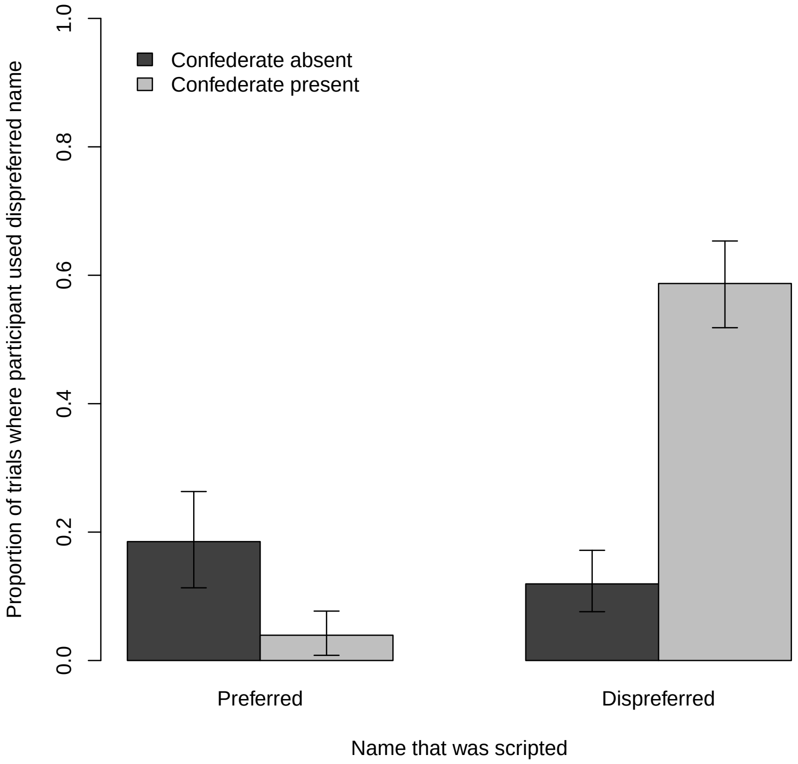


Figure 5.2: Experiment 2: The proportion of trials in which the dispreferred name was used to refer to an alignment image. In the dialogue condition, a preferred or dispreferred name was scripted and would be heard by the participant before they would name the image. In the monologue condition, no confederate was present and the scripted name was nominal only, in that the participant did not hear a name before they named the image themselves. Bars represent 95% confidence intervals estimated using bootstrap resampling (999 runs).

In 139 utterances, the participant did not use either the preferred or dispreferred name. Whilst this is a larger number of utterances than observed in the previous study (neither the preferred nor dispreferred name was used on 23 occasions), it is not unexpected as the blurring of images should make objects harder to correctly recognise (note also that the present study had four more participants than the previous one, providing a greater number of opportunities to use neither name). When a dispreferred name was scripted, participants were found to be over eighteen-and-a-half times as likely to use the dispreferred name ($p < .01$)

as when the preferred name was scripted. No effect of confederate presence was found (p = .51). Crucially, a significant interaction was observed between these two factors (p < .001), suggesting that participants were more likely to use the dispreferred name to refer to an image when they had previously heard the confederate refer to the image by that same name. Details of this model are given in Table 5.1.

Table 5.1: Experiment 2: Logistic mixed effects model of the likelihood of participants using the dispreferred name to refer to an alignment image.

					Random effects		
Fixed effect	β	SE	z	$p(\beta = 0)$	Group	Predictor	Variance
Intercept	-2.746	0.450	-6.102	$< .001$	Item	*Intercept*	3.960
Confederate present	0.298	0.457	0.653	.51		Confederate present	< 0.001
Dispreferred name scripted	2.920	0.941	3.103	$< .01$	Participant	*Intercept*	< 0.001
Confederate present × Dispreferred name scripted	6.780	1.047	6.479	$< .001$		Confederate present	0.101
						Dispreferred name scripted	1.940
						Confederate present × Dispreferred name scripted	6.670

Influence of confederate on hesitations

Figure 5.3 shows the proportion of trials in which participants produced a hesitation whilst describing a disfluency image. Participants were found to be over six-and-three-quarter times as likely to produce a hesitation when describing a hard-to-name image than an easy-to-name image (p < .001). No effect of confederate presence was found, while there was also no evidence of an interaction between these two factors (p = .38 and p = .93,respectively). Details of this model are given in Table 5.2.

As it is not clear that silent pauses could serve as a signal, we reran our analysis of hesitations without silent pauses. With silent pauses excluded, neither the effect of confederate presence nor an interaction with difficulty were found to reach significance (both ps < 1).

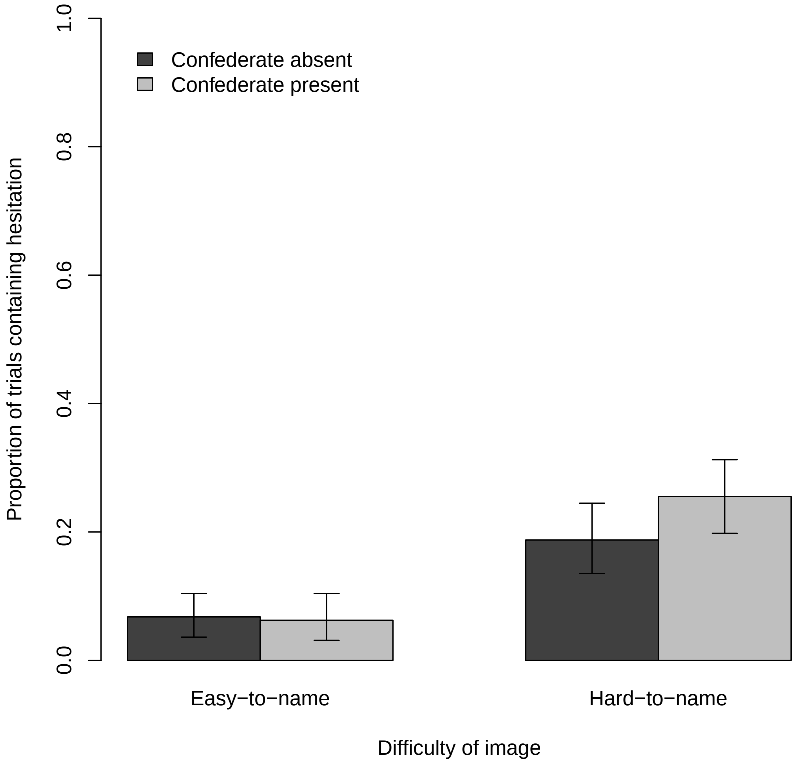


Figure 5.3: Experiment 2: The proportion of trials in which participants produced a hesitation whist naming a disfluent image. Bars represent 95% confidence intervals estimated using bootstrap resampling (999 runs).

Table 5.2: Experiment 2: Logistic mixed effects model of the likelihood of participants producing a hesitation during a trial.

					Random effects		
Fixed effect	β	SE	z	$p(\beta = 0)$	Group	Predictor	Variance
Intercept	−1.937	0.292	−6.642	< .001	Item	*Intercept*	1.090
Confederate present	0.216	0.246	0.875	.38		Confederate present	0.280
Hard-to-name	1.917	0.459	4.178	< .001	Participant	*Intercept*	0.843
Confederate present × Hard-to-name	−0.044	0.528	−0.083	.93		Confederate present	< 0.001
						Hard-to-name	0.322
						Confederate present × Hard-to-name	0.805

Table 5.3: Logistic mixed effects model of the likelihood of participants producing a hesitation during a trial in Experiment 2 and Finlayson and Corley (2012; Experiment 1).

					Random effects		
Fixed effect	β	SE	z	$p(\beta = 0)$	Group	Predictor	Variance
Intercept	−2.044	0.231	−8.836	< .001	Item	*Intercept*	0.645
Confederate present	0.225	0.196	1.143	.25		Confederate present	0.133
Hard-to-name	1.442	0.359	4.018	< .001		Experiment	0.386
Experiment	0.466	0.207	2.253	< .05		Confederate present × Experiment	< 0.001
Confederate present × Hard-to-name	0.198	0.363	0.545	.59	Participant	*Intercept*	0.582
Confederate present × Experiment	−0.043	0.337	−0.129	.90		Confederate present	0.119
Hard-to-name × Experiment	0.693	0.410	1.691	.09		Hard-to-name	0.345
Confederate present × Hard-to-name × Experiment	−0.417	0.670	−0.623	.53		Confederate present × Experiment	< 0.001

Finally, we analysed the combined data of the present experiment and that collected in the previous study. Our analysis contained an additional fixed effect for *experiment*, which was allowed to interact with all other fixed effects, and random slopes for experiment-by-items and for all licensed interactions by-items. The results of this analysis are shown in Table 5.3. As was the case when the data from each experiment was analysed individually, no effects were found for the presence of the confederate or its interaction with difficulty ($p = .25$ and $p = .59$, respectively). A main effect of experiment suggests that participants were over one and a half times more likely to produce hesitations whilst naming blurred images ($p < .05$), while a marginal interaction with difficulty suggests that this effect is strongest for hard-to-name images ($p = .09$). No other interactions with experiment were found to reach significance.

There remains the possibility that while the presence of the confederate had no effect on the overall likelihood of producing a hesitation, the distribution of hesitations may vary between conditions. Table 5.4 shows the counts observed for each type of hesitation in the monologue and dialogue conditions. A Fisher's exact test confirmed that the distribution of hesitations did not vary ($p = .47$).

Table 5.4: Experiment 2: Total numbers of hesitations observed in each of six categories across the experiment.

	Prolongation	*Uh*	*Um*	Silence	Repetition
Confederate Absent	79	3	21	141	12
Confederate Present	66	8	18	112	8

5.1.3 Discussion

The present experiment tested the hesitation-as-signal hypothesis by manipulating whether participants named images in a communicative situation, with an interlocutor, or in a non-communicative situation, in isolation, and examining the effect this had on fluency. As the hesitation-as-signal hypothesis suggests that hesitations are being designed to manage conversation with an interlocutor, we would expect that in a monologue the speaker should produce fewer hesitations than they would when they were in a dialogue with an interlocutor.

In order to be confident that the linguistic behaviour of our participants was sensitive to the presence of an interlocutor, we first demonstrated that participants took their interlocutor into account when naming images. The interlocutor, a confederate of the experimenter, was scripted to refer to certain images using a dispreferred name. When participants had to later name these images, they were more likely to use dispreferred names than they were when their interlocutor had instead previously used a preferred name. This reusing of referring expressions is similar to that which has been observed elsewhere (e.g. Clark & Wilkes-Gibbs, 1986).

After having established that our participants were taking their interlocutor into account when designing certain aspects of their speech, we were then able to investigate factors influencing the hesitations that they produced. As was predicted on the basis of previous studies (e.g. Schnadt, 2009; Schnadt & Corley, 2006), participants were more likely to produce hesitations when describing hard-to-name images (i.e. those with low name agreement and a commonly-used name of low frequency).

In Chapter 2, we highlighted the difficulty which arises when testing the hesitation-as-signal hypothesis: hesitations may be only symptoms of the difficulties that the hypothesis suggests that they signal. In other words, our participants could have been producing filled pauses to signal to their interlocutor that they were experiencing a delay in naming an image, or the filled pause may just be the sound that the language production system produces when it is disrupted because the name of the image is not readily accessible.

The consequence of this ambiguity is that simply showing that hesitations are associated with difficulty is not sufficient evidence to support the hesitation-as-signal hypothesis. If hesitations are merely symptoms of difficulty then we would expect them to be equally likely to be produced in dialogue and monologue, as the source of difficulty is present in both conditions. However, if hesitations are signals of difficulty that are designed to manage a conversation then we would expect that they should be more likely to occur when the speaker is engaged in a dialogue than when they are engaged in a monologue, as there is no conversation to manage.

When we examined the production of hesitations between dialogue and monologue, we found no evidence to suggest that participants were any more likely to produce hesitations when speaking to an interlocutor than when they were not. In our earlier study, which used a similar methodology, we also did not find any evidence that the presence of an interlocutor was having an effect on participants' fluency. Earlier in this chapter, we suggested that one possible explanation for the absence of an effect of interlocutor presence on hesitations in our earlier study was that it may have lacked statistical power due to a general scarcity of hesitations. In the present study, we blurred the images that participants had to name, which has been shown elsewhere to increase disfluency rates (e.g. Schnadt, 2009), in order to increase power. A cross-experiment analysis suggested that this had the desired effect, with participants in the present study more likely to produce hesitations than those in our earlier study.

There are three possible accounts of our results. Firstly, the reason that we did not observe a difference between monologue and dialogue may be because our dialogues were sufficiently structured (e.g. because participants and confederates had clearly defined turns, and the number of possible things that could be talked about in any turn was low) that participants did not need to use hesitations to manage conversation or account for their use of time (e.g. Clark, 1996, 2002). In other words, this account suggests that we did not observe a difference in hesitations between monologue and dialogue because our dialogue was not sufficiently unlike monologue. This may be a fair criticism (which we will address later), however it does beg the question of why participants produced any hesitations at all (at least hesitations which are not silent pauses). While the difficulties that co-occur with hesitations may be equally likely in monologue as in dialogue, there should be no need to produce signals in monologue (or "monologue-like" dialogue). Across conditions, participants in our experiment produced a hesitation in 17.5% of trials; while, excluding silent pauses, this drops to 7.5%. What remains unexplained, by both this account and hesitation-as-signal hypothesis more generally is the 10% of monologue trials where speakers produced "unnecessary" signals.

The second possible account for our null finding is that, rather than speakers performing in dialogue as if it were monologue, it may be that speakers were performing in monologue as if it were a dialogue. In other words, speakers may use hesitations as signals in dialogue and behave similarly in monologue. This may be out of habit, as, in spoken language at least, dialogue is much more frequent than monologue, or it may be because many speakers lack specific skills for monologue. This account has the appeal of explaining why speakers are disfluent in monologue (not just in our experiment but in many others); however, a consequence of this account would be that it becomes even more difficult to test the hesitation-as-signal account. One alternative source of evidence that could be explored is the developmental literature. Hudson Kam and Edwards (2008) examined the filled pauses of 3–4 year olds to see if they exhibited the difference in delays following "uh" and "um" observed by Clark and Fox Tree (2002). They found that, while longer silent pauses were more likely to be preceded by a filled pause than a shorter silent pause, the differences in pauses following filled pauses did not systematically vary depending on the realisation of the filled pause. If the use of hesitations as a signal is a skill that gradually develops before becoming habitual, then there may be a stage, before the habit is formed, where children perform differently in monologue than in dialogue.

The third, and most parsimonious, account of our null finding is that participants were no more likely to be hesitant in dialogue than in monologue because they were not designing their hesitations for the benefit of an interlocutor. Rather, consistent with the claim that hesitations are natural symptoms of difficulty, participants in our experiment were only more likely to produce a hesitation when the image they had to describe was hard-to-name.

The results of our experiment do not provide evidence to suggest that hesitations meet the first part of Kraljic and Brennan's (2005) third criterion for a signal: that their production vary according to the intention of the speaker towards their addressee. Regardless of whether or not there was an addressee for speakers to converse with, their likelihood of producing a hesitation remained constant. In the next section we will further investigate whether hesitations meet this criterion by using an analysis of a task-orientated corpus of dialogue to explore whether the situation in which a dialogue takes place influences the production of hesitations.

5.2 Corpus analysis 1: The influence of the situation on hesitations

According to the hesitation-as-signal hypothesis, hesitations are designed by speakers in order to manage conversations, for example by signalling when a speaker is experiencing difficulty and accounting for the speaker's use of time (Clark, 1996, 2002). Given this function, we might expect that as the situation in which a conversation takes place changes so too might the signals that are required to manage the conversation. Such a proposal is in line with the second part of Kraljic and Brennan's (2005) third criterion, that the production of a designed feature of speech should vary according to the speaker's intentions toward the situation. Furthermore, by varying the addressee, for example whether they are a friend of, or a stranger to, the speaker, we may also further test the first part of this criterion by investigating whether the production of hesitations vary according to the speaker's intentions toward their audience.

Nicholson (2007; Nicholson et al., 2003) conducted a series of experiments intended to investigate whether certain disfluencies, including repetitions and filled pauses, are being designed by speakers, or whether they are an automatic consequence of cognitive difficulty, by manipulating the situation in which speakers produced language. In a modified version of the map task (Anderson et al., 1991), participants described routes overlaid on maps presented on a computer screen, for the benefit of a listener who was attempting draw the route on their own copy of the map. In half of the trials, participants saw a moving icon on the screen which they were told represented the location of the listener's gaze. Unknown to participants, there was in fact no listener and the eye-movements had been programmed by the experimenter.

Nicholson predicted that if disfluencies are being designed by speakers as a helpful signal then they should be more likely to occur when the speaker is provided with feedback about their partner's understanding of the descriptions – perhaps because speakers would be better aware of when, and what, help was needed. The production of hesitations in her experiment was found to be insensitive to this feedback manipulation; however, speakers were more likely to produce deletions (where a speaker abandons an utterance entirely and begins to plan another) when they believed that they were being provided with feedback on their partner's understanding through having the ability to follow their gaze.

In a subsequent experiment, Nicholson investigated whether a speaker's motivation to be cooperative influenced the disfluencies they produced. She reasoned that one explanation for the lack of effects observed in the earlier experiment was that if speakers were experiencing cognitive burden due to performing the task then they may have insufficient cognitive resources to engage in cooperative behaviour (cf. Horton & Keysar, 1996). By giving half of the participants an incentive to perform the task well,[2] an additional £5 payment, Nicholson predicted that participants may be motivated to overcome the burden and produce an increased number of helpful disfluencies. An effect of this manipulation was observed for deletions, with motivated participants more likely to abandon utterances. Furthermore an effect in the opposite direction was found with substitutions, such that motivated participants produced fewer substitutions. However, as with feedback, the manipulation had no effect on hesitations.

As we suggested in Chapter 2 about repairs, it is not clear that deletions could be being produced as signals in a manner similar to that suggested by hesitation-as-signal hypothesis. Rather, abandoning an utterance likely reflects that the utterance that was intended is no longer appropriate or accurate. As it would seem unlikely that deletions are signals, and we know of no one who has suggested that they are, it is not clear that these specific effects speak to the validity of the hesitation-as-signal hypothesis. Taking together the results of both of Nicholson's experiments, there is little evidence to suggest that hesitations are being designed by speakers to be helpful.

While Nicholson's results provide little support for the hesitation-as-signal hypothesis, we would argue that tasks such as those she used provide a valuable resource for testing the hypothesis. The picture-naming task used in our Experiment 2 affords a high degree of control over what participants say; however, we would not dispute the assertion that partners taking turns to describe pictures bares little more than a slight resemblance to actual conversation. A task such as that used by Nicholson provides richer, more naturalistic, samples of speech, while still allowing us to control for differences in what may be said (for example

[2] Participants were instructed that their performance had to reach an unspecified criterion in order to receive the additional payment. It is not clear whether participants were informed what this criterion was.

whether the speaker is leading or following, the material being described, etc.). The Map Task Corpus (MTC; Anderson et al., 1991) provides us with such a sample, while also being large enough to allow us to control for much of the noise that may be found in "freer" dialogue. In this section of the chapter, we present a set of analyses of hesitations in the MTC. In particular, we focus on the effects of manipulations of the situation on hesitations, and whether these provide evidence which is compatible with the hesitation-as-signal hypothesis.

The Map Task Corpus

During the creation of the MTC, several aspects of the situation in which participants performed the map task were manipulated (for further information of these aspects, see Chapter 3 and Anderson et al., 1991). In the present study we focus on the effects on fluency of three of these factors: In a given dialogue, a given participant was either a giver of instructions, or a follower; they were able to see their partner, or their view of their partner was obscured by a screen; and their partner was either a friend or a stranger prior to performing the task. Two of these factors were manipulated within participants (speaker's role and familiarity with their partner), while the visibility of partners was manipulated between participants.

The effects on communicative behaviour of all three of these factors have previously been explored, both in studies of the MTC and in other data. The results of these studies provide reasons to believe that these factors may have effects on communicative strategies adopted by speakers. In the remainder of this section we will discuss relevant findings from this literature.

We would expect that performing the giver role should entail difficulty for participants in the MTC. Givers of instructions have to say more than followers, with over twice as many tokens in the MTC produced by givers as by followers. However, the cognitive burden faced by givers does not just result from the amount language they must produce. Givers of instructions must formulate, and reformulate where necessary, appropriate descriptions of their maps. They must also respond to and resolve difficulties encountered by the follower.

Previous research shows that both how much, and what, is said influences speakers' fluency. The likelihood of being disfluent has been found to increase with an increase in utterance length (e.g. Oviatt, 1995), while participants taking a similar leading role in other dialogue games have been shown to be more likely to produce disfluencies than those that they are leading (Bortfeld et al., 2001). Furthermore, Lickley's (2001) study of the distribution of disfluencies across different types of conversational moves produced by speakers in the MTC found that those types of moves which were more likely to contain disfluencies, for example *instructions* and *clarifications*, were produced by givers more frequently than by followers. In other words, not only do givers produce more speech than followers but the sorts of speech that they produce are more likely to be disfluent.

For reasons just outlined, speakers' roles in the MTC are expected have an influence on the likelihood that they will produce hesitations. However, such trends would be entirely consistent with an alternative account suggesting that, rather than hesitations being designed as signals, they are merely symptoms of difficulty. While the remaining two factors manipulated in the MTC, *visibility* and *familiarity*, may have moment-to-moment effects on the cognitive demands faced by participants, it is not clear that systematic difficulties would arise for a speaker as a result of being unable to see a partner or, having the partner be a stranger. Therefore, any effects of either of these factors on the production of hesitations may be seen to offer support for the hesitation-as-signal hypothesis.

Effects of the second factor, visibility of a partner, on some aspects of linguistic performance in the MTC have already been investigated (Boyle et al., 1994; Bull & Aylett, 1998). In their analyses of the MTC, Boyle et al. showed that partners who were unable to see each other produced more dialogue turns, and that those turns tended to be longer. These partners were also more likely to interrupt each other, and to produce back-channel responses (interjections produced by interlocutors to signal agreement and understanding), presumably because the manipulation deprived them of non-verbal means of communication (e.g. nodding to signify agreement). Boyle et al. suggest that visibility between conversational partners allows for greater efficiency in communication; however, in the absence of this visibility, speakers are still able to fall back upon their linguistic "flexibility and versatility" (p. 1) in order to successfully manage aspects of communication such as turn-taking.

One example of a non-verbal cue that may help manage turn-taking is gaze. Kendon (1967) investigated the role of gaze in unstructured dialogues between pairs of strangers, finding that speakers tended to look towards their partner at the end of their turn. Furthermore, when a speaker looked towards their partner, the partner was less likely to either delay in responding or to not respond at all than they were when the speaker did not end their turn by looking at their partner. This may suggest that partners were using gaze as a cue to determine if it is their turn to speak (although see evidence reviewed in 6.2.1 which suggests that the use of gaze may depend on context). Using the MTC, both Bull and Aylett (1998) and the analysis of inter-turn intervals presented in the Chapter 7, show that inter-turn intervals tend to be longer when partners are able to see each other than when they are not able to. Taken together, the results just summarised suggest that interlocutors take longer to respond (if they respond) if they *do not* receive a gaze cue and that when they *cannot* receive a gaze cue they respond more quickly. These different trends may reflect conversational partners adjusting the strategies they employ for managing turn-taking in response to changes in the situation (i.e. when they are deprived of the possibility of using gaze cues).

If participants in the MTC were forced to change their strategies when they are unable to see each other, then what alternative strategy might they have adopted? Both filled pauses and repetitions have been suggested to have a function in the management of turn-taking (by Maclay & Osgood, 1959; Clark & Wasow, 1998, respectively): Speakers may use filled pauses and repetitions to signal to their audience that they have not yet said all that they intended and so should not be interrupted (known in the literature on turn-taking as an attempt-suppressing signal; Duncan, 1972). If depriving interlocutors of the ability to see each other causes them to rely more heavily on verbal attempt-suppressing signals, then we might expect that speakers who are unable to see their partners will be more likely to produce filled pauses and repetitions. Such a relationship was observed by Kasl and Mahl (1965): Participants produced more filled pauses when they were being interviewed by an experimenter who was in another room, and who would therefore be unable to see them, than when the experimenter was in the room with them (repetitions were also recorded; however, in the authors' analyses they were conflated with other types of "disturbance", including repairs and speech errors, leaving us unable to determine whether they were also more common when participants could not see their interlocutor).

Boyle et al.'s results suggest that speakers may in general rely more heavily on verbal strategies, for example speaking more, when their interlocutors are unable to perceive their non-verbal cues. If speakers produce more speech when they are unable to see their partner then this greater planning burden could itself cause speakers to produce more hesitations. If this is the case then simply showing that speakers are more hesitant when they are unable to see their partner would remain consistent with an account suggesting that hesitations are symptoms of difficulty. By statistically controlling for the amount of speech that speakers produce (e.g. by including a predictor for utterance length in our regressions), we would be able to rule out the explanation that an effect of visibility was confounded by the amount of speech being produced.

Finally, prior familiarity between partners in the MTC has also been shown to have an impact on the language produced by speakers (Boyle et al., 1994). Friends produce a greater number of conversational turns, consisting of a greater number of words, and they are better at performing the task (as quantified by the amount of difference between the route on the giver's map and the route replicated by the follower) than strangers. Familiarity between partners also appears to have an effect on the manner in which speech is performed: Horton (2007) has shown that it is possible to predict from the prosodic features of *common ground units*, "dialogue segments in which discourse participants add content to their common ground" (Nakatani & Traum, 1999, p. 3), whether conversational partners were friends or strangers. Taken together, this suggests that there are differences in the ways in which people speak to friends and to strangers. If these differences extend to the hesitations that they produce then this would suggest that, consistent with Kraljic and Brennan's (2005) third criterion, the production of hesitations may vary according to speakers' intentions toward the addressee.

In an earlier analysis of the MTC, Branigan et al. (1999) explored the effects of factors including role, visibility and familiarity on the production of disfluencies. As would be expected, an effect of role was found, such that givers of instructions produced a higher rate of disfluencies than did followers. Although they did not reach significance, numerical trends suggested that visibility and familiarity both also influenced the rate of disfluencies a speaker produced, with partners who were able to see each other and those who were strangers producing fewer disfluencies. We must be cautious in interpreting Branigan et al.'s results as their analyses focused on only one factor at a time, conflating those factors which were not of interest. Using multiple regression, allowing for all of these factors to be considered simultaneously, Bard et al. (2001) explored the effects of these factors on MTC speakers' disfluency rates (disfluencies per conversational move). In addition to replicating the earlier observed effect of role, the effect of familiarity observed numerically by Branigan et al. was found to be significant in Bard et al.'s analysis. No effect of visibility was found.

It is not clear how well either Bard et al.'s or Branigan et al.'s findings speak to the hesitation-as-signal hypothesis. In their analyses, both Bard et al. and Branigan et al. conflated repetitions and repairs, the latter of which are not hesitations, while their measures of disfluency rates failed to include filled pauses, the type of hesitation most commonly suggested to be a signal (e.g. Clark & Fox Tree, 2002). Moreover, we have concerns about the statistical analyses presented in both studies. Bard et al.'s analysis considered only those moves (utterances which serve a specific purpose in the map task, for example requests for clarification or confirmatory responses; Carletta et al., 1996) where a speaker was likely to be responding to the content of their partner's previous move (with this likelihood assessed using the duration of inter-turn intervals); as a result, they excluded over one-quarter of the moves produced in the MTC. Furthermore, by analysing rates of disfluency (either per move or per 100 words) both Bard et al. and Branigan et al. may have violated the assumption of linearity (discussed in Chapter 3). Taken together, the results of both studies may be taken as indicative at best.

The purpose of the present study is to explore the factors which influence the production of different types of disfluencies. With the use of mixed effects regression, which allows us to control for possible sources of noise in the MTC, we will investigate whether variations in the situation in which a dialogue takes place influence speakers' fluency in ways which would be consistent with the claim that hesitations are being designed by speakers to perform a communicative function.

In addition to considering cases of hesitations, we will also analyse the effects of the situation on the production of repairs. As we argued in 2.2, repairs are not designed to be signals, rather, they are produced to correct infelicities in already uttered speech (and our reading of Clark, 1996, provides no reason to believe that the hesitation-as-signal hypothesis would contradict this view). For our purposes, repairs provide a "control" case of disfluency. If, for example, visibility was found to be having a similar influence on the production of filled pauses *and* the production of repairs then a parsimonious account of these findings would be that filled pauses were not being produced with the communicative function that is suggested in the hesitation-as-signal hypothesis (as the same factor has the same effect on the production of a "non-communicative" disfluency). Instead, if hesitations are being designed by speakers then we might expect them to be influenced by visibility, but speakers to be no more likely to produce a repair when they were unable to see their partner than when they were able.

Autonomous restart capability

While the primary focus of the corpus analyses presented in this chapter is to test the hesitation-as-signal hypothesis, the analyses provide the opportunity to resolve some issues with a previous set of analyses we conducted on the MTC. Finlayson, Lickley, and Corley (2010) found evidence of relationships between articulation rate and the production of different types of disfluencies: faster speakers produced more repairs and repetitions, but fewer filled pauses. If faster speakers are more likely to produce repetitions than slower speakers then this may provide support for Blackmer and Mitton's (1991) proposal that the articulator possesses an autonomous restart capability. If, as Blackmer and Mitton suggest, the articulator repeats material that has previously been uttered when a delay occurs between conceptualisation and formulation, and articulation, then we may expect that when a person speaks faster the articulator may finish producing all available material before the plan for the remainder of an utterance has been prepared (because conceptualisation and formulation are, Blackmer and Mitton argue, relatively slow processes, compared to articulation).

It would be reasonable to suggest that the measure of articulation rate used by Finlayson et al. may not be the appropriate measure for testing Blackmer and Mitton's proposal. Articulation rates for each speaker were averaged across the duration of each conversation. As a result, faster speakers were those who tended to speak faster on average rather than those who were speaking faster during the utterance where the repetition was produced. In order to test the hypothesised autonomous restart capability, in our analyses we will test a predictor for per-utterance articulation rate. Doing so allows us not only to investigate whether there is a relationship between the rate at which an utterance is produced and the likelihood that it will contain a repetition, but also to control for possible relationships between rate of speech and disfluencies which may confound our investigation of whether or not hesitations are designed by speakers.

5.2.1 Methods

The corpus analyses presented in this chapter are based on the MTC dataset prepared following the steps described in Chapter 3.

Outcomes

Our analyses were concerned with four types of disfluency annotated in the corpus: *substitutions*, *insertions*, *repetitions* and *filled pauses* (see Chapter 3 for further details of corpus annotation). As the first two of these disfluencies appeared to serve a similar function of modifying, or eliminating, previously uttered material we grouped them together as a single category: *repairs*. Two other types of disfluency were annotated in the MTC: *deletions* and *complex* disfluencies. Deletions were not considered as they represent cases where a speaker abandons an utterance and it is not straightforward from the corpus annotation alone to determine whether the speaker abandoned the utterance because it was infelicitous (with this deletion therefore a disfluency) or whether the utterance was abandoned because the speaker was interrupted. As described in Chapter 2, complex disfluencies may contain multiple types of disfluencies. As a result, they present as a heterogeneous group which cannot be neatly assigned to existing types of disfluency.

The corpus analyses presented in this chapter took individual tokens as units-of-analysis. Tokens are each word, or fragment of a word, produced by speakers in the MTC. The use of individual tokens provide advantages over other possible units-of-analysis. One alternative would be to consider the number of disfluencies, or number of words appearing in a disfluent context, per 100 words produced by a speaker. However, this may lead us to violate the assumption of linearity in the general linear model (discussed in 3.2.1). Another alternative would be to divide each conversation into individual utterances and then code whether or not each contained a disfluency. While this need not violate the assumption of linearity (we could use the generalized linear model to analyse our data), our analyses would be insensitive to the extent of disfluency of an utterance. For example, if givers tended to produce two filled pauses per utterance, while followers tended to produce only one, then both would be considered similarly disfluent despite it being the case that givers are arguably more disfluent that followers (by virtue of producing more pauses).

One concern that could be raised about using tokens as a unit-of-analysis is that our results could be confounded by systematic differences in the numbers of tokens in each type of disfluency (e.g. while a filled pause has only one token, the *uh*, repairs or repetitions may contain an unlimited number of tokens). However, this is not a problem for the present study as we do not statistically compare different types of disfluencies in our analyses.

For each token in the MTC, we coded whether or not it appeared as a *junk token* (J-token) for each type of disfluency (i.e. whether it appeared in the reparandum of a repair or repetition, whether it was a filled pause). In line with Fox Tree (1995), J-tokens were defined as tokens which "[did] not add propositional content to [the] utterance" (p. 709). In the case of filled pauses, it is clear that the filled pause itself does not offer propositional content to the utterance (although it may be commenting on the propositional content, as a collateral signal; Clark, 1996). For repetitions, only one mention of the repeated token(s) is adding propositional content while the other is not. For the purposes of our analyses, it would make no difference whether the first or second mention of repeated tokens were treated as junk. Considering the first mention to be junk was a decision made to remain consistent with the annotation of the MTC. Finally, for repairs we would suggest that the propositional content of the reparandum is not the content that is intended by the speaker (otherwise it would not be repaired) and therefore we consider it to be junk.

For each type of disfluency (repairs, repetitions and filled pauses) a discrete outcome variable was produced which represented whether or not the token was a J-token for that type of disfluency.

Random effects

Three random effects were identified in our data: *speaker*, *partner* and *map*. As we examined each token of the corpus individually, our dataset had a relatively large number of observations. As the number of observations increases, the computational feasibility of testing models with large, complex, random effects structures decreases (as the hardware requirements and time taken for models to converge will increase). In response to this, random slopes were only tested for predictors-of-interest which varied within participants.[3]

[3] We note that in all models the variance associated with random effects for speakers is larger than that associated with random intercepts for other grouping factors, although there is no reason a priori reason to believe that the amount of random intercept and random slope variance should correlate.

Fixed effects

A list of fixed effects considered in each of our analyses is shown in Table 5.5. Our analyses tested a set of control predictors which were intended to account for potential confounds due to aspects of the MTC which were not of theoretical interest. Two of these were measures of the number of tokens produced by the speaker both in the current turn (Length$_t$) and in the full conversation (Length$_{conv}$), intended to account for a possible relationship between the length (or potential length) of an utterance and the probability that it may contain a disfluency (Shriberg, 1996). A third measure (by-participant, by-conversation, *mean reparandum length*; Length$_{rep}$), used the mean number of tokens appearing in the reparandum of a repair or repetition (but not a filled pause, as their "reparandum" length will always be 1) for each participant in each conversation to control for the fact that taking each token as our unit-of-analysis meant that participants with a tendency to produce longer reparanda would have higher disfluency counts overall. Three further control predictors quantified the participants experience with the task (overall, with each map, and across the length of a single conversation). We reasoned that as participants became more experienced with the task it might become less difficult, and they may become less likely to be disfluent as a consequence. One of these predictors, experience with the task, did not improve the fit of any of the models tested and is not discussed any further.

Table 5.5: Corpus analysis 1: Fixed effects tested in each analysis. Predictors-of-interest are shown in bold.

Predictor	Type	Range
Length$_t$ (# of words)	Continuous	1–133
Length$_{conv}$ (# of words)	Continuous	35–2615
Length$_{rep}$ (# of words)	Continuous	0–4
Current turn (t)	Continuous	1–478
Experience with task	Continuous	1–4
Experience with map	Discrete	First/Second time
Role	Discrete	Giver/Follower
Visibility	Discrete	Visible/Not visible
Familiarity	Discrete	Friends/Strangers
Gender	Discrete	Male/Female
Partner's gender	Discrete	Male/Female
Gender match	Discrete	Match/Mismatch gender
AR (in syll/sec)	Continuous	0.50–20.04

Three discrete predictors-of-interest were tested in each of our analyses: speaker's role in the task (giver vs. follower of instructions); the visibility of the partner (visible vs. not visible); and the prior familiarity between the speaker and their partner (friends vs strangers). As Branigan et al. (1999) reported numerical trends suggesting gender differences in the production of disfluencies in the MTC, we also tested three gender-related predictors-of-interest: the speaker's gender, the gender of their partner, and whether or not their genders matched. While a random slope for one of these was found to improve the fit of the model for repetitions, none of these gender-related fixed effects improved model fit. We therefore do not discuss gender any further.[4] Finally, we tested articulation rate as a fixed effect in our models. In part, this was to control for possible relationships between rate of speech and different types of disfluencies, observed elsewhere (Finlayson & Corley, 2012; Oomen & Postma, 2001; Siegman & Pope, 1965). As we suggested in Chapter 2, per-turn articulation rate may be more theoretically meaningful than per-conversation articulation rate. Therefore, we used the former as a measure in our analyses (AR). As it remains an open question how the rate at which a person speaks influences their fluency we treated articulation rate as a predictor-of-interest, and consequently tested a within-speaker random slope.

Before performing our analyses, each predictor was prepared as described in 3.2.2.

5.2.2 Results

In each analysis, model construction was performed following the steps described in Chapter 3. In line with the exploratory nature of the analyses, we report on only those fixed effects which significantly improved the fit of each model.

Corpus Analysis 1a: Repairs

See Table 5.6 for the full model of the likelihood of producing a repair J-token. As would be expected, the tokens produced by speakers with longer mean reparanda were more likely to be repair J-tokens. The likelihood of producing a repair J-token was also found to increase as speakers planned and produced longer utterances, and as they produced more tokens, overall, throughout the conversation.

[4] Gender differences reported in the past (e.g., Binnenpoorte, Bael, Os, & Boves, 2005; Bortfeld et al., 2001; Lickley, 1994; Shriberg, 1994) may be the consequence of differences in approaches taken to analysis, for example in the types of regression or units-of-analysis used, or of the larger corpus sizes of up to 300,000 words (Binnenpoorte et al., 2005).

Table 5.6: Corpus Analysis 1a: Logistic mixed-effects model of the probability of a given token being a repair J-token. Fixed effects are given in the order in which they were included in the model. Predictors-of-interest are shown in bold.

					Random effects		
Fixed effect	β	SE	z	$p(\beta = 0)$	Group	Predictor	Variance
Intercept	-4.411	0.064	-68.513	$< .001$	Partner	*Intercept*	< 0.001
Length_t	0.269	0.018	14.566	$< .001$	Speaker	*Intercept*	0.198
Length_{rep}	0.146	0.041	3.593	$< .001$		Friend	0.231
Length_{conv}	0.131	0.058	2.259	$< .05$		Giver	0.300
Map Experience	0.140	0.079	1.770	.08		AR	0.047
Giver	0.351	0.118	2.979	$< .01$			

Table 5.7: Corpus Analysis 1b: Logistic mixed-effects model of the probability of a given token being a repetition J-token. Fixed effects are given in the order in which they were included in the model. Correlation was fitted between random intercept for *speaker* and random slope.

					Random effects			
Fixed effect	β	SE	z	$p(\beta = 0)$	Group	Predictor	Variance	Correlation
Intercept	-4.311	0.068	-63.276	$< .001$	Map	*Intercept*	0.011	-
Length_t	0.091	0.019	4.711	$< .001$	Partner	*Intercept*	0.048	-
Length_{conv}	0.176	0.044	3.980	$< .001$	Speaker	*Intercept*	0.178	-
Length_{rep}	-0.073	0.033	-2.220	$< .05$		Giver	0.232	-
						Matching gender	0.125	-
						AR	0.042	0.569

Speaker's experience with the map significantly improved the fit of the model; however, its coefficient was only marginally significant in the final model. After controlling for these trends, a significant effect of role was found: Givers of instructions were almost one-and-a-half times as likely to produce a repair J-token as were followers ($p < .01$), $\beta = 0.351$ (OR = 1.42).

Corpus Analysis 1b: Repetitions

See Table 5.7 for the full model of the likelihood of producing a repetition J-token. Repetition J-tokens were more likely to be produced during longer conversational moves and by speakers who produced more tokens. A relationship was found between mean reparandum length and the likelihood of producing a repetition J-token, with repetition J-tokens more likely to be produced by speakers with shorter mean reparanda. After controlling for these trends, none of our predictors-of-interest were found to significantly improve the fit of the model.

Table 5.8: Corpus Analysis 1c: Logistic mixed-effects model of the probability of a given token being a filled pause J-token. Fixed effects are given in the order in which they were included in the model. Predictors-of-interest are shown in bold.

						Random effects	
Fixed effect	β	SE	z	$p(\beta = 0)$	Group	Predictor	Variance
Intercept	−5.066	0.095	−53.127	< .001	Partner	*Intercept*	< 0.001
Length_t	0.085	0.023	3.608	< .001	Speaker	*Intercept*	0.455
t	−0.079	0.033	−2.380	< .05		Friend	0.205
Length_{conv}	0.057	0.075	0.771	.44		Giver	0.526
AR	−0.864	0.036	−23.735	< .001		AR	0.018
Giver	0.678	0.135	5.018	< .001			

Corpus Analysis 1c: Filled pauses

See Table 5.8 for the full model of the likelihood of producing a filled pause J-token. As each incidence of a speaker producing a filled pause can only have a length of one, the likelihood of producing a filled pause J-token is equivalent to the likelihood of a speaker producing a filled pause. Speakers were found to be more likely to produce filled pauses during longer conversational moves. An effect of turn was found, such that the likelihood of producing a filled pause decreased across the length of a conversation. The number of tokens that a speaker produced during the conversation significantly improved the fit of the model; however, its coefficient failed to reach significance in the final model. Givers of instructions were found to be almost twice as likely to produce a filled pause as were followers ($p < .001$), $\beta = 0.679$ (OR = 1.97). Finally, an effect of articulation rate was found, with filled pauses less likely to occur when speakers spoke slowly ($p < .001$), $\beta = -0.864$.

5.2.3 Discussion

The present study was intended to investigate whether the production of certain disfluencies showed sensitivities to manipulation of the situation in which a dialogue took place, which would be consistent with the claim that speakers design hesitations for their audience. Our analyses focused on three types of disfluencies, repairs, repetitions and filled pauses; with each analysis modelling the likelihood that a given token in the MTC would be a junk token (J-token) in each of these three types.

Using mixed-effects regression, which allowed us to control for anticipated confounds in noisy corpus data, we explored the effects on the production of J-tokens of three factors manipulated in the design of the MTC: Whether a given speaker was a giver or follower of instructions, *role*; whether pairs of speakers were able to see each other, *visibility*; and whether they were friends or strangers prior to the dialogue, *familiarity.* Performing the giver role in the MTC results in greater cognitive burden than performing the follower role, with this difficulty likely to lead to an increase in how disfluent speakers are. As expected, givers were more likely to produce both repair and filled pause J-tokens than were followers, although no effect of role was found in the case of repetitions. While *role* engenders a systematic difference in the difficulty that speakers in the MTC face, we would argue that such systematic effects should occur not with the *visibility* and *familiarity* manipulations; rather, any effects found for these two factors may be consistent with the claim that certain disfluencies are being designed by speakers to perform communicative functions. In particular, we expected that speakers who were unable to see each other would be more likely to produce filled pauses and repetitions, as they would be unable to rely on non-verbal strategies to manage turn-taking. Neither *visibility* nor *familiarity* were found to significantly improve the fit of the models tested for each type of disfluency. Before discussing these results further, we first discuss some other trends observed.

Speech rate and disfluency

In a similar set of analyses of disfluencies in the MTC, Finlayson et al. (2010) investigated the relationship between articulation rate and the production of repairs, repetitions and filled pauses. One of the motivations for their analyses was to test Blackmer and Mitton's (1991) proposal that the language production system possesses an autonomous restart capability. Blackmer and Mitton suggested that if, during speech, processes of planning fail to be completed before the articulator is ready to produce the plan then the articulator may repeat the last part of the utterance that it produced. Finlayson et al. predicted that one consequence of this proposed autonomous restart capability would be that fast speech may be associated with an increased likelihood of producing repetitions, as the articulator may be more likely to finish production before the next plan is ready. Consistent with this, they found that faster speakers were more likely to produce repetitions, as well as being more likely to produce repairs and less likely to produce filled pauses. However, we argued that a more appropriate test of the autonomous restart capability is not whether faster speakers produce more repetitions but whether speakers produce more repetitions when speaking faster. With this in mind, as well as to control for other possible relationships between rate of speech and disfluencies, we tested articulation rate in each of our models. While, for repairs and repetitions, there were by-speaker random slopes for articulation rate (suggesting that some speakers produce more of these disfluencies when speaking faster, whilst others produce more when speaking slower), only in the case of filled pauses did a fixed effect of articulation rate significantly improve the fit of our model: When participants spoke faster they were less likely to produce filled pauses than when they spoke slower.

While a similar trend for filled pauses was observed by Finlayson et al., we did not replicate their significant effects of articulation rate on repairs and repetitions. There are two possible explanations for this. The first is that per-turn and per-speaker, per-conversation, measures of articulation rate may be qualitatively different, and, from a statistical perspective, variance in the production of repairs and repetitions accounted for by one measure is not accounted for by the other. The second is that the models tested in the present study may better control for confounds which were present in Finlayson et al.'s original analysis. The models constructed in their study included three of the control predictors tested in the present study (turn and conversation token count, and mean reparandum length) but they did not include any of the three control predictors controlling for speakers' experience that were tested in the present study. Additionally, their model construction process did not include testing of random slopes for their predictors-of-interest.

In order to decide between these two explanations we tested per-speaker, per-conversation, articulation rate in our final models for repairs and repetitions. If the addition of this measure improved model fit then this would offer support for the first explanation; however, if the addition did not improve model fit then it would suggest that the trends reported by Finlayson et al. were confounded by noise that was controlled for in the present study. A log-likelihood ratio test showed that the addition of Finlayson et al.'s measure of articulation rate did not significantly improve the fit of models for either repairs ($\chi^2 = 0.735$, $p = .39$) or repetitions ($\chi^2 = 1.665$, $p = 0.20$). This lack of significant effects for per-speaker, per-conversation, articulation rate suggests that the trends for faster speakers to produce more repairs and repetitions, reported by Finlayson et al., were driven by confounds which were better controlled for in the present study, rather than a qualitative difference between the measures of articulation rate used in each study. Further support for the idea that there is not a qualitative difference between these two measures comes when we consider the observed effect articulation rate on filled pauses.

In contrast to Oomen and Postma (2001) and Shriberg (1994), but consistent with Finlayson et al., our analyses showed that when participants spoke faster they were less likely to produce filled pauses than slower speakers. One explanation for filled pauses being more likely to occur during slower speech may be that both phenomena (i.e. filled pauses and slow speech) share a similar cause. It is well established, both from the present study and the wider literature, that the likelihood of producing filled pauses increases when a speaker experiences cognitive difficulty. Similarly, the finding presented in Chapter 7, that givers of instructions tended to speak slower than followers, suggests that cognitive difficulty has an effect on rate of speech. While the observed relationship between articulation rate and filled pauses is not mediated by speakers' role, as this factor is controlled for in our model, unaccounted-for moment-to-moment causes of difficulty (e.g. factors related to lexical access, such as predictability) may covary with articulation rate in our analysis. Future research could explore this possibility by accounting for further sources of difficulty in the MTC and observing whether an effect of articulation rate on filled pauses remains.

Other influences on disfluency

Consistent in each of our analyses was the finding that the likelihood of producing J-tokens of each type of disfluency was elevated in longer utterances. These trends are consistent with previous literature suggesting a relationship between utterance length and the likelihood of a speaker being disfluent (e.g. Oviatt, 1995; Shriberg, 1996). We would, however, raise a note of caution about interpreting that the burden of planning longer utterances was causing speakers in the MTC to be disfluent. Rather than longer utterances being more likely to contain disfluencies, it may simply be that the additional tokens that disfluencies engender are exaggerating the recorded lengths the utterances. For example, a fluent utterance of 13 tokens, "Where is the top of the lemon grove in relation to the pyramid?", would instead contain fourteen tokens if it contained a filled pause, "Where is the top of the lemon grove *uh* in relation to the pyramid?". Regardless of whether the observed trends are due to the burden of planning longer utterances or a confound originating from our use of single tokens as a unit of analysis, what is important for the present study is that these trends are controlled for when we come to interpret our predictors-of-interest. As the nature of the relationship between utterance length and disfluencies is not of interest to this study, we will not attempt to test these two explanations of the observed trend.

Our analyses found little evidence that experience with aspects of the MTC, either the map being discussed or the map task itself, had an influence on participants' fluency. Participants' experience with the map was found to significantly improve the fit of the model for repairs; however, in the final model its coefficient was only marginally significant. If speakers were, in fact, more likely to produce repair J-tokens during their second conversation about a map then this may seem counter-intuitive, as we might imagine that it would be easier to describe the map on the second attempt. An alternative explanation for such an effect is that with past experience of describing the map speakers have an increased awareness of the right and wrong way to describe landmarks, and produce more repairs to refine their descriptions.

While we did not find any significant effects of between-conversation experience, we did find a significant effect of turn, such that speakers were less likely to produce filled pauses as a conversation progressed. This may reflect participants finding the task less difficult as the dialogue proceeds, perhaps because of a common ground being developed (e.g. Clark, Schreuder, & Buttrick, 1983; Clark & Wilkes-Gibbs, 1986). The lack of any significant effects of task experience may suggest further that this common ground must be redeveloped when performing the task with a new partner.

In order to prevent the likelihood of producing a J-token being confounded by participants who tended to produce longer reparanda (the parts of speech which are repaired or repeated), each speaker's mean reparandum length was included as a control predictor. As would be expected, longer mean reparanda were associated with an increased likelihood of producing repair J-tokens; however, the opposite trend was observed for repetition J-tokens. This latter trend may appear counter-intuitive: If our mean reparandum length predictor is a measure of the number of tokens repeated (as well as the number repaired), then we ought to expect that speakers with longer mean reparanda will be more likely to produce repetition J-tokens. One explanation for the observed trend could be that speakers with shorter mean reparanda were tending to produce a greater proportion of repetitions to repairs. If this is the case then we would expect that the reparanda associated with repetitions should be shorter than those associated with repairs.

Using linear regression which regressed the length of reparanda against whether a disfluency was a repetition, rather than a repair, we found evidence in support of this claim. The length of reparanda for repetitions was found to be significantly shorter than that of repairs ($\beta = -0.491, t = -4.472, p < .001$). Furthermore, consistent with the claim that our measure of mean reparandum length represents the ratio of repetitions to repairs, our results suggest that the measure is unrelated to the production of filled pauses.

Are hesitations designed?

For both repairs and filled pauses, our analyses showed that speakers were more likely to produce J-tokens when they were givers of instructions rather than followers. Disfluencies are known to be associated with difficulties experienced during planning of utterances and lexical access (e.g. Hartsuiker & Notebaert, 2010; Schachter et al., 1991; Schnadt, 2009), and we argued that the requirement of the giver role to take the lead in planning utterances that are appropriate for their partner would involve greater cognitive burden than that experienced by their partner, for example, because givers formulate utterances which can subsequently be reused by followers.

We found no evidence, however, that givers were any more likely to produce repetition J-tokens than followers. The existing literature provides mixed evidence for a relationship between difficulty and repetitions. Using the network task (Oomen & Postma, 2001) with Dutch speakers, Hartsuiker and Notebaert (2010) found a frequency effect for repetitions; however, this effect was restricted to the frequency of determiners, while manipulation of the name agreement of the images described had no effect on the production of repetitions. Schnadt's (2009) experiments using the network task similarly did not show evidence that difficulty in lexical access influenced the number of repetitions produced by speakers, although he observed very few repetitions overall (occurring in $< 2\%$ of utterances in all experiments), raising the possibility that the experiments lacked the power to detect an effect. Given the size of the MTC, it is unlikely that the absence of any clear effect of difficulty on the likelihood of producing repetition J-tokens[5] is due to a lack of power. Rather, we would conclude that either the frequency

[5] Our analyses did show that longer utterances were more likely to contain repetition J-tokens however, for reasons discussed in 5.2.3, we cannot be sure whether this effect is due to the burden of producing longer utterances or whether it is just due to repetitions extending the length of an utterance.

differences exploited by Hartsuiker and Notebaert (2010) are not present in English or that givers were no more likely, after controlling for the number of tokens produced, to produce low frequency determiners than were followers.

Finding that sources of difficulty, for example speaker's role, are associated with increased disfluency is entirely consistent with accounts that suggest that hesitations are either symptoms, or signals, of difficulty, as the difficulties that could automatically result in hesitations may be the same difficulties which speakers may wish to signal. In order to differentiate between these two accounts we suggested that if *visibility* and *familiarity*, two factors which we argued would not systematically engender differences in difficulty, were found to influence the production of J-tokens then it may suggest that these tokens were being produced as signals of difficulty, rather than as symptoms. Although we observed several random effects for these factors, suggesting that certain speakers may have been more likely to have produced J-tokens in certain conditions, when they were tested as fixed effects neither was found to significantly improve the fit of the models for any of the types of disfluency considered.

Following Maclay and Osgood's (1959) and Clark and Wasow's (1998) suggestions that speakers use certain hesitations to manage turn-taking, we reasoned that when partners were unable to see each other, which has been shown to influence turn-taking in the MTC (Boyle et al., 1994; Bull & Aylett, 1998), they may switch from non-verbal turn-taking cues (such as gaze or gestures; e.g. Duncan, 1972; Kendon, 1967) to verbal strategies. With a partition depriving partners of the ability to provide non-verbal cues, speakers would be expected to use filled pauses and/or repetitions to ensure that they retained their turn until they had finished. However, we found no evidence to suggest that filled pauses were any more likely to occur when partners were unable to see each other.

Our analyses of disfluencies in the MTC did not find evidence to suggest that speakers' production of hesitations varied in manners consistent with them being designed with a communicative function. Only speaker's *role* was found to influence the production of J-tokens, with this effect observed for repairs and filled pauses. Given that different roles in the MTC entail differences in the cognitive difficulties faced by speakers, the same types of difficulties which are known to be associated with the production of certain disfluencies, a parsimonious account of these findings would be that speakers were more likely to be disfluent when faced with difficulty because disfluencies are an automatic consequence of difficulty.

5.3 General discussion

In this chapter we presented an experiment, and an analysis of a corpus of task-orientated dialogue, which were intended to evaluate hesitations against different aspects of Kraljic and Brennan's (2005) third criterion for recognising a feature of speech as being designed by a speaker for their audience.

In Experiment 2, we reasoned that if speakers produce hesitations for the benefit of their audience then the likelihood that they will produce hesitations would be expected to decrease in the absence of an audience. Participants described pictures which were either easy-to-name (high agreement, high frequency) or hard-to-name (low agreement, low frequency), either alone or as part of a card sorting task with an interlocutor. While speakers' language was influenced by the presence of an interlocutor (as indexed by their choice of referring expressions), they were no more likely to produce hesitations with an audience than without one.

In Corpus Analysis 1, we reasoned that if speakers use hesitations to manage aspects of conversation then changes to the situation in which a dialogue takes place, which may change the aspects that need to be managed, or the means by which they can be managed, may change the nature of the hesitations they produce. Using the Map Task Corpus, a corpus consisting of dialogues between pairs of participants taking turns to direct each other through a route-finding task, we explored whether manipulation of *visibility* (whether or not participants could see each other) or *familiarity* (whether partners were friends or strangers) had an influence on the disfluencies that speakers produced. Our analyses found no evidence that either factor was having any systematic effect on the production of either repairs, repetitions, or filled pauses.

Neither the experiment nor the corpus analysis provided any evidence to suggest that speakers produce hesitations in a manner which would suggest that they meet Kraljic and Brennan's (2005) third criterion. However, in both studies we observed clear associations between the difficulties experienced by speakers and their likelihood of producing certain disfluencies. In Experiment 2, participants were more likely to produce hesitations when naming items with low name agreement and low frequencies, consistent with existing findings suggesting that certain disfluencies may be associated with difficulty in lexical access (e.g. Hartsuiker & Notebaert, 2010; Schnadt, 2009). In Corpus Analysis 1, participants were more likely to produce repairs and filled pauses (but not repetitions) when they filled the role of giver of instructions rather than follower. We have argued above that this role places greater planning demands on speakers, and the finding that givers were more likely to produce certain disfluencies is consistent with previous studies showing an association between the production of hesitations and choice (e.g. Schachter et al., 1991), and with leadership in cooperative tasks (e.g. Bortfeld et al., 2001).

Earlier, we suggested that one reason that it has been difficult to differentiate between a symptom and a signal account of hesitations is that those difficulties which are likely to cause hesitations are those that a speaker may want to signal to their listener. Using Kraljic and Brennan's (2005) third criterion, this chapter presented two distinct attempts to test predictions which should differentiate between these two accounts. Both cases failed to show any evidence to support the hesitation-as-signal hypothesis; however, both studies did show clear associations between cognitive difficulties and the likelihood of producing certain disfluencies, including hesitations. Whilst we ought to always be cautious in the interpretation of null results, especially when the object of study is a relatively rare phenomenon, we would suggest that the results of both of the studies presented in this chapter are entirely consistent with the parsimonious account that hesitations are an automatic symptom, rather than a signal, of difficulty.

One of the functions that hesitations have been argued to perform is to allow a speaker to keep hold of their conversational turn when their speech was disrupted. In the following part of this thesis we turn our attention to what happens when another party in the conversation takes a turn. Here, another aspect of how speech is performed has been implicated in having a role in the coordination of conversation. In their oscillatory theory, Wilson and Wilson (2005) argue that the smooth transitions between turns, argued to be commonplace in conversation, is achieved through entrainment of the rates at which conversational partners speak. In the following chapter, we will review the oscillator theory, as well as earlier theories of turn-taking which have informed it, before presenting a set of corpus analyses which test several predictions derived from the theory.

Part II

Taking a turn

Chapter 6

Theories of turn-taking

It is one of the defining features of conversation that the parties involved take their turn to speak. Often one party will not begin a turn until the previous turn has ended, and the gaps between these turns are often so short as to be imperceptible. Turn-taking appears to proceed seamlessly, like the movements of a dance performed by experienced dancers. Unlike a dance, however, there is little or no choreography behind the smooth organisation of turn-taking. The order of who will speak, how long they will talk for, and what they say, is not predetermined. Rather, these aspects must be managed by parties on the fly during the conversation. Not surprisingly, it has long been of interest how turn-taking is managed, and in particular how the smooth timing of turn-taking is achieved in conversation. In this chapter we will review evidence that has been collected, and theories that have been developed, over the past fifty years which have offered explanations of how turn-taking in conversation comes to be so precisely organised.

One recent theory proposes that the precision exhibited in the timing of turn-taking is achieved by the entrainment of one particular aspect of linguistic performance, rate of speech (Wilson & Wilson, 2005). This theory is informed by two older theories of turn-taking. One that suggests that conversational partners are able to anticipate that a turn is likely to end (and are therefore able plan to produce a subsequent turn), and another that suggests that conversational speech contains a variety of different types of cues that partners can use to determine when a turn has or is about to end. According to Wilson and Wilson's theory, partners may rely on a variety of different cues in order to make coarse predictions about when a turn is likely to end. These predictions can then be "refined" by the precision timing that Wilson and Wilson argue is afforded by the entrainment of rate speech.

The structure of this chapter is as follows. We will first discuss these two early theories of turn-taking which have influenced most of the subsequent work on the subject. Both of these theories place importance on different cues that listeners may use to anticipate, or react to, the end of a turn, and as such we will then review evidence of the types of cues that may be used in turn-taking. Finally, we will present Wilson and Wilson's theory of the timing of turn-taking.

Before continuing, we will make a few brief notes on terminology. Throughout this part of the thesis, we will use the term *turn exchange* to refer to the change from one speaker to another in a conversation. The duration between the first speaker finishing their turn and the second speaker beginning their turn will be referred to as the *inter-turn interval* (hereafter, ITI). When the second turn begins following the end of the first turn (when ITI ≥ 0) we will refer to this as a *gap*. While, when the second turn begins prior to the end of the first turn (when ITI < 0), producing overlapping speech, we will refer to this as an *overlap*. Consistent with the literature (e.g. Duncan, 1972; Kendon, 1967), we will use the term *auditor* to refer to all parties in a conversation that are not currently speaking (i.e. those people who could begin the next turn). In two-party conversations, such as in our own analyses presented in the following chapter, the auditor will always be the next person to speak; however, in conversations with a greater number of participants the auditor is everyone other than the current speaker (regardless of whether or not they are the next person to speak). Finally, for consistency with the oscillator theory that we test in the subsequent chapter, throughout this part of the thesis we will refer to articulation rate as syllable rate.

6.1 Early theories of turn-taking

We may expect that, given the flexibility in the content and structure of conversation, turn-taking would be problematic for people. Many aspects of conversations may vary. Before a conversation begins it is often not known how many turns will occur during the conversation or how long the conversation will last. When a speaker begins a turn, interlocutors may not know how long the turn will last, nor what its contents will be. Finally, while it is often (but not always) known how many parties will be involved in a conversation, neither the order in which each party will speak nor the distribution of turns (i.e. how many each party will take) are decided in advance. Rather, these matters are often managed one turn at a time. Despite its free-form nature, described elsewhere as "anarchistic"

(Wilson & Wilson, 2005, p. 957), conversation generally gives the impression of being very well organised. A finding consistent across many studies has been that many ITIs fall in a range of 0–200 ms (e.g. Beňuš, 2009; De Ruiter, Mitterer, & Enfield, 2006; Heldner & Edlund, 2010; Wilson & Wilson, 2005; for evidence that this tendency holds across a variety of languages and cultures, see Stivers et al., 2009), where they may frequently be imperceptible to listeners (i.e. the turn exchange is perceived as a smooth transition; Walker & Trimboli, 1982). Given this seamlessness, it is of interest how people manage to achieve such organisation of turn-taking in a type of social interaction which would appear so difficult to organise, and in particular how people know when one turn will end with enough precision that they are able to produce a subsequent turn so quickly.

One of the earliest and most influential theoretical treatments of turn-taking in conversation was provided by Sacks, Schegloff, and Jefferson (1974). After spending over half a decade collecting recordings of conversations they reported anecdotal evidence that confirms not only the anarchic nature of conversation, but also how well it is organised.[1] Sacks et al. offer a set of "grossly apparent facts" (p. 700–701) which any theory of turn-taking should accommodate:

1. Speaker-change recurs, or at least occurs.
2. Overwhelmingly, one party talks at a time.
3. Occurrences of more than one speaker at a time are common, but brief.
4. Transitions (from one turn to a next) with no gap and no overlap are common. Together with transitions characterized by slight gap or slight overlap, they make up the vast majority of transitions.
5. Turn order is not fixed, but varies.
6. Turn size is not fixed, but varies.
7. Length of conversation is not specified in advance.
8. What parties say is not specified in advance.
9. Relative distributions of turns is not specified in advance [i.e. the number of turns to be produced by each party is not predefined].
10. Number of parties can vary.

[1] See O'Connell, Kowal, and Kaltenbacher (1990) for a critique of Sacks et al.'s use of anecdotal evidence.

11. Talk can be continuous or discontinuous [i.e. sometimes the end of one turn is followed by the beginning of another, at other times a turn ends with nobody stepping in to continue].
12. Turn-allocation techniques are obviously used. A current speaker may select a next speaker (as when he addresses a question to another party); or parties may self-select in starting to talk.
13. Various 'turn-constructional units' are employed; e.g. turns can be projectedly 'one word long', or they can be sentential in length.
14. Repair mechanisms exist for dealing with turn-taking errors and violations; e.g. if two parties find themselves talking at the same time, one of them will stop prematurely, thus repairing the trouble.

In Sacks et al.'s (1974) theory, turns are made up of *turn-constructional units* (TCUs). What constitutes a TCU can vary from individual words (e.g. "yes" or "no", in response to a polar question), through to phrasal, clausal and sentential constructions. On the basis of evidence suggesting that turns tend to begin at points of syntactic completion (rather than beginning at any point in an utterance), Sacks et al. suggest that syntax is an important source of information for demarcating TCUs (subsequent research has further demonstrated the importance of syntax in turn-taking, e.g. Ball, 1975; Caspers, 2003; Gravano, 2009; Koiso, Horiuchi, Tutiya, Ichikawa, & Den, 1998; Wennerstrom & Siegel, 2003). However, Sacks et al. also suggest that intonation may play a role (for example, whether or not a single word can serve as TCU may depend on its intonation contour), although they generally give very little attention to intonation.

Sacks et al. claim that when a speaker begins a turn they are entitled to produce one TCU (although they may produce more). At the end of each TCU is a *transition-relevance place* (TRP). A TRP is a point at which the speaker may opt to yield their turn to a different party in the conversation. While, at a TRP, a speaker may choose to continue and produce another TCU, if they wish to yield their turn then there is a set of rules which describe the process by which the next speaker is selected:

1. The original speaker may select the next speaker, giving that chosen speaker the obligation to take a turn.

2. If the original speaker does not select the next speaker then any party in the conversation may self-select. The first party to self-select then has the right to begin the next turn.
3. If no other party self-selects then the original speaker has the right to continue.
4. If the original speaker does not continue then the options recycle back to 2.

Wilson and Zimmerman (1986) report evidence which they argue supports the existence of such selection rules. They reasoned that if options 2 and 3 are open to parties for similar periods of time then the ITIs in conversation should be multiples of this duration. This would be because only one option is available at any point, therefore the first option would have to have been open and passed before the second option could open. Quantitatively, this would mean that the distribution of ITIs should exhibit a periodicity, peaking in frequency at recurring multiples of these durations. They tested this prediction using ITIs taken from recordings of seven nine-minute conversations between dyads. It was expected that there would be more short ITIs than long ITIs, they therefore took the step of "differencing" the ITIs (ranking each in increasing order of size and then subtracting each value from the value that followed; e.g. $2-1$, $3-2$, $4-3$, etc.) in order to remove a possible linear decline in frequency. After this step, they used a set of time-series analyses which confirmed that there was periodicity in the distribution of ITIs, with a period of 120 ms on average. Wilson and Zimmerman suggest that the period of the ITIs represents the duration for which each option is open to parties in the conversational. While such a conclusion would appear to us to be rather premature (e.g. we know of no evidence that the strictly serial set of options proposed by Sacks et al. are present in conversation, nor were these results confirming a prediction made by Sacks et al. about the duration for which each option would be open), their results are at least consistent with the claim that parties in conversation cycle through selection options at turn exchanges.

Further evidence argued to be compatible with the existence of these options comes from a study of turn-taking in conversations of more than two people (Wennerstrom & Siegel, 2003). Wennerstrom and Siegel examined all TRPs in each conversation and found a non-linear relationship between ITIs and the probability of a turn exchange taking place. Exchanges were found to be more likely at very short ITIs, decreasing as ITIs increased to 500 ms, before increasing again. They suggested that just after 0 ms we see the first two options being exercised (the next speaker self-selecting, or being selected) with the dip in likelihood at 500 ms reflecting the original speaker exercising the third option by beginning a new TCU, and the subsequent increase in likelihood reflecting the recycling to option 2 where a new speaker can begin a TCU. While the authors suggest that their results are consistent with Sacks et al.'s rules for self selection, we would note that, assuming each option is open for the same amount of time, the dip at 500 ms is inconsistent with Wilson and Zimmerman's suggestion that each option is open for 120 ms (Wilson and Zimmerman may predict a dip at 240 ms, rising again towards 360 ms). One possible explanation for this inconsistency could be that the 120 ms is not universal, and that different durations occur in different types of conversations (e.g. different numbers of participants, different relationships between participants, different topics, etc.) in different contexts. Unfortunately, neither author provides sufficient details of their data collection to allow us to identify all of the systematic differences between their conversations.

Sacks et al.'s (1974) system provides a possible answer to the question of how speakers manage to achieve the minimal gaps and minimal overlaps which are common in conversation (if they even produce a gap or overlap at all). Because one of the rules of the system is that turns must be constructed of one or more complete TCUs, and that a TRP occurs at the end of each TCU, it is possible for auditors to project the end of a turn, because they know that the completion of the TCU could result in the end of the turn (although when exactly such a projection could be made will depend much on the content of the TCU). Because auditors, it is argued, are able to project that a turn will end, they should be able to plan so that they can begin a new turn of their own close to the end of the previous turn (leaving only a minimal gap or overlap). This capability for projection is in contrast to reactive ideas of turn-taking which were being proposed at a similar time, most prominently in the work of Duncan and colleagues (1972, 1974; Duncan, Brunner, & Fiske, 1979; Duncan & Niederehe, 1974). In Duncan's account of turn-taking, it is assumed that speakers produce cues that they wish to yield their turn, to which auditors can then react. Duncan uses the term "signals" to refer to the cues that are being produced; however, as Cutler and Pearson (1986) suggest, it is not always obvious that Duncan is asserting that they are signals in a Gricean sense (1957). While Cutler and Pearson prefer to term them "correlates of end of speaking turn", for the sake of brevity we will refer to them as cues, except where directly quoting Duncan.

Duncan focuses on six possible types of cue which may be involved in turn-taking: (1) The completion of a grammatical clause; (2) the termination of any hand gestures; (3) sociocentric sequences, such as "but uh", "or something" or "you know", that "do not add substantive information to the speech content that they follow" (p. 287); (4) any phrase-final intonation that is not a sustained, intermediate pitch level, with neither rising nor falling intonation; (5) a drawl on the final syllable of a terminal clause (i.e. a clause with either rising or falling intonation); and (6) a drop in pitch and/or intensity in conjunction with a sociocentric sequence. Although it is not made clear how Duncan arrived at these six categories, Cutler and Pearson (1986) suggest that they may have been identified through the somewhat circular approach of identifying turn exchanges and then examining what behaviours were being exhibited by the speaker yielding their turn.

In his account, Duncan (1972) proposes two rules which specify the correct usage of different cues in the organisation of turn-taking. The first of these rules says that a speaker who is ready to yield their turn will produce one or more of the six cues described above:

> "The auditor may take his speaking turn when the speaker gives a turn-yielding signal. Under proper operation of the turn-taking mechanism, if the auditor acts to take his turn in response to a yielding signal by the speaker, the speaker will immediately yield his turn." (p. 286)

Duncan further suggests that speakers may produce an attempt-suppressing cue to stop an auditor's attempt to take a turn. Such cues are, it is argued, able to override turn-yielding signals. In Duncan's account, engaging one or both hands in gesticulation forms the attempt-suppressing cue, and auditors should respond to such gesticulation by ceasing their attempt to begin a turn:

> "An attempt-suppressing signal displayed by the speaker maintains the turn for him, regardless of the number of yielding cues concurrently being displayed. Auditors almost never attempted to take their turn when this signal was being displayed." (p. 287)

From the two rules described above, Duncan derived two predictions about the effects of cues on turn-taking. As speakers produce turn-yielding cues when they wish to yield their turn, then, conversely, if they do not wish to yield their turn then they will not produce turn-yielding cues. The first prediction was therefore that if the auditor begins overlapping with the speaker then the speaker must not have produced any turn-yielding cues (because if they had then they would have yielded the turn and stopped speaking, and there would therefore have been no overlap). The second prediction was that as speakers are able to "cancel" their turn-yielding cues with an attempt-suppressing cue then auditors should be less likely to make a turn-taking attempt when a turn-yielding cue is followed by an attempt-suppressing cue than when it is not.

In order to test both of these predictions, Duncan (1972) video-recorded two interviews between psychotherapists and their clients. The first 19 minutes of each of these interviews were transcribed, and extensively coded for the occurrence of vocal and bodily gestures. Consistent with his first prediction, overlaps were more likely to happen occur when the speaker did not produce a turn-yielding cue. Furthermore, it was found that the percentage of turn-taking attempts by the auditor increased as the number of cues being concurrently produced increased. Although Beattie (1981) has shown that Duncan's study was badly underpowered (if only one instance of a turn-taking attempt after the production of six cues had occurred then the strength of the correlation between number of cues and likelihood of a turn-taking attempt would be reduced dramatically), this relationship has been replicated elsewhere (Gravano, 2009; Hjalmarsson, 2011) Taken together, these results suggests that auditors respond to turn-yielding cues as if they were an invitation to begin a new turn, with increasing numbers of cues produced concurrently forming a "stronger" invitation to take a turn.

Duncan also examined the production of attempt-suppressing cues. Consistent with the second prediction that such cues would override turn-yielding cues, there were fewer turn-taking attempts when turn-yielding cues co-occurred with attempt-suppression cues than when they did not. This would suggest that the cues were successfully suppressing turn-taking attempts.

While Yngve (1970) suggested that the smoothness of turn-taking must mean that conversational partners were exchanging signals, the high frequency of short ITIs (minimal gaps and minimal overlaps) has been viewed as an important source of evidence in support of projection, such as in Sacks et al.'s (1974) theory, over reaction as an explanation of the general mechanism that underlies turn-taking. For example:

> If turn taking were reactive, these brief transitions and slight overlaps should not have occurred. It is cognitively impossible to react to a stimulus in less than 0.2 seconds and logically impossible to do so before the stimulus even exists. (Clark, 1996, p. 322)

However, it has recently been called into question whether or not in general turn-taking is actually so quick that it could not possibly be reactive (Heldner & Edlund, 2010). Heldner and Edlund examined ITIs (including overlapping turns) across three corpora (including the MTC), in three languages (English, Dutch and Swedish). There were two motivations for their analyses: firstly, to establish how common were 0 ms inter-turn intervals, so called no-gap-no-overlap intervals, and secondly, to establish how many inter-turn intervals were greater than 200 ms, the point at which, they argued, the auditor could possibly be reacting to a signal. In each of the three corpora, and using several measures of central tendency, the averages of ITIs were found to be greater than 0ms. Across all of the corpora, 41%–45% of ITIs were found to be over 200 ms. With a sizeable number of intervals being the value claimed as a minimum reaction time, Heldner and Edlund suggested that either turn-taking is reactive or that the projections that auditors make are imprecise, and later concluded that evidence of the distribution of ITIs can neither be used as evidence for projection or against reaction (it is not clear how the authors would explain the finding that the majority of ITIs were under 200 ms).

The estimate that it will take 200 ms for an auditor to begin speaking in response to a turn-yielding signal was obtained from experimental tasks where participants were instructed to produce a neutral vowel as quickly as possible in response to a cue (e.g. Fry, 1975; Izdebski & Shipp, 1978, cited in Heldner & Edlund, 2010). There are obvious differences between producing a single phoneme in a reaction time test and producing entire utterances in conversation, not least the differences in length and, consequently, syntactic complexity, which have been shown to affect how long it takes to begin speaking (Ferreira, 1991). Heldner and Edlund (2010) recognise that the reaction time task is relatively simple, and cite evidence from a more complex task, producing a phoneme as a response in a tone discrimination task, where the time to initiate speech rises to almost 500 ms (Ferrand & Blood, 1991, cited in Heldner & Edlund, 2010). While a more complex task may be closer to the complexity of actually producing speech, it is still not clear how similar they really are. The results of Kuriki, Mori, and Hirata (1999) suggest that 500ms may be at the bottom end of the range of times to initiate speech (with 6 participants, they observed a range of 500–800 ms); however, their task also involved producing single words. Therefore, it is still not clear how long it would take to initiate a longer utterance.[2] At a conservative estimate of 500ms, 70–82% of ITIs were too short to be reactions. Furthermore, the findings of Wilson and Zimmerman (1986), who argue that the durations of ITIs in part reflect the cycling of selection options, might suggest that even longer ITIs could still have occurred through projection (as long as some options are passed). For all of these reasons, it is not clear that Heldner and Edlund's results are actually inconsistent with the projection account.

While theories such as Sacks et al.'s and Duncan's are sometimes presented as being in opposition (e.g. Wilson, Wiemann, & Zimmerman, 1984), it is not always clear that this is the case. If, as Sacks et al. suggest, intonation contours are sometimes used to determine whether a word or phrase could be a TCU then it is surely the case that the auditor is *reacting* to the intonation contour. They may react before the TCU is complete, but it is not clear that this is in any way different to what is suggested by Duncan.

It is clear that we are not alone in seeing the distinction between projection and reaction based theories as being something of a false dichotomy. Heldner and Edlund (2010) make the point that these theories of turn-taking need not be mutually exclusive. Several authors have also discussed the use of cues within what are ostensibly projection-based theoretical frameworks (e.g. Clark, 1996; Taboada, 2010; Wilson & Wilson, 2005), for example by suggesting that cues guide auditors' anticipations (as Wilson and Wilson do). We would argue that cues are an important factor in auditors' decisions about whether a turn has or is about to end. Therefore, in the next section we will review in more detail the sorts of cues that auditors may use in turn-taking.

6.2 Cues in turn-taking

In our earlier discussion of Duncan's (1972) study of turn-yielding cues, we saw that several mediums have been proposed for the transmission of cues used for turn-taking. Many of the investigations of cues that have appeared since the work of Duncan have also tended to follow his lead by examining multiple modalities

[2] Indefrey and Levelt's (2004) meta-analysis of single word production studies derived an estimate that it may take approximately 600 ms to begin speaking, although, again, it is not clear how similar the timing of producing a single word is to the timing of beginning a multi-word utterance.

of cues in the same studies. For the sake of clarity, we will focus on different cues individually.

Before reviewing evidence for the use of different types of cues in turn-taking we will first make brief mention of what might seem to be the most obvious cue that a turn has ended, silence. If we were to ask the average person on the street how they know that an interlocutor has finished their turn then we may imagine an initial response that at least implies that the silence following the end of a turn would be a strong cue. However, as Yngve (1970) observes, there are many turn exchanges with gaps that are not perceivable and there are many long pauses that occur without a turn exchange taking place. Furthermore, as Walker and Trimboli (1984) point out, the notion that pauses are cues which allow for seamless turn exchanges would make little sense, as the pause itself would stop the exchange from qualifying as seamless. With a few exceptions (e.g. Local & Kelly, 1986; Wennerstrom & Siegel, 2003), pauses have largely been ignored by those interested in cues that are used in turn-taking.

6.2.1 Visual cues

One of the earliest cues to be suggested as playing a role in smooth turn-taking is the gaze direction of the person in the conversation who is currently taking a turn. Kendon (1967) investigated the role of gaze in conversations between pairs of friends and pairs of strangers. He observed that participants showed a tendency to look away from their interlocutor as they began a turn. This was followed by a tendency to look back at their interlocutor as they ended the turn. The interlocutor then tended to look away, as they began their own turn. Furthermore, he found that when a speaker ended their turn by looking toward their interlocutor, that interlocutor was less likely either to delay in responding or to not respond at all, than they were when the speaker did not end their turn by looking at their interlocutor. Kendon concluded that interlocutors were using speakers' gaze as a sign that they were welcome to begin a new turn.

Attempts to replicate Kendon's findings have met with mixed success. Beattie (1978) examined turn-taking and gaze in four conversations taking place in academic settings, either between colleagues or between supervisors and supervisees (comparison of reported methodologies suggest that three of the four conversations were also investigated in Beattie, 1977, discussed in 2.2.1), and found that, overall, gaze had little effect on the proportion of immediate switches from one speaker to another. Furthermore, he found that, contrary to Kendon, pauses actually tended to be *longer* when the first speaker gazed at the second. Beattie suggests that a possible reason for the difference between his results and those of Kendon were that Kendon may have included turns that ended when a speaker was interrupted (rather than ending when the speaker was finished saying what they wanted) and that these interruptions may have confounded the results observed. In support of this, he highlights that the proportion of immediate switches was almost twice that observed elsewhere (Jaffe & Feldstein, 1970, cited in Beattie, 1978). In our reading of Kendon's (1978) response, it is not clear that he either confirms or denies Beattie's accusation; however, he does state that the duration used as a threshold for immediate switches, 500 ms, was longer than the 300 ms used by Jaffe & Feldstein. Therefore it is to be expected that there were more immediate switches in his study.

Rutter, Stephenson, Ayling, and White (1978) investigated the possible relationship between gaze and turn-taking with two experiments with pairs of participants in conversation. In the first experiment, participants were strangers, who were instructed to discuss their interests; while, in the second experiment, pairs of different participants discussed sociopolitical issues. Rutter et al. found that the majority of turns ended with the speaker gazing at their interlocutor, as Kendon would predict. However, the likelihood of the new speaker looking away at the beginning of their turn was only found to increase in the second experiment, while the opposite was found in the first (although this did not reach the level of statistical significance).

In responding to both Kendon's and Rutter et al.'s studies, Kendon (1978) suggests that Rutter et al.'s second experiment was most similar to his own. The data analysed by Kendon (1967) came from the first and final thirds of the conversations between friends and the penultimate fifth of the conversation between strangers. As a result of these selections, Kendon (1978) argues that his data would not include the "getting to know you" stages that would have been present in Rutter et al.'s first experiment. Kendon continues by asserting that the use of gaze as a cue may be moderated by the conversational context, finally concluding with the suggestion that future research into turn-taking and gaze should take into account possible differences between the situations in which conversation occurs.

We have seen that there is some evidence to suggest that people in conversation use visual cues to help organise turn-taking. It is obvious, however, that visual cues alone would not suffice. Anecdotally, turn-taking generally does not degenerate in non-visual modes of conversation such as on the telephone. This suggests that there may be other non-visual cues which are more important than gaze or gesture. In the remainder of this section we will review evidence that some of the cues used in turn-taking are present in the speech itself.

6.2.2 Acoustic cues

Although Duncan (1972) did not consider the relative importance of the individual sets of cues that he suggested were being used in turn-yielding, the one cue that he identified that has received perhaps the most subsequent attention has been intonation. In his study, he identified any phrase-final intonation that deviates from an "intermediate pitch level, which is sustained, neither rising nor falling" (p. 286) as being a turn-yielding cue. Both Beattie (1981) and Cutler and Pearson (1986) have however raised concerns with the quality of the annotation of intonation performed by Duncan. Beattie has noted that the system used for transcription, the Trager-Smith scheme, is known for exhibiting poor reliability (Lieberman, 1969, cited in Beattie, 1981), while Cutler and Pearson have speculated that Duncan's "subjective impression of what he heard" (Cutler & Pearson, 1986, p. 141) may have been influenced by syntactic and lexical content of the utterances.

Using a more precise system for annotation of intonation, the ToBi system, Gravano (2009; Gravano & Hirschberg, 2011) found support for Duncan's claim that certain intonation contours may be turn-yielding cues by analysing a corpus of 12 dyads playing a series of computer games that involved communication (the Columbia Games Corpus). For each IPU, it was recorded whether or not a turn exchange occurred. He found that both falling and high-rising intonations were associated with a change in speaker (a similar trend has been found elsewhere for high-rise; Wennerstrom & Siegel, 2003), while plateaus, corresponding to Duncan's sustained and intermediate pitch level, were found to be much more likely to occur when a speaker change did not take place, as Duncan would predict.

Gravano's investigation of intonation cues was part of a wider study of turn-yielding cues which tested several other acoustic cues, including reduction of intensity and pitch, and final-syllable lengthening (assumed to be the same as Duncan's "drawl"; Cutler & Pearson, 1986). While Duncan suggested only that a reduction in intensity or pitch was a cue when it occurred for a sociocentric sequences, Gravano considered whether such reduction could be a more general cue of turn-yielding. Measures of both intensity and pitch were taken across entire IPUs, and in the final 1000ms and the final 500ms of each IPU. For pitch, he found a reduction when there was a turn exchange compared to when the same speaker produced the next IPU. While, for intensity, he found not only the same reduction but also that the difference in intensity between when a turn exchange occurred, and when it did not, appeared to be increasing on the approach to the end of the IPU. Both pitch and intensity may therefore provide relatively early cues to a possible turn-ending.

Recall that Duncan claimed that drawl, lengthening of a phrase-final syllable (Cutler & Pearson, 1986), served as a turn-yielding cue. However, phrase-final lengthening has been argued to occur at all prosodic boundaries not just those where a turn is yielded (e.g. Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992). Is there any evidence that lengthening is any greater when a turn is yielded? Gravano calculated rates of speech for entire IPUs, and the final word of IPUs (in both syllables and phonemes per second). In measures of rate, lengthening would be reflected in a slowing down of speech (although the degree of slowing down would be moderated by the length of the IPU; see the point made about turn length and syllable rate in 7.3.1). Contrary to Duncan's claim that lengthening is a turn-yielding cue, Gravano found that when a turn exchange occurred the lengthening appeared to be less pronounced (i.e. rates for each segment were higher for a turn exchange). This reduction of natural lengthening, Gravano suggests, may in fact be a turn-yielding cue in itself.

6.2.3 Which cues are actually used in turn-taking?

We have seen so far that there are a variety of cues which conversational partners could use to anticipate when a turn is about to end. There are, however, two questions that up to this point we have avoided. The first, and certainly most important, is: Do people actually use these cues? It is one thing for TRPs to be accompanied by, for example, a particular intonation contour, it is another thing entirely for people to actually interpret this contour as a sign that a turn will end. The second (contingent, of course, on the answer to the first being "yes"), is: Are some cues better than others?

In general, most studies investigating cues share a similar methodology. Researchers examine conversations for moments where they expect that a turn exchange may occur (usually either TRPs or IPUs), they code the cues present at these points and whether or not an exchange actually occurs, and finally they test whether the likelihood of a change occurring depends on the presence or absence of a particular cue (or cues) of interest. While such analyses are adequate for illuminating the cues that are present at turn exchanges, they give little insight into which cues people are actually using. Better suited to answering questions about the sorts of cues that people actually use are experimental tasks where participants are asked to make relative judgements about (typically, manipulated) recordings of speech. In this section we will exclusively discuss studies making use of such tasks.

One difficulty with assessing the relative value of different cues, such as cues at syntactic and intonation boundaries, is that they tend to co-occur (cf. Caspers, 2003). We know of at least two studies which have teased apart linguistic and acoustic cues, both showing the importance of linguistic cues, but finding mixed results for acoustic cues.

Stephens and Beattie (1986) provide evidence which suggests that linguistic content may be important for determining whether or not a turn has ended. They took recordings of travel enquiries made over the telephone to three different operators and extracted one set of utterances which were turn-final and another set which were turn-medial. Participants in a detection experiment were then instructed to determine whether or not an utterance was turn-final. Half of the participants heard the recordings, with the other half reading transcriptions. While those participants who heard the recordings were able to correctly judge that an utterance was turn-final, participants who read transcriptions were only able to make this judgement above chance for one operator. Further analyses showed that particular topics of utterances (e.g. those related to times and costs of journeys), produced in particular types of syntactic “frames” (e.g. an impersonal sentential form, “The eleven forty-five from Charing Cross gets to Tunbridge Wells at twelve forty-two”, p. 216) were more likely to be turn-final. It was these types of utterances that were being produced more by the operator whose transcriptions could be accurately judged, explaining why participants were able to determine when their utterances were turn-final from a transcription alone. On the basis of these results, they concluded that linguistic cues are important, particularly the interaction between meaning and structure. While they do not give much attention to the importance of acoustic cues when drawing conclusions, the finding that participants were able to correctly determine when an utterance was turn-final for operators who did not produce many of the typically turn-final sentences when they heard recordings would seem to suggest that acoustic cues may be of even more importance than linguistic cues for turn-taking (after all, it did not seem to matter what sorts of utterances were being produced as long as they could be heard).

The conclusion that acoustic cues are more important than linguistic cues during turn-taking contradicts the results of a more recent study comparing these two types of cues. De Ruiter et al. (2006) conducted an experiment where possible effects of linguistic and prosodic cues could be tested independently. Participants performed a task where they would hear recordings of speech, taken from spontaneous conversations, and were instructed to indicate when they anticipated that the current turn would end by pressing a button at the moment they expected the turn to end. The authors believed that by emphasising to participants that they should *anticipate* the turn ending, rather than simply *reacting* to the (possible) turn ending (as was all participants were required to do in Stephens and Beattie's, 1986, study), they would engage in the type of projection which is argued to occur in actual conversation. Some of the recordings that participants heard had been altered in a variety of ways. In one condition participants heard the original recordings, in another two either the prosody and pitch or words had been removed (by flattening the pitch or low-pass filtering, respectively), in the fourth condition both prosody and words were removed (by flattening and filtering), and, in a fifth, prosody, words and rhythm were "removed" (by creating white noise with the duration and frequency spectrum of the original recording).

In all five conditions, participants showed a tendency to anticipate the ending before it had actually occurred, suggesting that they were in fact projecting, rather than simply reacting to the end. De Ruiter et al. assessed the relatively importance of each cue by looking at the consistency of participants' responses. Rather than telling us how accurate they were (although the results for accuracy were the same as the results for consistency for all but the no-pitch-and-no-words condition), this measure tells us how consistently participants were relying on each cue. Participants' responses were equally consistent with the original, unprocessed, recordings and with the recordings where pitch had been flattened. This would suggest that pitch and intonation alone are not used to anticipate turn endings. The use of filtering to remove the words was found to reduce consistency, suggesting that participants were using linguistic information in order to anticipate when a turn was likely to end. While also removing pitch from these recordings did not reduce consistency, participants did perform more consistently in this condition than they did when listening to the white noise. These results suggest that when it comes to anticipating when a turn will end, linguistic information is the most important source of information.

Given Stephens and Beattie's (1986) finding that, overall, participants were better able to determine whether or not an utterance was turn-final in an audio, rather than written, form it may come as a surprise that De Ruiter et al. found that the elimination of pitch information had no detrimental impact on speakers' ability to anticipate turn endings. De Ruiter et al. suggest that, compared to intonation, linguistic information is much more restrictive, and therefore easier to make predictions from. This may be true; although, an alternative explanation for the lack of a significant difference between the unprocessed and the no-pitch conditions may be that flattening of pitch had no effect on other acoustic cues, such as the reductions of intensity and of lengthening observed by Gravano (2009). Alternatively, the difference in results may also be due to the different tasks employed. As we earlier suggested, the participants' task in De Ruiter et al. may be more similar to what actually happens in conversation than the task in Stephens and Beattie, so it is possible that while acoustic cues are valuable when participants are making offline judgements about entire utterances, they have less value when trying to make a judgement about turn endings on the fly. The extent to which acoustic cues can help listeners anticipate turn endings therefore remains an open question, and future research could focus on investigating the effects of different types of acoustic cues on anticipation of turn endings as well as investigating whether or not there are effects of the types of tasks used (e.g. by using De Ruiter et al.'s materials, and having participants make judgements similar to those in Stephens and Beattie).

6.2.4 Discussion

In the previous section we reviewed two theories of turn-taking which both suggest that cues may be involved in achieving the sorts of seamless turn exchanges that have consistently been observed. In this section we have seen that conversational partners may be able to rely on a variety of cues to determine the end of a turn. These include gaze; acoustic cues, such as intonation contours; and, De Ruiter et al. (2006) might suggest most importantly, linguistic cues, such as a speaker reaching a syntactic boundary. While some cues may be better than others, it would seem unlikely that there is one sure-fire turn-yielding cue that is suitable in any form of conversation. Rather, we might expect that, in the course of conversation, people will largely rely on any available cue to determine whether a turn is ending (similar suggestions have been made elsewhere, e.g. Heldner & Edlund, 2010; Wilson & Wilson, 2005).

There is one issue surrounding the use of cues in turn-taking, which seems to have been overlooked in much of the literature: There are two possible explanations of what cues are actually cues to. The first is that they are a cue that a turn *will* end. This would seem to be the view held by Duncan. Cues, he suggests, are produced by speakers when they wish to yield their turn. Therefore, a speaker should only produce cues when a turn will end. The second explanation is that they are phenomena which occur at points at which turn exchanges may be more likely to occur. Or, in the terms of Sacks et al.'s (1974) system, they are cues that appear at the ends of TCUs (and remember that a turn need not end following a TCU). It is possible that different cues may have different explanations. Cues such as syntactic completions are cues of the latter category. It is clear that speakers do not only complete syntactic constituents when they are finished their turn. Other cues, such as the reduction of normal phrase-final lengthening, could plausibly fall into either category. Given that cues appear to play an important role in turn-taking we would suggest that future research should be focused on attempting to explore why cues are produced.

In the next section, we will discuss another theory of turn-taking which attempts to relate the precision of the timing of turn-taking to entrainment of rhythm, which has been found to occur between conversational partners.

6.3 The oscillator theory of turn-taking

In their theory of turn-taking, Sacks et al. (1974) suggest that the very short ITIs (sometimes even as small as 0 ms) observed at turn exchanges result from partners in conversation being able to project the endings of each others' turns. Such projection is argued to be accomplished through rules which specify what a turn can and cannot consist of, as well as cues of different forms which help partners to anticipate that a turn is going to end (e.g. intonation contours). Wilson and Wilson (2005) suggest that what cannot be explained by theories such as Sacks et al.'s, nor by attempts to catalogue turn-yielding cues, is how partners could come to know with great precision (sufficient to regularly produce < 200 ms ITIs) when a turn is going to end. It is only by knowing precisely when a turn will end, they suggest, that conversational partners are able to time the initiation of utterances to produce seamless turn exchanges.

Wilson and Wilson provide a theoretical account of the cognitive processes that may underlie the ability of conversational partners to precisely time turn exchanges. Like much of the work on turn-taking, Wilson and Wilson's account is informed by the system proposed by Sacks et al.; however, they also see a role for cues in turn-taking. Knowledge about what may form a TCU, or a particular intonation contour (amongst other possible cues), may help auditors to determine that a turn will end. However, this provides only a "coarse" prediction. From an intonation contour, an auditor may decide that the current word will end a turn; however, they may still not know precisely *when* the word will end. Wilson and Wilson intend their theory to explain how an auditor could come to refine their predictions about timing.

Wilson and Wilson's theory builds upon two observations about the timing of turn-taking that were discussed in 6.1. Firstly, that there are a large proportion of relatively short ITIs in conversation. Secondly, that there is an observable periodicity to ITIs, thought to reflect the cycling of options for speaker selection (Wilson & Zimmerman, 1986). They argue that this combination of precision and cyclical patterning suggests that endogenous oscillators, internal to each conversational partner, are involved in the timing of turn-taking. Oscillators are thought to perform timing-related functions in the brain, particularly in coordinating activity in distinct cortical regions (e.g. Fries, 2005), and have been shown to be involved in cognitive processes including memory (for reviews, see Jensen, Kaiser, & Lachaux, 2007; Klimesch, 1999), attention (see Schroeder & Lakatos, 2009, and references within), language comprehension (see Bastiaansen & Hagoort, 2006), and consciousness (see Ward, 2003; for general reviews of endogenous oscillators, see Buzsáki & Draguhn, 2004; Ward, 2003).

There is reason to expect that oscillators could be involved in both the production and perception of speech. As Ghitza (2011) points out, there are several similarities in the timings of speech and the timings of oscillators in the brain:

> Phonetic features (duration of 20–50 ms) are associated with gamma (>50 Hz) and beta (15–30 Hz) oscillations, syllables, and words (mean

> duration of 250 ms) with theta (4–8 Hz) oscillations, and sequences of syllables and words embedded within a prosodic phrase (500–2000 ms) with delta oscillations (<3 Hz). (p. 1)

Furthermore, it has been suggested that the specific timings of speech may arise as a result of the oscillatory patterns of parts of the brain responsible for speech perception and production (e.g. Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009). In his *asymmetric sampling in time* (AST) model, Poeppel (2003) proposes that the auditory cortex preferentially samples at rates which correspond to the timing of different aspects of speech. In particular, he argues that the left auditory cortex samples at a high frequency that corresponds to the production of phonemes, while the right auditory cortex samples at a slower rate which may correspond to prosodic phrases. Such asymmetry would account for findings that, while structures implicated in speech perception are bilaterally distributed (e.g. Hickok & Poeppel, 2000), the processing of acoustic transitions in short time-scales, such as within the duration of a single phoneme, tends to be left-lateralised (e.g. Belin et al., 1998; Johnsrude, Zatorre, Milner, & Evans, 1997).

Giraud et al. (2007) tested the AST model in a study that used simultaneous recording of EEG and fMRI. The AST model would predict hemispheric differences in oscillations, with those in the left auditory cortex faster than those in the right. In the AST model, the proposed sampling rates are an intrinsic property of the cortices; therefore, participants did not need to hear any recordings or perform any type of task during recording. Consistent with the AST model, in areas of the brain overlapping with the primary auditory cortex oscillations between 28–40 Hz were observed in the left hemisphere (similar to the frequency of phonetic features, 20–50 Hz, suggested by Ghitza, 2011), while slower oscillations, between 3–6 Hz, were observed in the right hemisphere.

While much attention has been given to the role of oscillators in speech perception (e.g. Giraud & Poeppel, 2012; Ghitza, 2011; Ghitza & Greenberg, 2009), relatively little has been given to their role in speech production. In Wilson and Wilson's (2005) theory, each person in a conversation, whether they are currently speaking or not, has a readiness to initiate production of a syllable which rises and falls in cycles over time. The timing of turn-taking is determined by this oscillatory pattern of the readiness to initiate a syllable for each party in the conversation. The period of these oscillations correspond to the duration of a single syllable, and as such the frequency of the oscillation follows the speaker's syllable rate (the number of syllables they produce per second). At the peak of the oscillation, the speaker is maximally ready to initiate producing a syllable. Their readiness decreases until the mid-point of the syllable, when it then begins to rise again. By the end of the syllable, the speaker has returned to their maximal readiness. They are then ready to potentially begin producing the next syllable.

Even when they are not speaking, parties in the conversation are also going through a periodic cycle of readiness to initiate producing a syllable. The frequency of listeners' cycles are entrained with the cycle of the current speaker through perception of the speaker's speech, although the cycles are in anti-phase (i.e. 180° out of phase). As a result, when the speaker is at the peak of their cycle (most ready to begin a syllable) their listeners are lowest point of their cycles (least ready to begin a syllable). This property of being in anti-phase may explain, at least for the case of dyads, Sacks et al.'s (1974) observation that simultaneous speech is rare. As one person will be most ready to begin speaking when their partner is least ready, then they should be unlikely to begin speaking at the same time. Where there are three or more participants in a conversation, we might expect to see an increase in the likelihood of producing simultaneous starts will increase as there will be at least two parties, the two auditors, who will reach their peaks at the same point.

Wilson and Wilson are not alone in suggesting that the production of syllables has an oscillatory basis (e.g. in theories of how ordering of syllables is achieved in speech; Harris, 2002; Vousden, Brown, & Harley, 2000). In his frame/content (F/C) theory, MacNeilage (1998; MacNeilage & Davis, 2001) argues that the ontology of the syllabic property of speech lies in the cyclic activity of the mandible bone (e.g. during chewing). Noting the close proximity of the area of the brain responsible for ingestion (the frontal perisylvian region) to Broca's area, he plots an evolutionary course where the cycles of ingestion (the closing and opening of the jaw during chewing) were "borrowed" to provide cycles which allowed for the production of syllables in speech (the closing and opening of the mouth and vocal tract to produce consonants and vowels, respectively), via communicative behaviours such as lipsmacks and teeth chatters—also exhibited by other primates—which lack the syllabic property of speech. In the F/C theory, the cycles of closing and opening of the mouth provide a syllabic frame into which phonemes can be inserted.

Giraud et al. (2007) suggest that according to a theory such as MacNeilage's, typical syllable rates observed in speech may reflect the intrinsic rhythm parts of the motor system responsible for movement of the mouth. Consistent with this, they found oscillations of 3–6 Hz (corresponding to syllable rates of 3–6 syll/sec; cf. Tauroza & Allison, 1990, who report a mean range of syllable rates of 3.45–5.45 for conversation in British English) occurring in areas of the motor cortex overlapping with areas responsible for generating mouth movements. When considered together with the AST model discussed earlier, there is evidence that oscillators may be involved in both production and perception of speech. However, to our knowledge, there is no evidence of entrainment between production and perception related oscillations (although, as will be reviewed in 7.3.4, there is considerable evidence of overlaps in the areas of the brain responsible for production and for comprehension/perception).

What does the oscillator theory mean for the timing of turn-taking? Recall that the period of oscillations corresponds to the duration of an utterance (with each syllable lasting from peak to peak). When the speaker reaches the end of their turn they will be at their maximal readiness to initiate a syllable. However, as a consequence of being entrained in anti-phase, their partner will be at their minimal readiness. If their partner has anticipated that the speaker may yield their turn around this point—perhaps on the basis of one of the cues discussed in the previous section—then they will be maximally ready to initiate the production of the first syllable of a new turn either half a syllable before or after the end of the speaker's turn. If the partner does not begin speaking half a syllable after the end of the turn, perhaps because they are still planning what they will say, then they will not be able to begin speaking again until another full cycle of their readiness has been completed (i.e. after one full duration of a syllable). Both parties will remain entrained with each other for several cycles, although at some point entrainment will break down. Wilson and Wilson do not specify how long it takes for entrainment between speakers to break down, although, on the basis of Jefferson's (1989) observation that simultaneous starts are rare in silences of 1 sec and their own claim that it is entrainment that prevents simultaneous starts, they assume that partners remain entrained for at least 1 second.

Wilson and Wilson's idea that the rhythm of speech may be used as one source of information to anticipate the timing of a turn ending is not a new one. Walker and Trimboli (1984) have previously suggested that rhythm and intonation may be used together in order to achieve smooth turn-taking. In their briefly-sketched account, rhythm allows listeners to anticipate the position of TRPs, while intonation contours provide information about whether or not the TRP is actually the end of the turn. Consistent with this account, and the oscillatory theory, De Ruiter et al. (2006) found evidence to suggest that people may use rhythm to anticipate when a turn will end. However, their findings that the removal of prosodic information had little effect on people's ability to anticipate the end of a turn would be inconsistent with Walker and Trimboli's account. De Ruiter et al.'s findings are less problematic for Wilson and Wilson, as the oscillator theory does not rely on intonation as the only possible turn-yielding cue. Rather, Wilson and Wilson suggest that people are likely to be opportunistic in the types of cues they exploit during turn-taking.

A crucial requirement of the oscillator theory is evidence that speakers in conversation do actually become entrained on the rates at which they speak. If such entrainment did not occur, then the theory would obviously be untenable. In the following section we will present evidence suggesting that entrainment occurs extensively during conversation, as well as discussing one prominent account for why this occurs that does not consider oscillators.

6.3.1 Entrainment in conversation

In the course of conversation, people demonstrate a tendency to become increasingly similar across a broad range of dimensions. In what is said, partners may come to repeat the referring expressions previously used by their partner (e.g. Clark & Wilkes-Gibbs, 1986; Finlayson & Corley, 2012), to reuse the syntactic structures that they have recently heard (e.g. Branigan et al., 2000; Cleland & Pickering, 2003; Levelt & Kelter, 1982), and, in at least one case, to even become entrained on the median frequencies of the words that they use (Levin & Lin, 1988). However, the similarities between conversational partners are not limited to *what* is said. People in conversation may also come to become more similar in *how* they speak, for example the rate at which speech is produced.

There has been a long history of studies demonstrating entrainment of rate of speech during conversation (see Street, 1982, and references within). Webb (1969) provide one demonstration that the rate at which a person speaks in a dialogue is related to the rate at which their partner is speaking. He compared mean speech rates of interviewers and interviewees, and found a correlation between speech rates within conversations, such that faster-speaking interviewees were associated with faster-speaking interviewers. A similar relationship between rates of speech has also been observed in the MTC (Finlayson et al., 2010). Evidence of entrainment has also come from more fine-grained analyses, such as those of Street (1984). Again using interviews as a source of data, Street divided dialogues into one-minute slices and calculated the speech rate of each participant in each slice. Across slices, there was a positive relationship between the rates at which each participant spoke, demonstrating that entrainment occurs across the length of a conversation.

The entrainment of rates of speech reflect a more general trend for conversational partners to become more similar in the ways in which they speak. Pardo and colleagues (Pardo, 2006; Pardo, Jay, & Krauss, 2010; Pardo, Gibbons, Suppes, & Krauss, 2012; Pardo, Jay, et al., 2013) have shown across a series of studies that conversational partners show a tendency to sound similar (similar results have been observed in studies using shadowing, e.g. Miller, Sanchez, & Rosenblum, 2010; Shockley, Sabadini, & Fowler, 2004). In these studies, the extent of entrainment is generally assessed using the AXB task (Goldinger, 1998), where naïve participants rate the similarity of tokens produced by one speaker before and during conversations with a token produced by the other speaker in the conversation. One consequence of this methodology is that it is often not clear which particular aspects of their linguistic performance participants were becoming entrained on (although, note that a recent study suggests that it is unlikely that any one aspect is driving participants' perceptions of similarity; Pardo, Jordan, Mallari, Scanlon, & Lewandowski, 2013). Those studies with a narrower focus have observed entrainment occurring across a variety of aspects of linguistic performance, including intensity (Coulston, Oviatt, & Darves, 2002; Levitan & Hirschberg, 2011; Ward & Litman, 2007), pitch (Gregory, 1990; Heldner, Edlund, & Hirschberg, 2010; Levitan & Hirschberg, 2011), response latencies (Cappella & Planalp, 1981), utterance durations (Matarazzo, Weitman, Saslow, & Wiens, 1963) and accent (e.g. Giles, 1973; Gregory & Webster, 1996).

Entrainment during conversation is not just restricted to what is said, and how it is said. McFarland (2001) has shown that partners in spontaneous conversation show similar patterns of breathing, particularly around turn exchanges and during overlapping non-verbal acts such as laughter. Like syllable rate, breathing is also a cyclical behaviour (although across a larger time-scale) and Wilson and Wilson (2005) suggest that entrainment of breathing may come to strengthen the entrainment of syllable rate, as slower oscillators have been found to influence oscillators at higher frequencies (Buzsáki & Draguhn, 2004).

Finally, Shockley, Santana, and Fowler (2003) had pairs of participants perform a communicative task (a "spot the difference" task where each participant could only see one of the pictures), either with each other or with pairs of co-present confederates. They found that there was more similarity in the postural movements between the two participants when they were performing the task together than when they were performing it with their respective confederates. In a later study, Shockley, Baker, Richardson, and Fowler (2007) investigated the effect of articulation on postural similarity, in particular whether similarity at the postural level arose from similarity in rate of speech. Pairs of participants performed a task which involved reading aloud (either simultaneously, or alternating) words presented on a screen. Participants were not able to see each other, and words were presented either individually at a fast rate or a slow rate, or all at once (allowing speakers to choose their own rate). The authors expected that where participants were free to speak at their own tempos they would become entrained. Participants were found to exhibit more similar postures in the fast condition than in either the slow or natural conditions. The authors concluded that the increased postural entrainment in the fast condition was because participants produced a greater number of words, and so were more heavily influenced by articulation. They also concluded that the entrainment of posture does not arise from the entrainment of rate of speech; however, as they did not present any evidence to show that participants' rates of speech were actually entrained in the natural condition (nor do they present evidence to show that participants were *not* entrained in the fast and slow conditions) we would suggest that this conclusion is speculative at best.

In the oscillator theory, the entrainment of rates of speech occurs as a consequence of the entrainment of endogenous oscillator. As Wilson and Wilson (2005) put it, their theory provides a mechanism where entrainment of rate occurs "for free". Previous theoretical accounts of why conversational partners come to be more similar have also tended to focus on the consequences of the similarity; however, they have suggested that the consequences may be more wider-reaching. Increased similarity in linguistic content has an obvious effect on communicative success: It should be easier to express ideas if partners share mutually comprehensible ways of talking about facets of those ideas (cf. Clark, 1996; Garrod & Pickering, 2004; Pickering & Garrod, 2004). Entrainment of rate of speech could have similar effects. For example, speech that is more similar in style to your own may be more comprehensible (cf. Giles & Powesland, 1975). Alternative theories of entrainment, most prominently Communication Accommodation Theory (CAT; Coupland & Giles, 1988; Giles, Coupland, & Coupland, 1991; Giles & Powesland, 1975; Giles & Smith, 1979), have generally focused on further-reaching social consequences of the similarity between conversational partners. CAT began as an attempt to marry social psychology with psycholinguistics, with an aim of developing an understanding of the diversity of speech in social settings (Giles et al., 1991). While it has since vastly expanded in both its theoretical content and the range of phenomena and settings considered (for a review, see Giles et al., 1991), in its early history CAT relied heavily on *similarity attraction*, the idea that people tend to prefer others that they see as being similar to themselves (e.g. Byrne, Griffitt, & Stefaniak, 1967). In CAT, entrainment (or convergence, as it is termed in the literature of the theory) is a means by which people are able to strengthen social relations with others. By speaking similarly to a conversational partner, for example becoming entrained on their rate of speech, it is argued that the speaker may cause their partner to form a positive impression of them. Consistent with this suggestion, studies by Street and colleagues (Street, 1984; Street, Brady, & Putman, 1983) have shown that, in interview settings, people were found to have more positive impressions of competence and social attractiveness for conversational partners who spoke more similarly to themselves. Similarly, Chartrand and Bargh (1999) found that participants who had their posture and mannerisms mimicked by a confederate during a conversation liked the confederate more than participants who were not mimicked. They argue that people naturally show a tendency to mimic those with whom they are interacting, what they have termed *the chameleon effect*, which may function to enhance group cohesion.

While Chartrand and Bargh suggest that the chameleon effect may be largely unconscious, in CAT, entrainment may have a more conscious aspect. People who are striving to appear more likeable may actively become entrained with interlocutors. Putman and Street (1984) recruited participants to take part in interviews who were instructed that they should attempt to give off a particular impression while being interviewed. Half of the participants were told that they should appear likeable, while the other half were told they should appear "not likeable". Participants who had been instructed to be likeable were found to become entrained on speech rate (as well as on turn duration) with their interviewer, while those told not to be likeable produced the opposite behaviour, becoming dissimilar to their interviewer (known in CAT as divergence). Similarly, Natale (1975a, 1975b) showed that speakers who rate highly on their need for social approval entrained more closely to their partners vocal intensity and pause durations than those who rated lower. Taken together, these studies suggest that when a speaker wants to give a positive impression they become more similar to their partner, while evidence presented earlier suggests that these strategies appear to be successful.

While CAT may explain *why* people come to be entrained, what is lacking is an account of *how* they come to be entrained. In particular, an explanation of the cognitive mechanisms that underlie entrainment. In the oscillator theory, entrainment is a lower-level mechanistic process which results from a natural propensity of oscillators; however, CAT talks only of the higher-level motivations for entrainment. It is not clear that the tendency for oscillators to become entrained can be modulated on the basis of motivation, nor that motivations can stop entrainment altogether and cause speakers to become dissimilar, as divergence would seem to require.

There is evidence that entrainment occurs when there is no apparent motivation to make a positive impression. Jungers and Hupp (2009) were interested in whether rate of speech could be primed in a similar fashion to syntactic structure. Participants heard recordings of prime sentences which were produced at either slow or fast rates. After each prime, they then described a target image. It was found that participants produced a faster spoken description following a fast prime than during a slow prime, suggesting that rate of speech was being primed. The presentation of Jungers and Hupp's methodology gives us no reason to believe that participants thought that the primes were anything other than recordings (rather than someone actually speaking directly to them in real time), while the recorded speaker was not the same person as the experimenter (M. K. Jungers, personal communication, 12th August 2013) so participants were unlikely to believe that becoming entrained with the recordings would lead to the experimenter forming a positive impression (as they might have if the speaker and the experimenter was one and the same person). If participants knew that the speaker producing the primes was not able to hear them, then it is not clear why they would be becoming entrained to their rate of speech, at least not if the goal of entrainment is to cause the person to whom you are becoming entrained to develop a positive impression of you. The rate priming effect observed in this study may therefore be difficult to explain within the framework of CAT; however, it would be entirely consistent with the oscillator theory.

Before moving on to the following section, where we will more extensively evaluate the oscillator theory, we will first briefly pause to mention that there is little existing support for an oscillator based account of rate of speech entrainment. To our knowledge there is no evidence showing entrainment of oscillators between speakers during conversation,[3] nor do we know of any evidence to show that within a single brain there is entrainment between oscillators in areas involved in production and those involved in perception (although, as will be reviewed in 7.3.4, there is considerable evidence of overlaps in the areas of the brain responsible for production and for comprehension/perception).

6.3.2 Evaluating the oscillator theory

Despite making several strong predictions about the timing of turn-taking, there have been very few studies which have tested the oscillator theory in the almost ten years since it was presented by Wilson and Wilson (2005). In the remainder of this chapter we will review the small number of studies which have tested predictions of the theory, and one previously unmentioned study which appears consistent with a prediction of the theory.

Using the Columbia Games Corpus, Beňuš (2009) tested four different sets of predictions which follow from the oscillator theory. As will be seen, when taken together, his results do not provide strong support for the oscillator theory. Beňuš's study provides what is, as yet, the most comprehensive test of the oscillator theory; therefore, we will give considerable attention to each of the predictions and his subsequent findings.

Firstly, if the production of syllables follows a periodic rhythm then adjacent IPUs produced by the same speaker (i.e. adjacent IPUs where a turn exchange does not occur) should have syllable rates that correlate, as the rhythm in the first IPU should carry into the second IPU. While, as the readiness to initiate speech continues to cycle even when a person is not speaking, the pauses that separate IPUs should also correlate with the person's syllable rate. These pauses should also be in phase with the person's syllable rate, as the pause must begin at the

[3] Dumas, Nadel, Soussignan, Martinerie, and Garnero (2010) report entrainment of alpha band oscillators in the centroparietal region within dyads performing a motor imitation task. Oscillations in this band (as well as in the mu band), located in this region, are thought to index social coordination (Tognoli, Lagarde, DeGuzman, & Kelso, 2007).

peak of a cycle and end at a later peak (Beňuš operationalises phase by dividing the duration of pauses by the syllable rate; therefore, in phase pauses should produce values around 1, 2, 3, etc., while, in anti-phase, pauses should produce values around 0.5, 1.5, 2.5, etc.). When comparing within speakers, Beňuš found a correlation between the syllable rates of adjacent IPUs, and between syllable rates and pause durations, consistent with some of these predictions; however, he found no evidence that IPUs were in phase.

Secondly, if speakers become entrained then their syllable rates on either side of a turn exchange should correlate. The pauses between IPUs at a turn exchange should also correlate with the second speaker's syllable rate, as the time it takes them to begin speaking in part reflects the time it takes them to reach the peak of their cycle, although the pauses should be in anti-phase with the syllable rate. When Beňuš compared IPUs produced by different speakers, the correlations between syllable rates did not reach significance, nor were the rates correlated with pause durations, suggesting that partners were not becoming entrained. The absence of evidence of entrainment is surprising, given the evidence reviewed in the previous section. We are unable to explain the absence of this effect, and Beňuš himself does not offer an explanation.

Thirdly, the pauses between IPUs should be bimodally distributed around zero. This is because, while an auditor in anti-phase cannot begin speaking as soon as the speaker has finished, they should be equally likely to begin speaking at half of period before or half a period after the end of the speaker's turn (this prediction is explicitly made by Wilson and Wilson, although it is not clear why they do not expect people to show a bias for speaking after a turn has ended, rather than before). Examination of the distribution of ITIs offered only mixed support for the oscillator theory. There was no indication that ITIs were bimodally distributed. However, ITIs were found to follow a unimodal distribution which peaked around 100-200ms, which is generally consistent with the findings presented at the beginning of this chapter.

Finally, the number of simultaneous starts should begin to rise after a pause of 1 second, as, according to Wilson and Wilson (2005), this may be the earliest point at which entrainment could begin to break down, and the breakdown of entrainment (and consequently of being in anti-phase) is suggested to be when simultaneous starts become more likely. Examination of ITIs for simultaneous speech revealed that they reached actually reached a peak at around 500ms, suggesting that, if the oscillator theory is correct, then Wilson and Wilson may have overestimated how long conversational partners remain entrained.

While Beňuš's study generally offered only mixed support for the predictions of the oscillator theory, some evidence that is consistent with the oscillator theory has come from examining the relationship between the ITIs produced by each partner in a conversation (Ten Bosch, Oostdijk, & Boves, 2005). A consequence of Wilson and Wilson's theory is that we should see a degree of entrainment in the durations of gaps produced by speakers in conversation. Such entrainment should come as a natural consequence of entrainment of rate of speech, as the duration of gaps will reflect the rates of speech of the partners involved in the turn exchange. Consistent with this prediction, Ten Bosch et al. found that there was a correlation between the durations of the gaps produced by speakers in a corpus of 93 telephone conversations. We must, however, be cautious when drawing conclusions from this study. It is possible that the trend that they observed was confounded by the position in the conversation of each gap: If there was a consistent trend in their data for gaps to decrease (or increase) in duration across the conversation then this would produce a similar correlation between consecutive turn-intervals (e.g., the gap between turn t and turn $t-1$ would be smaller than the gap between turn $t-1$ and $t-2$, which would be smaller than the gap between turn $t-2$ and turn $t-3$, and so on).

6.4 Conclusion

Despite the "anarchy" of conversation, turn-taking appears to proceed with a seamless organisation. In this chapter we have reviewed accounts of how this may be achieved. We first introduced two theories of turn-taking, Sacks et al.'s (1974) projection-based theory, where a series of rules about when people can take turns and what turns must consist of allow people to anticipate turn-endings. This could be viewed as being in contrast to Duncan's (1972) reaction-based theory of turn-taking, where speakers respond to cues that indicate when a turn is ending. The dichotomy between projection and reaction is, we would suggest, a false one. A variety of different cues, from gaze, to intonation, to syntactic completion, are likely to be used to help people anticipate when a turn is about to end.

The last section of this chapter introduced a more recent theory of turn-taking, where the entrainment of oscillators representing people's readiness to initiate the production of a syllable allows conversational partners to precisely time the beginning of their turns to coincide with the ends of others' turns. Such an account is not in opposition to those of Sacks et al. or Duncan. Rather, it expands on their proposals about projection and the role of cues by providing a mechanism by which auditors can come not only to anticipate that a turn will end, but also when it will end.

In the next chapter we will test several predictions of the oscillator theory using the MTC. While, as we have seen, the oscillator theory has so far met with mixed empirical support, we know of no study which has tested what would seem to us to be its most fundamental claim: That partners who are more entrained will produce more precisely timed turn exchanges. It is this prediction, among others, which we will test in the following chapter.

Chapter 7

Corpus analysis 2: Testing the oscillator theory of turn-taking

In the previous chapter we introduced Wilson and Wilson's (2005) oscillator theory of turn-taking. In this theory, it is argued that there are endogenous oscillators in the brains of conversational partners which represent each partner's readiness to initiate production of a syllable. The periods of these oscillations are the durations of a single syllable, with their frequency therefore representing the person's syllable rate. During conversation, these oscillators become entrained, which allows partners to precisely time their own turns to begin close to the end of each other's turns.

In this chapter we will test three predictions derived from the oscillator theory against data from a corpus of task-orientated dialogue, the MTC. In particular, our analyses will focus on the relationship between the entrainment of rate of speech between conversational partners and the precision that is exhibited in their turn-taking. While there have already been several studies which have tested aspects of the theory (Beňuš, 2009; Włodarczak, Juraj, & Wagner, 2012), ours is the first to directly test the relationship between entrainment of rate of speech and the precision of turn-taking.

The first of our analyses will test the prediction that the rates of speech of conversational partners become entrained. Previous research leads us to expect that participants in the MTC would become entrained on rate of speech with their partners. Finlayson et al. (2010) measured the articulation rate per conversation for each participant in the MTC and found a relationship between each speakers' articulation rates, such that faster speaking participants tended to have faster speaking partners. However, Finlayson et al.'s analyses compared global measures of articulation rate (i.e. comparing each speaker's mean rate across a conversation), while entrainment is described as a local process by Wilson and Wilson (i.e. a relationship that holds between turns, not just between pairs of participants). Similarity between partners' overall rates in each conversation need not imply a similarity between the turns (e.g. each pairing could start at the same rate and become increasingly dissimilar across the length of the conversation, and as such they would not be entrained; however, because each pairing started in the same place, when averaged across the conversation there may be a trend for each pairing to share a similar rate). Therefore, we will test for the presence of entrainment by comparing the syllable rates between subsequent turns produced within a conversation (similar to the approach used by Beňuš).

The second of our analyses will test the prediction that there is a relationship between the rate at which a turn is spoken and the duration of the ITI that preceded it. In the oscillator theory, the period of oscillations corresponds to the duration of single syllables. Consequently, faster speech will be associated with shorter periods. When periods are short the next peak will be reached sooner than when periods are longer. If we assume that people tend to begin speaking on the first peak following the end of their partner's turn, as would appear to be the case given the high frequency of < 200 ms ITIs, then in general faster speakers, with shorter periods, should begin speaking sooner than slower speakers, with longer periods. Therefore we would expect to see that the ITIs that occur before faster spoken turns will on average be shorter than those that occur before slower spoken turns.

As a direct test of the oscillator theory, our final analysis will test the crucial prediction that precise turn-taking is achieved through the entrainment of oscillators. If this is indeed the case then we would expect that conversational partners who are more tightly entrained will tend to begin new turns closer to the ends of previous turns, as they will be able to more precisely time the start of their turn relative to the end of each other's previous turn.

7.1 Methods

The corpus analyses presented in this chapter are based on the MTC dataset prepared following the steps described in Chapter 3. As our outcomes were all concerned with features of turns, or of the relationships between consecutive turns, we collapsed across individual tokens, taking each turn as our unit of analysis. This reduced our data to 20,974 observations. As some of our predictors

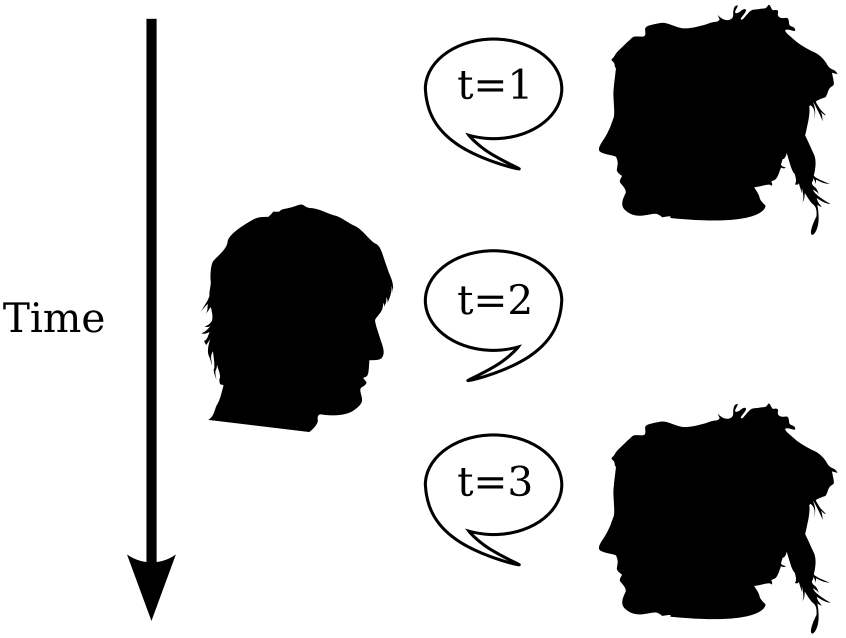


Figure 7.1: An illustration of turn taking between two conversational partners.

described features of previous turns (for example, syllable rate in the previous turn), our analyses were restricted to only those turns where such information was available. This necessitated the removal of the first turn of each conversation (as there was no information on the prior turn from which to generate predictors), leaving a remaining 20,846 observations.

In order to illustrate the outcomes and predictors that are investigated in this chapter, Figure 7.1 shows a sequence of three turns (t). Each turn in a conversation serves both as an observation of an outcome and as a predictor for the next turn (except for the first turn, which serves only as a predictor, and the final turn, which serves as only an observation). As an example, in Corpus Analysis 2a the syllable rate in the first turn is used as a predictor of the syllable rate of the second turn, which is then used as a predictor of the syllable rate of the third turn, and so on.

7.1.1 Outcomes and parameters

The analyses of the MTC presented in this chapter modelled continuous outcomes, therefore linear mixed effects regression was used for each of three analyses. Analysis 2a investigated factors influencing syllable rate in the current turn (SR_t). There are several possible measures of rate of speech (including words or phonemes per second) however we used syllable rate (excluding silent pauses) as it maps directly onto the oscillators in Wilson and Wilson's (2005) theory (e.g. 5 syll/sec would equate to a frequency of 5 Hz). The steps taken to calculate syllable rate were described in 3.1.

Analysis 2b investigated factors influencing the durations of ITIs. As described in Chapter 3, turns were annotated by the Spoken Dialogue Parser (McKelvie, 1998). Each turn is produced by one speaker, and ITIs were measured as the start time of the current turn minus the end time of the previous turn ($\text{start}_t - \text{end}_{t-1}$). As participants in the MTC could interrupt one another, producing overlapping speech, turn-intervals could have negative values.

Finally, Analysis 2c investigated factors which influence the precision of turn-taking. We operationalised precision as being how close the duration of an ITI was to having a value of zero (a no-gap-no-overlap turn exchange, in Sacks et al.'s, 1974, terminology). As some ITIs would have a negative duration (in the case of overlaps, where the second speaker began speaking before the first had finished), we used the absolute value of inter-turn intervals as a measure of precision ($|\text{start}_t - \text{end}_{t-1}|$). Therefore, a turn that began 100 ms after the end of a previous turn would be considered to be as precise as one that began 100 ms before the end of the turn (an ITI of -100). As turn-exchanges became more precise the values of our measure of precision tend towards zero (although, if entrainment occurs in anti-phase then ITIs would not be expected to reach zero).

Random effects

As the analyses presented in this chapter are largely concerned with the dynamics of speech between conversational partners, we tested a random effect for each dyad (*conversation*) in addition to the three random effects tested in the analyses of the MTC presented in Chapter 5: *speaker*, *partner*, and *map*.

Fixed effects

While the focus of our analyses was on the entrainment of syllable rate, and its relation to the timing of turn-taking, there is reason to consider the effects of other features of the design of the MTC on rate of speech and turn-intervals. We would anticipate that our data, coming from a corpus of relatively unstructured speech, could be noisier than that arising from a tightly-controlled experiment. By statistically controlling for as many factors as possible, particularly those where we may have reason to expect relationships with our outcomes to exist, we reduced the possibility of relevant effects being obscured by the noise.

A list of the predictors considered in each of our analyses is shown in Table 7.1. As detailed in Chapter 3, all predictors were centered, while continuous predictors were subsequently standardised. Similar to the analyses of the MTC presented in the previous chapter, we tested a set of control predictors which were intended to account for potential confounds in the data which we were not interested in. The first of these was the word count for the current turn (Length_t). This was intended to account for relationships between turn length and speech rate (e.g. Yuan, Liberman, & Cieri, 2006), and initiation times (Ferreira, 1991), observed in previous studies.

A further two control predictors represented the speakers' experience with both the task and the map. While, in the previous chapter, we found no clear evidence for a link between experience and speakers' production of repairs or hesitations, we may still expect increased experience to reduce the difficulty of performing the map task, and we know of at least one study suggesting that difficulty leads people to speak slower (Lay & Paivio, 1969). Increased experience may reduce the amount of planning that is necessary and this may reduce the length of time it takes to plan a turn.

Finally, we included a predictor representing the progress through the conversation (t). This control served two purposes. Firstly, including progress provides a third measure of experience. Secondly, if a positive relationship was found between the syllable rates of both partners in Analysis 2a then an alternative explanation could be that both partners were speeding up (or slowing down) across the length of the conversation rather than actually becoming more similar (as both would be getting faster an increase for one person would be followed by an increase for the other person). By including a control for their progress, we would be able to take into account any general trend for speeding up or slowing down.

Table 7.1 presents all of the predictors tested in our analyses. For each outcome variable, we tested a model with six standard predictors-of-interest, representing different aspects of the MTC. These were predictors which previous studies of corpora, presented in Table 7.2, had given us reason to believe may influence rate of speech and ITIs. These were: speaker's role in the task (giver vs. follower of

Table 7.1: Corpus analysis 2: Fixed effects tested in each analysis. Predictors-of-interest are shown in bold. The final three predictors were those related to our testing of the oscillator theory, and they were only tested in the stated analyses.

Predictor	Type	Range
Length$_t$ (# of words)	Continuous	1–133
Experience with task	Continuous	1–4
Experience with map	Discrete	First/Second time
Current turn (t)	Continuous	2–478
Role	Discrete	Giver/Follower
Visibility	Discrete	Visible/Not visible
Familiarity	Discrete	Friends/Strangers
Gender	Discrete	Male/Female
Partner's gender	Discrete	Male/Female
Gender match	Discrete	Matching/ Mismatching gender
$\mathbf{SR}_{t-1}$ (in syll/sec) [Corpus Analysis 2a]	Continuous	0.50–20.04
$\mathbf{SR}_{t}$ (in syll/sec) [Corpus Analyses 2b]	Continuous	0.50–20.04
$\lvert\mathbf{SR}_t - \mathbf{SR}_{t-1}\rvert$ [Corpus Analyses 2c]	Continuous	0–16.82

instructions); partners' ability to see each other (visible vs. not visible); prior familiarity between the speaker and their partner (friends vs. strangers); and three gender-related predictors, speaker's gender, their partner's gender, and whether or not their genders matched.

Each model that we tested also included a unique predictor-of-interest related to our tests of the oscillator theory: syllable rate in the previous turn (SR$_{t-1}$), syllable rate in the current turn (SR$_t$), and the difference between SR in current turn and SR in previous turn ($|\text{SR}_t - \text{SR}_{t-1}|$).

The outcome variables in two of the analyses presented in this chapter were not expected to be normally distributed. In both cases, this was because they were bounded at zero. In the case of Corpus Analysis 2b, this was because we eliminated all overlaps (for reasons given below) and in the case of Corpus Analysis 2c, this was because we took an absolute value of ITIs and therefore they were all non-negative. We this in mind, for all of our analyses we first performed a Cramér-von Mises test on the outcome variable, and where outcomes were found not to be normally distributed we performed a Box-Cox transformation, as described in Chapter 3.

In keeping with the model construction process detailed in Chapter 3, for each DV we first constructed a model with only random effects, before incrementally

Table 7.2: Corpus Analysis 2: Previous studies suggesting relationships between the exploratory predictors-of interest and rate of speech and turn-taking behaviour

Factor	Finding	Source
Role	Givers produce shorter gaps[a]	Bull & Aylett, 1998
Visibility	Longer gaps when partners able to see each other	Bull & Aylett, 1998
	More overlaps when partners able to see each other	Ten Bosch et al., 2005
Familiarity	People speak slower to friends than to strangers	Yuan et al., 2006
	Friends produce more overlaps	Yuan et al., 2007
Gender	Males speak faster than females	Binnenpoorte et al., 2005; Verhoeven et al., 2004 Whiteside, 1996 Yuan et al., 2006
	Females overlap more than males[b]	Yuan et al., 2007
Partner's gender /Gender match	Females are overlapped more than males	Yuan et al., 2007

[a] Recall that gaps are turn exchanges with ITI ≥ 0

[b] However, see Anderson and Leaper (1998), and the papers cited within, which suggests that the relationship between gender and overlapping is likely to be heavily influenced by context.

testing the fixed effects (in the order of control predictors and then predictors-of-interest).

7.2 Results

As with the analyses presented in the previous chapter, we report on only those fixed effects which significantly improved the fit of the model. The only exception to this is are the three unique predictors-of-interest related to our tests of the oscillator theory.

It has been suggested by Auer, Couper-Kuhlen, and Müller (1999) that in order to perceive isochrony (the rhythm of speech) listeners must hear at least three metrical units (e.g. syllables, in the case of the oscillator theory) of speech. To ensure that for each observation participants had heard enough speech to perceive its isochrony, we eliminated all observations where the *preceding* turn did not contain at least three syllables (a similar step was taken by Beňuš, 2009). This left a remaining 14,640 observations.

Table 7.3: Corpus Analysis 2a: Final linear mixed-effect model of syllable rate. Predictors-of-interest are shown in bold. The reported correlation is between the by-speaker random slope and the by-speaker random intercept; where no correlation is reported it is because the inclusion of a correlation did not significantly improve model fit.

					Random effects			
Fixed effect	β	SE	t	$p(\beta = 0)$	Group	Predictor	Variance	Correlation
Intercept	3.104	0.033	92.713	$< .001$	Conversation	*Intercept*	0.021	-
Length$_t$	0.208	0.008	25.256	$< .001$		SR$_{t-1}$	0.003	-
t	0.032	0.009	3.351	$< .001$	Speaker	*Intercept*	0.059	-
Task experience	0.034	0.016	2.153	$< .05$		SR$_{t-1}$	< 0.001	-
SR$_{t-1}$	0.035	0.010	3.380	$< .001$		Male partner	0.034	-
						Giver	0.112	-0.392
					Residual		0.857	

7.2.1 Corpus analysis 2a: Entrainment of rate of speech

In the oscillator theory, endogenous oscillators, representing people's readiness to initiate a syllable, become entrained doing conversation. Such entrainment would mean that the rates at conversational partners speak should be similar, in particular, faster speech in one turn should lead to faster speech in the subsequent turn.

Syllable rate in the MTC was found not to be normally distributed ($W = 1.153$, $p < .001$). Before performing the model construction process a Box-Cox transformation was applied to syllable rate ($\lambda_1 = 0.694$).

See Table 7.3 for the full model of speakers' articulation rate in the MTC. The first prediction that we tested was that conversational partners should become entrained in their rates of speech. As a result, the faster one person spoke in one turn, the faster their partner would speak in the next turn. In line with this prediction, we found a positive relationship between syllable rate in the previous turn (SR$_{t-1}$) and syllable rate in the current turn ($p < .001$). This effect is shown in Figure 7.2. In order to rule out the possibility that this effect was only due to both speakers either speeding up or slowing down across the length of the conversation, we tested a measure of partners' progress in the conversation. The rate at which partners spoke was found to increase across the length of each conversation (t; $p < .001$). As this was tested in the model before our predictors-of-interest, we can be sure that speakers were actually becoming entrained (rather than both shifting independently in the same direction). Longer turns were found

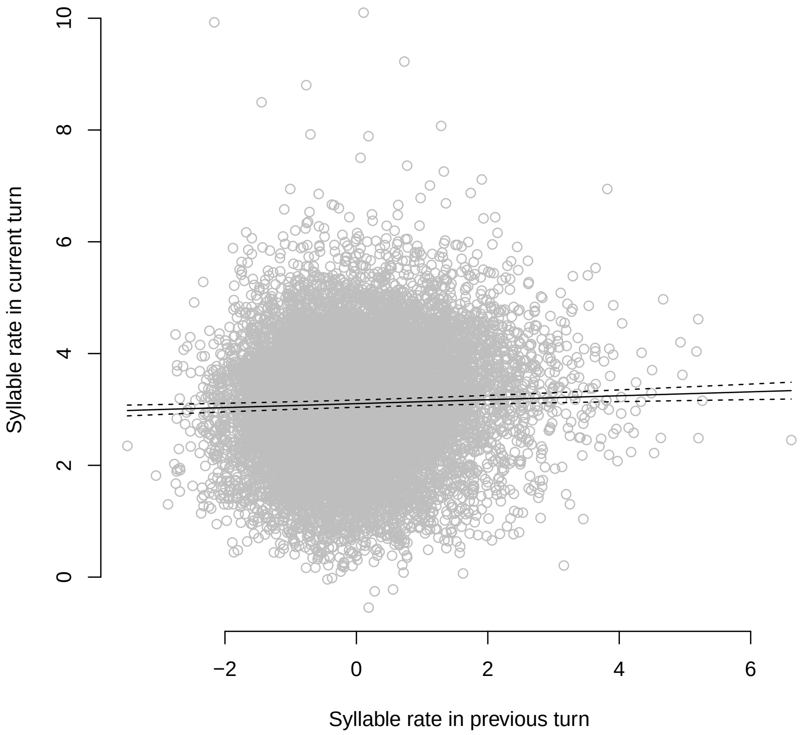


Figure 7.2: Corpus Analysis 2a: Syllable rate in the current turn by syllable rate in the previous turn. The line represents the estimated effect, with confidence intervals. Syllable rates in the current turn have been Box-Cox transformed, while syllable rates in the previous turn were standardised prior to the model construction process.

to be produced at a faster rate than shorter turns (Length$_t$; $p < .001$). Finally, participants were found to speak faster each time they performed the map task ($p < .05$). No other predictors significantly improved model fit.

While several of the predictors that we tested were found to influence rate of speech, it was striking that other predictors which we expected to influence the rate at which participants spoke were not observed to have an effect (e.g. speaker's role or their gender). One possible explanation for this could be that the measures of syllable rate obtained from shorter utterances may not accurately represent the rate at which a person generally speaks in a given situation (as many factors could influence the durations of single words; e.g. Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Fowler & Housum, 1987; Wightman et al., 1992). While this would not be a problem for our testing of the oscillator theory, as the periodicity of even short utterances would be expected to bear relation to the periodicity of the utterance that came before, it may introduce noise which obscures the effects of factors we may consider to have more global effect on the rate of speech (e.g. the difficulty experienced by a speaker, or their gender).

Goldman-Eisler (1954) has suggested that any measure of rate of speech obtained from a stretch of fewer than 5 syllables is likely to be an inaccurate representation of a speaker's typical rate of speech. In order to allow for this possibility we reran Analysis 2a with the exclusion of any turn which contained fewer than five syllables. This reduced the number of observations to 7,477. The results of this second analysis are shown in Table 7.4.

Table 7.4: Corpus Analysis 2a: Final linear mixed-effect model of "accurate" syllable rate. Predictors-of-interest are shown in bold.

					Random effects			
Fixed effect	β	SE	t	$p(\beta = 0)$	Group	Predictor	Variance	Correlation
Intercept	3.467	0.032	107.592	$< .001$	Conversation	*Intercept*	0.006	-
Task experience	0.034	0.011	3.120	$< .01$	Speaker	*Intercept*	0.058	-
$\mathbf{SR_{t-1}}$	0.058	0.009	6.258	$< .001$		SR_{t-1}	0.001	-
Giver	-0.212	0.045	-4.675	$< .001$		Giver	0.111	-0.370
Male	0.200	0.060	3.348	$< .001$	*Residual*		0.441	

Of critical importance, we first note that in this second model of syllable rate the entrainment effect was still present (faster speech in the previous turn predicted faster speech in the current turn; $p < .001$), while the absence of an effect of the current turn suggests that there was no trend to either speed up or slow down across the length of each conversation.

Consistent with findings presented in Table 7.2, males were found to speak faster than females ($p < .01$). Givers of instructions were found to speak slower than followers ($p < .001$), while participants were found to increase in the rate at which they spoke each time they performed the map task ($p < .05$)

7.2.2 Corpus analysis 2b: Rate of speech and durations of ITIs

As Wilson and Wilson (2005) suggest that the durations of ITIs reflect the period of oscillators, we reasoned that faster speech (where syllables would be shorter) should tend to follow shorter gaps. In order to test this prediction, for this analysis only we first eliminated all overlaps (i.e. where ITI < 0), leaving 10,290 observations.

Table 7.5: Corpus Analysis 2b: Final mixed-effect model of ITI durations. Predictors-of-interest are shown in bold. The reported correlation is between the by-speaker random slope and the by-speaker random intercept; where no correlation is reported it is because the inclusion of a correlation did not significantly improve model fit.

					Random effects			
Fixed effect	β	SE	t	$p(\beta = 0)$	Group	Predictor	Variance	Correlation
Intercept	-0.816	0.026	-31.885	$< .001$	Conversation	*Intercept*	0.017	-
Length$_t$	0.100	0.010	9.970	$< .001$		SR$_t$	0.004	-
Task experience	-0.076	0.015	-5.032	$< .001$	Speaker	*Intercept*	0.027	-
t	-0.025	0.011	-2.293	$< .05$		SR$_t$	0.002	-
Giver	-0.221	0.029	-7.719	$< .001$		Giver	0.024	-0.687
Visible	0.197	0.047	4.180	$< .001$	*Residual*		0.839	
SR$_t$	-0.051	0.013	-4.014	$< .001$				

ITIs did not follow a normal distribution ($W = 134.277$, $p < .001$). This was expected as it has been suggested elsewhere (e.g. Campione & Véronis, 2002; Heldner & Edlund, 2010) that the durations of the gaps between turns tend to follow a log-normal distribution. Before performing the model construction process a Box-Cox transformation was applied to articulation rate ($\lambda_1 = 0.190$).[1]

See Table 7.3 for the full model of ITIs in the MTC. Consistent with the prediction of the oscillator theory, a relationship was observed between the syllable rate of speech produced in a turn and the duration of the ITI that preceded it ($p < .001$), with faster syllables rates being associated with shorter ITIs. This effect is shown in Figure 7.3.

The durations of ITIs were found to be related to the lengths of turns of the turns that followed ($p < .001$), with longer turns associated with faster ITIs. ITIs were found to reduce in duration as speakers gained more experience, both within one conversation ($p < .05$) and through repeated performance of the map task ($p < .001$). Givers of instructions were found to produce shorter ITIs than followers ($p < .001$). Finally, consistent with Bull and Aylett (1998), participants who were unable to see each other were found to produce shorter ITIs than those who were able ($p < .001$). No other predictors were found to significantly improve the fit of the model.

[1] We note that finding a value for λ_1 that was close to zero is consistent with the suggestion that these intervals show a tendency to be log-normally distributed.

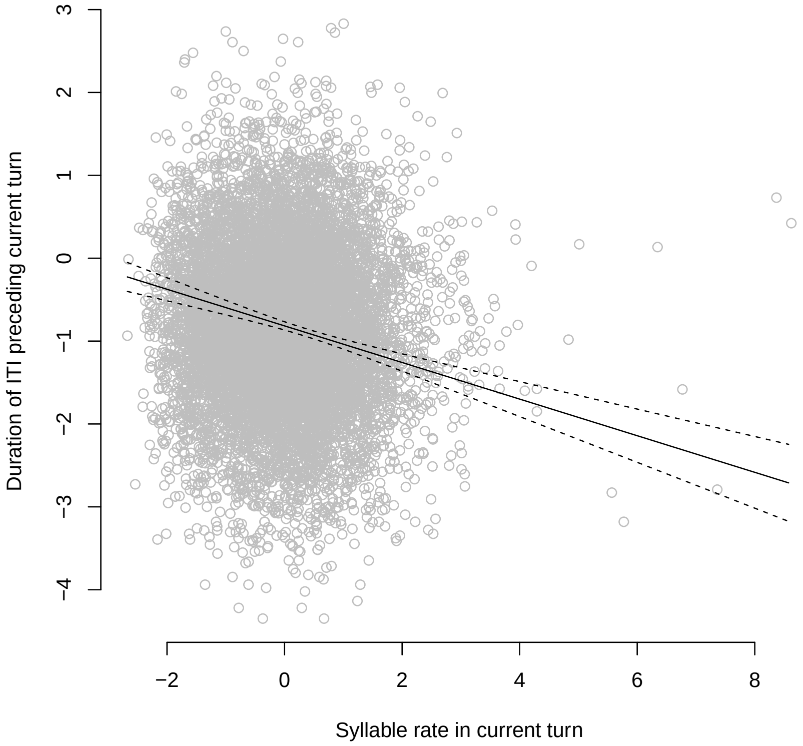


Figure 7.3: Corpus Analysis 2b: Duration of ITIs by the syllable rate of the following turn. The line represents the estimated effect, with confidence intervals. ITI durations have been Box-Cox transformed, while syllables rate in the current turn were standardised prior to the model construction process.

Wilson and Wilson's (2005) oscillator theory suggests that faster speakers respond quicker because their syllables have a shorter duration and therefore they reach their maximum readiness to speak earlier than slower speakers. An alternative explanation may be that faster speakers begin speaking sooner not because they have reached the peak of their cycle sooner, but because they are able to plan their utterance quicker. If faster speakers are able to prepare their utterances faster than slower speakers, then we may find that they are more likely to produce overlaps than slower speakers (because they are more likely to have their next utterance planned before their partner has finished). To test this possibility, we went back to the earlier dataset, which included overlaps, and tested a model of the likelihood that a speaker would produce an overlap.

Table 7.6: Corpus Analysis 2b: Final mixed-effect model of likelihood that a turn will be an overlap. Predictors-of-interest are shown in bold.

					Random effects		
Fixed effect	β	SE	t	$p(\beta = 0)$	Group	Predictor	Variance
Intercept	-0.987	0.048	-20.594	$< .001$	Conversation	*Intercept*	0.085
Map experience	0.137	0.070	1.958	.05		SR_t	< 0.001
t	0.083	0.022	3.811	$< .001$	Speaker	*Intercept*	0.075
Length_t	0.039	0.020	1.975	$< .05$		SR_t	0.010
Giver	0.353	0.056	6.352	$< .001$		Giver	0.021
$\mathbf{SR}_t$	0.096	0.024	3.982	$< .001$			
Male partner	-0.130	0.065	-1.999	$< .05$			

The full logistic mixed effects model of overlaps is shown in Table 7.6.[2] Consistent with the hypothesis that faster speakers are simply able to prepare their next utterance earlier than slower speakers, faster speakers were found to be more likely to produce overlaps ($p < .001$). Several other predictors were found to influence the likelihood of a turn being an overlap. As the conversation progressed, it was found that overlaps became increasingly likely to occur ($p < .001$). Givers were found to be more likely to produce overlapping turns than followers ($p < .001$). An effect of experience with the map was found to significantly improve the fit of the model; however, its coefficient was not significant in the final model. Consistent with Yuan et al. (2007), we found that male partners were less likely to be overlapped than females. Finally, a relationship was found between turn length and the likelihood that it would be an overlap ($p < .001$), with shorter turns more likely to be overlaps than longer turns. This could be because shorter turns were more likely to be backchannel responses. We would note that the inclusion of backchannel responses should not provide a challenge for our conclusion that faster speakers are more likely to begin speaking earlier than slower speakers, as we do not know of any evidence suggesting that faster speakers are any more likely to produce backchannel responses.

7.2.3 Corpus analysis 2c: Entrainment and the precision of turn-taking

In the oscillator theory, it is suggested that precision timing in turn-taking is achieved through the entrainment of rates of speech between conversational partners. This would suggest that speakers who were more entrained (i.e. those with

[2] Note that in order to count as an overlap a turn only had to begin within the previous turn. Therefore, what we term overlapping turns need not overlap in their entirety.

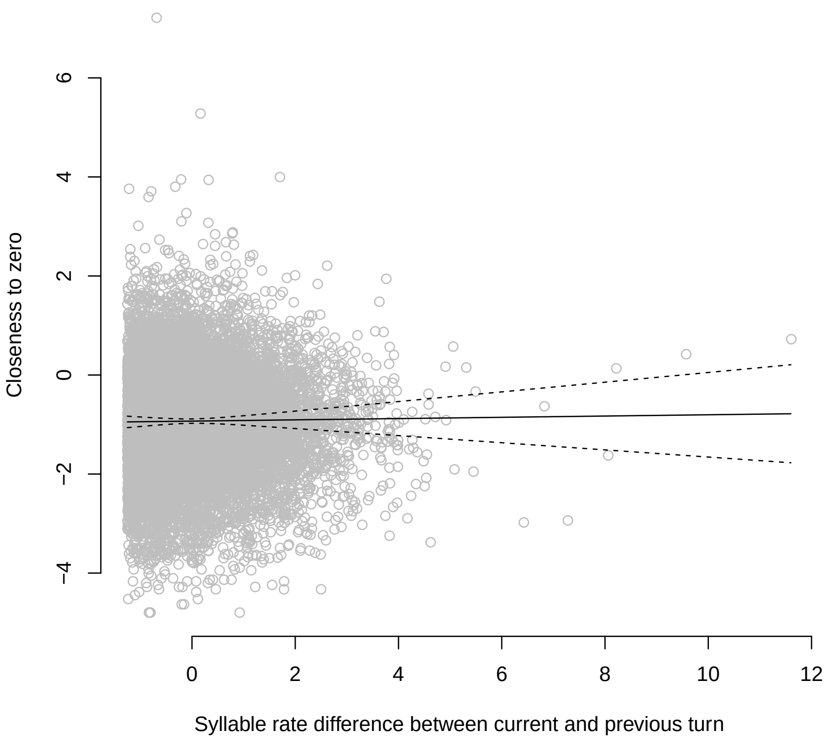


Figure 7.4: Corpus Analysis 2c: Closeness of ITI to zero by difference between partners' syllable rates. The line represents the estimated effect, with confidence intervals. Closeness of ITIs to zero have been Box-Cox transformed, while differences were standardised prior to the model construction process.

less difference between their syllable rates) would exhibit more precise turn-taking (achieving turn exchanges with ITIs that are close to zero).

As would be expected given that we used absolute values (therefore bounding the variable at 0), our measure of precision was found not to follow a normal distribution ($W = 433.442$, $p < .001$). Before performing the model construction process a Box-Cox transformation was applied to the measure ($\lambda_1 = 0.161$).

As is clear from Figure 7.4, no relationship was observed between the difference between syllable rates on either side of a turn exchange and the closeness of the ITI to zero ($\chi^2(1) = 0.27$).

The final model for the closeness to zero of ITIs is given in Table 7.7. The turn exchanges that preceded longer turns were found to be less close to zero than those preceding shorter turns ($p < .001$). The closeness to zero of ITIs was found

Table 7.7: Corpus Analysis 2c: Final mixed-effect model of closeness to zero of ITIs. Predictors-of-interest are shown in bold. The reported correlation is between the by-speaker random slope and the by-speaker random intercept; where no correlation is reported it is because the inclusion of a correlation did not significantly improve model fit.

					Random effects			
Fixed effect	β	SE	t	$p(\beta = 0)$	Group	Predictor	Variance	Correlation
Intercept	-0.931	0.022	-41.382	$< .001$	Conversation	*Intercept*	0.010	-
Length$_t$	0.104	0.009	11.711	$< .001$	Speaker	*Intercept*	0.022	-
Task experience	-0.077	0.016	-4.945	$< .001$		Friend	0.013	-
Map experience	0.013	0.044	0.294	.77		Giver	0.033	-0.771
t	-0.019	0.010	-1.997	$< .05$	*Residual*		0.969	
Giver	-0.205	0.036	-5.689	$< .001$				
Visible	0.168	0.039	4.337	$< .001$				
$\mathbf{\lvert SR_{\mathit{t}} - SR_{\mathit{t}-1} \rvert}$	-0.009	0.008	-1.095	.27				

to decrease across the length of the conversation ($p < .05$) and with repeated performances of the map task ($p < .001$). As in the previous analysis, the predictor for experience with the map itself was found to significantly improve the fit of the model; however, it's coefficient did not reach significance in the final model. Finally, givers of instructions, and participants who were not unable to see each other, were found to exhibit more ITIs that were closer to zero (both *ps* $< .001$).

Wilson and Wilson (2005) suggest that entrainment between partners may break down after a certain period of time, leading to an increase in simultaneous starts. In our data, this would result in disproportionately more ITIs that are close to zero occurring when the difference between speakers' rates is large (when entrainment has broken down). Wilson and Wilson suggest that this breakdown should occur after at least 1 second has passed; however, Beňuš (2009) observed that the peak occurrence of simultaneous starts was around 500ms. In order to ensure that a genuine effect of entrainment on precision was not being obscured by cases where entrainment had broken down we reran our model construction process, firstly excluding all observations where ITIs were over 1 second (leaving 12,674 observations) and secondly excluding all observations where ITIs were over 500ms (leaving 10,295 observations). In both cases, our measure of entrainment did not improve the fit of our models when it was tested for inclusion ($\chi^2(1) = 0.24$ and $\chi^2(1) < 0.01$, respectively).

7.3 Discussion

The present study tested three predictions derived from Wilson and Wilson's (2005) oscillator theory of turn-taking. Firstly, we tested the prediction that the rates of speech of conversational partners should become entrained. Secondly, we tested the prediction that people who are speaking fast will be quicker to begin a new turn than slower speakers. Finally, we tested the prediction that the entrainment of rates of speech will lead partners to achieve more precise turn-taking (i.e. closer to no-gap-no-overlap turn exchanges).

These predictions were addressed separately, each in a mixed effects regression performed on data from the MTC. Consistent with the first prediction, the rates of speech produced by participants in the MTC were found to be entrained, with faster speech in one turn leading to faster speech in the subsequent turn. Consistent with the second prediction, shorter ITIs were found to be followed by turns that exhibited faster speech. However, we found no evidence to support the critical third prediction of the oscillator account, that entrainment between speakers should produce ITIs that are closer to zero. Before further discussing these results, and their implications for theories of turn-taking, we will first discuss some other trends that were observed.

7.3.1 Other influences on rate of speech

In our first analysis of syllable rate we found that several of the factors which, on the basis of the existing literature, we strongly expected to influence rate of speech did not significantly improve the fit of our model. We reasoned that one possible explanation for this was that the effects of these factors may have been obscured by the inclusion of short turns, which, while not posing a problem for our test of the oscillator theory (as the rates would still capture the periodicity of the syllables produced, even for short turns), may be an inaccurate representation of the rate at which a person generally speaks. We subsequently ran a second analysis which excluded all turns containing fewer than 5 syllables (following the suggestion of Goldman-Eisler, 1954).

Consistent with previous findings (e.g. Binnenpoorte et al., 2005; Yuan et al., 2006), male participants in the MTC were found to speak faster than female speakers. The role that participants were performing in the MTC was also found to influence their syllable rate, with givers of instructions tending to speak slower than followers. In Chapter 5, we argued that performing the role of giver in the

MTC is likely to entail an increased level of cognitive difficulty compared to the follower role. This suggests that MTC participants may speak slower when they are under a cognitive burden (cf. Lay & Paivio, 1969). Seemingly consistent with this, the rate at which participants spoke was found to decrease with repeated performances of the map task. One possible explanation for this may be that as participants perform the task more often its difficulty is reduced (perhaps because they are able to reuse successful strategies from previous performances); however, an alternative explanation may be that through repeated performance participants become more comfortable with the task and it is this "relaxing" which causes speakers to speak slower. Similarly, a plausible alternative account for the relationship between role and rate of speech would be that givers speak slower in order to ensure that followers are able to follow what they are saying. As difficult ideas are likely to be those which are harder to follow, we might expect speakers to purposefully slow down when explaining difficult ideas to a partner. We would therefore suggest that any future research into the nature of the relationship between cognitive difficulty and speech rate should focus primarily on monologue tasks, where the speaker may have no reason to slow down to be more easily understood.

Another possible explanation is that the slow syllable rates are an artefact of the content that speakers produce. As they have to take the lead, givers are likely to be the first to mention particular words. As a result, if and when the follower repeats these words they are likely to be reduced in duration (e.g. Fowler & Housum, 1987), and their speech may appear slower than the giver's.

7.3.2 Other influences on the timing of turn-taking

At the outset of their study of ITIs in the MTC, Bull and Aylett (1998) make the point that the amount of time it takes for a person to begin a turn may largely reflect the cognitive pressures of planning the content of that turn. In this section we will discuss the findings of our analyses of the timing of turn-taking, which supports Bull and Aylett's suggestion.

We conducted several analyses on different aspects of turn-taking: the duration of gaps, the likelihood of producing an overlap, and the precision of turn exchanges. We begin by highlighting the general similarity between the results of our analysis of gap durations and of precision. This should perhaps not come as a surprise given that the majority of turn exchanges (69.6%) had an ITI of zero or above (consequently, the data analysed in Analysis 2b represents over two-thirds of the data analysed in Analysis 2c).

Before discussing our findings, we will first briefly make mention of familiarity. Previous studies have found some evidence that familiarity between conversational partners has an effect on their rate of speech and their turn-taking (Yuan et al., 2006, 2007). We are unable to explain why these effects were not replicated in the present study; however, we would speculate that, while having a casual, everyday, conversation with a friend may be quite different from having a conversation with a stranger, the difference between friends and strangers may be less marked when performing a cooperative task such as in the MTC (we also note the absence of effects of familiarity on the production of disfluencies in Chapter 5; although, some effects of familiarity have previously been found in the MTC; Boyle et al., 1994)

Our analysis of gap durations found several trends which were consistent with previous evidence, including some which replicate Bull and Aylett's (1998) findings. In an experiment where participants had to read a sentence and then recite it from memory, Ferreira (1991) found that participants took longer to begin speaking when reciting longer utterances than when reciting shorter utterances. As we suggested in Chapter 2, it is not entirely clear how comparable this task is to spontaneous speech; however, we observed a similar trend with the gaps prior to longer turns longer in duration than those before shorter turns. While we would not argue that participants were planning each of their turns in their entirety before they began speaking, our results do suggest that participants take longer to begin producing longer turns. Future research could explore whether Ferreira's other finding, that initiation times were longer before more syntactically complicated utterances, can also be replicated in spontaneous speech by building upon the coding of parts-of-speech and syntax already present in the MTC annotation.

Our analysis replicated Bull and Aylett's (1998) findings that participants who were givers of instructions in the MTC, and those who were unable to see their partner, produced shorter gaps than followers and those who could see their partner. The direction of the effect of role is perhaps surprising: Given our expectation that they face more cognitive difficulty, we might have predicted that givers should take longer to begin producing a turn, as they need more time to plan what they are going to say. In the present study, we observed the opposite. While we cannot be sure why this effect is in the direction that was observed, one possibility is that it is an artefact of the map task itself. When a follower ends a turn, and a giver begins a new turn, all that the new speaker has to do is to either plan a response to what has been said or to plan the next step in the instructions; however, when a giver ends a turn the follower might need to convert the instructions they have received into a route which they draw on their map. The reason followers may take longer to respond may therefore be that they take up time drawing a route. As the giver does not have to do this, they may be able to begin a new turn sooner.

7.3.3 Testing the oscillatory theory

Each of the analyses presented in this chapter were intended to test one of three predictions of Wilson and Wilson's (2005) oscillator theory of turn-taking. In the theory, it is suggested that each speaker in a conversation's readiness to initiate production of a syllable follows an oscillatory function, with the periods of these oscillations being the duration of a single syllable, and therefore the frequency being the syllable rate. At the peak of these oscillations the person is at their maximal readiness to speak, while at their lowest they are at their minimal readiness. The oscillators of each participant in the conversation are argued to become entrained in anti-phase, and consequently when the current speaker is at their maximal readiness, their interlocutors are at their minimal readiness. For the purpose of turn-taking, being entrained in anti-phase means that interlocutors will be ready to begin a new turn within half a syllables' range of the end of the previous turn.

In Corpus analysis 2a, we tested the prediction that interlocutors will become entrained, and consistent with the theory, we found a positive relationship between the syllable rates of turns on either side of a turn exchange (i.e. faster spoken turns are followed by faster spoken turns). Such entrainment of rate of speech is consistent with previous findings (e.g. Street, 1984; Webb, 1969), including a previous analysis of syllable rate in the MTC (Finlayson et al., 2010).

In Corpus analysis 2b, we tested the prediction that faster speech will tend to be preceded by gaps of shorter duration (as faster speech reflects faster cycles of oscillations and shorter gaps reflect reaching the peak of a cycle faster). Again we found evidence consistent with the theory, with shorter gaps being followed by turns that were produced at a faster rate. As earlier suggested, finding evidence in support of this prediction would not provide unequivocal support for the oscillator theory: One possible explanation for this finding could be that faster speakers may be those who are able to prepare utterances quicker than slower speakers. We suggested that being able to prepare an utterance quicker may result in an increase in overlaps, as the utterance may be prepared before the previous speaker has finished their turn. While Wilson and Wilson's (2005) presentation of their theory provides no reason for us to believe that they would predict a relationship between rate of speech and the likelihood of producing an overlap (although it would make predictions about the timing of overlaps), our alternative explanation would predict that faster speakers may be more likely than slower speakers to produce overlaps. Consistent with an account suggesting that rate of speech reflects the amount of time necessary to plan an utterance, rather than the time needed to reach the peak of a cycle of readiness, faster speakers were more likely to produce overlaps in the MTC. For this reason we would suggest that the results of Corpus analysis 2b cannot be taken as offering unequivocal support for the oscillator theory.

If the entrainment of oscillators is the means by which people become able to precisely time the beginnings of their turns, relative to the ends of their partners' turns, then we would expect that people who are more closely entrained should achieve more precisely timed turn exchanges. Testing this prediction was the purpose of Corpus analysis 2c. Two oscillators that are perfectly entrained will possess the same frequency. As, in the oscillator theory, the frequency of oscillations corresponds to speakers' syllable rates, in order to quantify the degree of entrainment we calculated the absolute difference between both participants' syllables rates at each turn exchange. Given that Wilson and Wilson's theory is intended to explain how people come to precisely time the beginnings of their turns, as a measure of precision we took the absolute difference between the beginning of a new turn and the end of the previous turn. Our analysis did not find any evidence to support the prediction that speakers whose syllable rates were more closely entrained exhibited any more or less precision in their turn-taking.

The previous empirical investigations of the oscillator theory that we presented in 6.3.2 could perhaps be generously described as offering mixed support for the oscillator theory. In a study that combined inferential and descriptive methods, Beňuš (2009) did not find evidence that speakers were becoming entrained on their syllable rates, nor that ITIs were bimodally distributed around 0ms. However, consistent with the claim that simultaneous starts become more common when entrainment breaks down, he did find evidence that simultaneous starts became more common after 500ms (although this is at least 500ms earlier than Wilson and Wilson predict). Ten Bosch et al. (2005) found evidence that appears to show that conversational partners' gap durations become entrained, as would be predicted by the oscillator theory, although it is not clear that this finding is not confounded by a general trend for gaps to become shorter across the length of a conversation (as we observed in our own Corpus analysis 2b). Finally, Włodarczak et al. (2012) found that, as the oscillatory theory would predict, the onset of overlaps are not randomly distributed throughout the duration of a syllable, however, their data would be consistent with a version of the oscillator theory where people become entrained in phase rather than in anti-phase, as Wilson and Wilson suggest.

Taking together each of our analyses, our results do not provide much in the way of support for the oscillatory theory. We did observe entrainment of syllable rate; however, this alone is not evidence that entrainment of syllable rate occurs through the entrainment of oscillators (e.g., syllable rates could have been becoming entrained because participants were making a conscious effort to speak at a similar rate to their partner). Faster spoken turns were preceded by gaps of shorter duration; however, further analyses suggest that this may just reflect a tendency for faster speakers to prepare their utterances earlier. Finally, in our strongest test of the oscillator theory, we found no evidence that entrainment of articulation rate was having any effect at all on the timing of turn-taking. Given the general lack of support found in the present study, and in previous studies, we would suggest that it may be worth considering alternative accounts of the timing of turn-taking. In the remainder of this chapter we will discuss an account of turn-taking that has emerged from a recent psycholinguistic account of the architecture of the language systems.

7.3.4 Precision through prediction

Gambi and Pickering (2011) have recently suggested an alternative account of the timing of turn-taking which is compatible with evidence observed both in previous studies and in our own. In this account, conversational partners are able to predict the ends of each others' turns because they develop predictions about what they expect the other to say and can consequently develop predictions about how long the speaker will take to produce their utterance. Such predictions should therefore allow each person to time their own turn to begin close to the end of the other's turn.

Gambi and Pickering's (2011) account of turn-taking is part of a larger theoretical framework which proposes an architecture for the language system where production and comprehension are entwined, and where both involve the use of prediction (Pickering & Garrod, 2007, 2013). Integral to Pickering and Garrod's theory is the claim that during language production speakers construct forward models which allow them to monitor the content of their utterances. In theories of motor control, forward models are generated from an efference copy of the original action command (see Miall & Wolpert, 1996), and they provide a prediction of the sensory feedback of the action. For example, if a person was reaching out to grab a cup then the forward model would provide an expectation of where the person's arm might be at any point, the velocity at which it would be travelling, the position of the fingers etc. By comparing these predictions with the actual sensory percepts of reaching for the cup, the person is able to evaluate the progress of the action and make necessary adjustments to avoid errors (e.g. "falling short" of the cup).

Pickering and Garrod (2013) propose that forward models perform a similar monitoring role during language production. When a speaker initiates a production command, the intended utterance, an efference copy of this command is run through the forward model and speakers subsequently generate predictions of the percepts of the utterance (these may include its semantics, its syntax, and its phonology). These predictions can then be compared with the percepts of the actual utterance (e.g. comparing actual and predicted semantics, actual and predicted syntax, and so on) and if any errors are detected then they can be corrected.

Within Pickering and Garrod's framework, the systems of production are not used only to produce the speaker's own utterances. It is argued that production processes are used in the comprehension of the utterances of interlocutors. While it has long been considered that comprehension systems may be employed during production (for example, the use of the external loop for monitoring; Levelt, 1989; Levelt, Roelofs, & Meyer, 1999), less consideration has been given to whether the production systems are employed during comprehension. There is, however, an increasing amount of evidence to suggest that the production system is activated when listening to speech. Listening to speech has been found to lead to increased activation in the articulators (for examples, the tongue and lips; Fadiga, Craighero, Buccino, & Rizzolatti, 2002; Watkins, Strafella, & Paus, 2003) and in areas of the motor cortex thought to be involved in speech (Pulvermüller et al., 2006; Watkins et al., 2003). Furthermore, there is increasing evidence that production and comprehension processes share regions of the brain (e.g. Heim, Opitz, Müller, & Friederici, 2003; Menenti, Gierhan, Segaert, & Hagoort, 2011; Wilson, Saygin, Sereno, & Iacoboni, 2004; for reviews, see Mar, 2004; Pulvermüller, 2010; Pulvermüller & Fadiga, 2010; Scott, McGettigan, & Eisner, 2009).

In Pickering and Garrod's theory, the proposed predictive abilities of forward models and the production system are applied during comprehension of language (similarly, forward models have been proposed to be used for prediction in action-perception; see Wolpert, Doya, & Kawato, 2003). During comprehension, it is argued that listeners use the production system to generate predicted percepts of incoming speech. This is thought to occur through a series of stages. Listeners covertly imitate utterances as they unfold, firstly deriving the production command which produced the utterance as it has already unfolded and then deriving the production command that they anticipate to unfold. By running this second production command through their own forward production model (i.e. the forward production model they use to generate predicted percepts during their own production) they can generate predictions of percepts of the speaker's utterance (similar to during production, these may include predictions of semantics, syntax and phonology), and, as forward modelling is assumed to be faster than generating actual production commands, these predictions are obtained before the speaker has produced the utterance.

There is now considerable evidence that prediction occurs during language production. Predictions may occur at a variety of levels of linguistic representations, including phonology (DeLong et al., 2005); syntax (Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005; Wicha, Moreno, & Kutas, 2004); and semantics, where studies using the visual world paradigm have allowed researchers to observe anticipatory eye movements (i.e. looking at an item before it is mentioned; e.g. Altmann & Kamide, 1999; Kamide, Altmann, & Haywood, 2003; Kamide, Scheepers, & Altmann, 2003; Knoeferle, Crocker, Scheepers, & Pickering, 2005) (for a review of prediction in language comprehension, see Kutas, Delong, & Smith, 2011). Further evidence argued to reflect prediction has come from a study where participants were recorded telling stories during an fMRI scan, "as if telling the story to a friend", while separate participants listened to these recordings, also during an fMRI scan (Stephens, Silbert, & Hasson, 2010). When the scans of speakers and listeners were compared, a great amount of overlap in areas of activity was observed with a delay between speakers and listeners (i.e.

activity observed in the speaker at time n was observed in the listener at time $n+1$). There was also, however, some evidence of anticipatory activation in some regions (i.e. activity observed in the listener at n and in the speaker at $n+1$). Furthermore, the amount of anticipatory activation was found to correlate with the listeners' understanding of the story (operationalised as the amount of the story they could recall), suggesting that prediction may facilitate communicative success (at least when success is defined as remembering what was said).

Gambi and Pickering (2011) develop Pickering and Garrod's idea further by proposing that listeners use the predictions that they generate about speakers' utterances in order to further generate predictions of the timings of these utterances. If listeners are able to make predictions about the timings of utterances during comprehension then we might expect that they should make similar predictions about the timings of their own utterances during production. Consistent with this, Griffin (2003) found that when asked to name two pictures, at least one of which, the second picture to be named, depicted an item with a polysyllabic name, participants were faster to begin naming when the first image also depicted an item with a polysyllabic name than when the item was monosyllabic. She suggested that participants were delaying naming the first image when it was monosyllabic because they knew that they would not have enough time to prepare the second image while naming the first image; however, naming a polysyllabic item would provide enough time to prepare the name of the second polysyllabic image (for an alternative account of this, and related findings, see Meyer, Belke, Häcker, & Mortensen, 2007)

If listeners are able to predict the percepts of a speakers' utterances, and the amount of time necessary for various stages of producing the actual utterances, then, Gambi and Pickering suggest, they may be able to predict when an utterance will end. For example, if you hear someone say "the day was breezy so the boy went outside to fly...", then you might not only predict that they will say "a kite" but also how long it will take them to produce it. By doing so, you may then be able to precisely time the production of your own utterance to begin after the speaker has finished the sentence. We will call this the *precision through prediction* account.

Magyari and De Ruiter (2008) provide some evidence which is consistent with the precision through prediction account. They took unfiltered utterances from De Ruiter et al. (2006) and divided them into those where participants had been generally good at anticipating a turn ending (high bias) and those where they had been generally poor (low bias). New participants heard versions of these utterances which cut off at various points throughout the utterances. They were then asked to predict the next word in the utterance. Participants were found to be more accurate for high bias items than for low bias items. Furthermore, when asked to predict how many words would follow the cut-off, participants were found to predict a greater number of words for low bias items. Taken together, this may suggest that participants were using their predictions of upcoming words to anticipate turn endings; while, when participants were poor at anticipating turn endings it may have been because they expected more words to be produced.

While our own analyses of the timing of turn-taking do not allow us to test the precision by prediction account, several of our findings could be explained by such an account. Our results reveal two sets of effects of experience on the timing of turn-taking. Firstly, participants in the MTC were found to produce shorter, more precise, turn exchanges each time they performed the map task. We might expect that repeated performance of the map task would increase the predictability of utterances, as participants could have begun to anticipate the ways in which others may have tended to talk about the task. Secondly, turn exchanges became shorter and more precise across the length of each conversation. Similar to the effects of task experience, participants may have been learning how their partner tended to talk about the map task but moreover we might expect that increased exposure to a particular speaker would increase the accuracy with which a listener can predict the timing of their production. Future research could test whether increased exposure to a speaker improves the accuracy of predictions by investigating whether accuracy improves across trials when listening to multiple utterances produced by the same speaker in turn-ending anticipation tasks (such as in De Ruiter et al., 2006) (multiple speakers may be required to avoid the possible confound of a general learning effect).

The ability to predict features of upcoming words would be compatible with Sacks et al.'s (1974) projection-based account of turn-taking. If listeners are able to predict syntactic features of upcoming words then this may allow them to predict when a TCU will end. For example, if an utterance could be syntactically completed by a noun then if the listener has predicted that a noun will follow then they may further anticipate that a TRP will occur after the next word. Similarly, the precision by prediction account is not incompatible with existing evidence for the possible use of cues in turn-taking. Being able to predict upcoming words may not always be sufficient to predict the end of a turn. As Wennerstrom and

Siegel (2003) suggest, from the perspective of the listeners there may be many moments of possible syntactic completion before the speaker actually completes their utterance. In cases like these, or in cases where strong predictions about upcoming material cannot be made (the extent of prediction in comprehension, and how common it is outside of tightly controlled laboratory conditions, are both still open questions), listeners may still rely on cues to help determine that a turn is ending. For example, if a listener has predicted the timing of an upcoming word and when it is being produced its intonation contour is consistent with a potential turn-ending then the listener may be more likely to anticipate that it will be the end of a turn and plan the timing of their own next utterance.

Before concluding, we will first raise two speculative points which we would argue are worthy of future investigation. The first of these is that is worth briefly noting that the imitation which Pickering and Garrod (2013) argue occurs during comprehension could provide an explanation for entrainment of rate of speech. We would expect that imitation should proceed at the same rate as the speaker is speaking. A speaker obviously could not imitate an utterance before it has been produced, and they would likely not lag behind the speaker, else the listener would be redundantly generating predictions about utterances which have already been produced. As the listener's own production system is employed for imitation, the rate at which this system is running will be similar to the rate of the speaker's own production system. When the listener begins to speak after listening to the previous speaker the consequence of imitation during comprehension may be that their rate of speech will be similar to the rate at which they were just comprehending (and consequently, the rate at which the previous speaker spoke). An account where entrainment is the by-product of imitation would explain Jungers and Hupp's (2009) priming effects, as there is no reason to believe that listeners will only imitate the speech of speakers with who they are engaged in conversation. However, we would also note that such an account would likely make very similar predictions to an account where entrainment of rate occurs through the entrainment of oscillators (although this need not have any direct consequences for turn-taking). Therefore, testing this explanation, or at least finding evidence which unequivocally supported this explanation, may prove difficult. One possible difference between these accounts which may be worthy of future research would lie in the predictions that each account makes about entrainment to speech where words have been filtered out (which could be achieved using similar approaches to those employed by De Ruiter et al., 2006). If participants heard this speech as part of a priming task (similar to that in

Jungers & Hupp, 2009) then, as long as rhythm was retained, an oscillator-based account would likely predict that priming should still occur, while an imitation-based account would predict that there would be no (or at least reduced) priming, as there would be fewer linguistic representations to imitate.

Our second point is that it is possible that the organisation that turn-taking is thought to exhibit may be, in part, illusory. One of the analyses we carried out investigated the likelihood that a particular turn would be an overlap. What was particularly striking about the results of this regression was that many of the factors which predicted shorter gaps (e.g. faster syllable rate in the subsequent turn, a shorter subsequent turn, being further along in the conversation, and being the giver) were also found to predict an increased likelihood of producing an overlap. Taken together, this suggests that participants were not just beginning new turns as close to the end of the last turn as possible; rather, they were beginning new turns as early as possible, regardless of whether or not their partner had finished speaking. Such a claim would be somewhat in contrast to the precision through prediction account; however possessing the ability to make predictions about others' utterances may mean that listeners need not always have wait for a speaker to complete an utterance before they can plan their own response. Similarly, we could imagine a situation where a person decides that they no longer need to listen to what a speaker is saying. For example, a person asking "what is the time?", and receiving the response "I don't know. Unfortunately my watch broke last week after it got wet." may decide soon after hearing "I don't know" that they have heard all that they need and may begin another turn (e.g. by saying "Thanks anyway").

The claim that people begin speaking as soon as they are ready, regardless of whether their partner has finished their turn, would also seem to be in contrast to the idea, at least implicit in several prominent theories of turn-taking (e.g. Duncan, 1972; Sacks et al., 1974; Wilson & Wilson, 2005), that the "goal" of turn-taking is to begin speaking as soon as the previous speaker has finished their turn.[3] However, it need not necessarily be incompatible with the evidence cited at the beginning of the previous chapter that the majority of turn exchanges feature short gaps. While listeners may be able to predict what the speakers will

[3] O'Connell et al. (1990) make a similar claim that people in conversation are not concerned with producing smooth transitions between turns. They note that there may be situations where people regularly produce simultaneous speech (e.g. in having an argument about politics), or produce long gaps (e.g. two old men talking in a pub), without the conversation "breaking down".

say some of the time, there may be many more cases where they have to wait until the end of the speaker's turn. In these cases, projection that a turn will soon end (even if they are unsure as to exactly how it will end) could still be employed.

Considering the existing literature on turn-taking, it would be rather radical, of course, to claim that people are concerned with beginning their own turn as soon as possible, rather than with waiting for their partner to finish speaking. Therefore, we would suggest that further research should be undertaken to determine whether this apparent tendency to speak as soon as possible occurs only in task orientated dialogue, such as in the MTC, or whether it is in fact a general tendency in conversation. We might expect that if there is a general tendency for people to begin turns as soon as possible, rather than as soon as their interlocutor has finished their turn, then this may vary according to context. For example, while such overlaps may be relatively common in conversations between friends, people may be less likely to interrupt the turns of an interlocutor in a position of authority, such as in a meeting with their boss. If a general tendency is identified, then the possibility of such contextual variability may also be worthy of investigation.

7.4 Conclusion

In this chapter we tested Wilson and Wilson's (2005) oscillator theory with analyses of the MTC. While, consistent with the oscillator theory, we found evidence that participants in the MTC were becoming entrained on their syllable rates, and that turns produced at faster rates tended to be preceded by gaps of shorter duration, we crucially found no evidence to suggest that the entrainment of rate of speech led to increased precision at turn exchanges. In light of this, we suggested that Gambi and Pickering's (2011) precision through prediction account of the timing of turn-taking, where listeners are able to predict when a speaker's turn will end by first predicting features of the words that they are likely to use, may be worthy of consideration. Such an account is consistent with findings observed in the analyses presented in this chapter (e.g. that increased experience leads to more precise turn exchanges). The broader framework in which this account originates (Pickering & Garrod, 2007, 2013) may also explain the entrainment of rate of speech, where such entrainment could be a "by-product" of using the production system during speech comprehension.

Part III

General discussion

Chapter 8

General discussion

This thesis presented a series of experiments and corpus analyses which were intended to test two different theories about the ways in which speakers coordinate conversation, in particular about how turn-taking is managed. In the first theory, termed by us the hesitation-as-signal hypothesis, it is argued that speakers design certain hesitations, such as filled pauses and repetitions, in order to perform communicative functions during conversation (Clark, 1996, 2002). In particular, hesitations may be produced as signals of difficulty that allow speakers to account for their use of time when language production is disrupted. One reason why speakers might want to be able to account for themselves during a disruption is that the delay that the disruption provides could be interpreted by a partner in conversation as the end of the speaker's turn. By signalling that their silence is only temporary, the speaker may be able to keep hold of their turn.

In the second theory, Wilson and Wilson's (2005) oscillator theory of turn-taking, it is argued that each person in a conversation is able to make precise predictions about when others will finish their turns. This allows them to time the initiations of their own turn to achieve seamless turn exchanges. Such precision timing is argued to be afforded by endogenous oscillators in each partner, reflecting their readiness to initiate the production of a syllable (and, therefore, the speakers' rate of speech), which become entrained during conversation.

Before discuss our findings, we would first comment that when we introduced the Map Task Corpus (MTC) in Chapter 3, we mentioned that the motivation of its creators was to provide a resource which would allow many different questions about language and communication to be answered. True to this ideal, in this thesis we presented research which used the MTC to investigate factors influencing the likelihood of producing disfluencies, the rate at which people speak, and the timing of turn-taking. We would suggest that there is much more that can be learnt from the MTC, particularly with the use of mixed effects regression, which allow researchers to control for many of the possible sources of noise that may be present in a corpus of this type.

8.1 Are hesitations designed?

If people are designing their hesitations for the benefit of their audience in conversation, then we might expect that hearing a hesitation should have some effect on listeners' linguistic processing. There is an increasing amount of evidence to suggest that hearing filled pauses can have effects on linguistic processing; however, there is relatively little evidence that similar effects are produced by hearing a repetition.

In Experiment 1 (Chapter 4) we used a change detection paradigm to investigate whether hearing a repetition has an effect on listeners' attention which may lead them to form fuller semantic representations of subsequent words. While such effects have previously been observed for filled pauses, we did not finding similar evidence for repetitions.

It has been argued elsewhere that effects of hesitations on listeners could result from the delay for processing that they provide (i.e. it is the pause, rather than the *uh*, that drives the effects). In order to control this possible confound, we also tested the effects of silent pauses (which only provide a delay) on attention. Interestingly, while effects were lacking for repetitions, we found evidence suggesting that hearing a silent pause heightened participants' attention which suggests that they were semantically representing the following word in greater detail. Taken together, these two results lend further support to the suggestion made elsewhere (MacGregor et al., 2009) that any benefit that may arise from hearing a hesitation is not present when that hesitation is filled with linguistic material (such as in a repetition).

The focus of Chapter 5 was on directly testing the hesitation-as-signal hypothesis. We suggested that one problem that arises when testing the hypothesis that hesitations are being designed is that they may alternatively be being produced as symptoms of the difficulties that they are argued to signal. As a result, it is not enough to show that speakers produce hesitations when they encounter difficulty, as this tells us little about whether they are symptoms or signals of the difficulty.

We reasoned that if speakers are designing their hesitations as signals for their audience then they should be more likely to produce hesitations in dialogue, where they are communicating with an interlocutor, than in monologue, where they are not. In Experiment 2, we found that while speakers were designing one aspect of their speech for their interlocutor (they showed a tendency to reuse their interlocutor's referring expressions), they were equally likely to produce hesitations in dialogue and in monologue.

In Corpus Analysis 1, we used data from the MTC to investigate whether the production of hesitations varies according to manipulations of aspects of the dialogue in manners which suggest that they are designed by speakers to have a communicative function. If filled pauses and repetitions were being designed by speakers in order to perform communicative functions then we would expect them to be performed more frequently when those functions were needed. One function that filled pauses and repetitions have been argued to serve is to allow speakers to account for their use of time when their speech is disrupted. This may be particularly important if there was a possibility that the listener may interpret the speaker as having finished speaking (rather than just hesitating). If this is the case, then we would expect speakers to be more likely to produce hesitations when partners are unable to see each other (depriving the speaker of the chance to use non-verbal cues). Our analyses did not, however, find any evidence that the ability for speakers to see each other was having any effect on their hesitations. Nor were speakers any more or less likely to produce hesitations when speaking to a friend than when speaking to a stranger.

Consistent in both Experiment 2 and Corpus Analysis 1 was the finding that speakers were generally more likely to produce hesitations at points when we expected that they would be experiencing difficulty (e.g. when identifying a hard-to-name image, or performing the cognitively demanding giver role in the MTC). As we found evidence that the production of hesitations is associated with difficulty in language production, but no evidence that they are designed to signal this difficulty, we argued that the most parsimonious account of our findings is that hesitations are symptoms, but not signals, of the difficulty that speakers experience. While hesitations may help a speaker to hold onto the floor (by signifying to the listener that they have not finished), they are not designed by speakers for this purpose.

8.1.1 Testing the hesitation-as-signal hypothesis

Earlier, we reiterated the point that one difficulty with testing the hypothesis is that evidence that would be consistent with the claim that hesitations are designed as signals of difficulty (e.g. producing hesitations when experiencing difficulty) would also be consistent with the claim that hesitations are natural symptoms produced by a disrupted language production system. An even greater challenge is that there is little indication in the literature about what exactly would be required in order to falsify the claim that hesitations are being designed by speakers. Instead, the general trend has been to collect evidence which is consistent with the claim and then infer the speakers' intentions (e.g. observing that there are systematic differences in the delays following different realisations of filled pauses, and then inferring that the speaker intends to signal this difference).

Building on suggestions made elsewhere (e.g. Kraljic & Brennan, 2005; Nicholson, 2007; Schober & Brennan, 2003), we reasoned that if hesitations are being designed with a communicative function then manipulation of aspects of the context of communication (e.g. whether or not there was someone to communicate with, whether a person could see the person they were communicating with) should have an influence on the hesitations that speakers produce. In light of our null results, a proponent of the hesitation-as-signal hypothesis could perhaps suggest that our manipulations were not suitable for testing the claim; however, if this was the case then it is far from clear what the appropriate manipulations would be. We would therefore suggest that if the hesitation-as-signal hypothesis is to remain a viable account of why speakers produce hesitations then it is imperative that its proponents make explicit the predictions it would make.

8.1.2 The heterogeneity of hesitations

In both Experiment 1 and Corpus Analysis 1, as well as our review of the disfluency literature in Chapter 2, we saw demonstrations that repetitions are different in several respects from other types of hesitations, such as filled pauses. From the standpoint of comprehension, we do not know of any example of an effect that has been observed to result from hearing a filled pause that has also been observed to result from hearing a repetition. Similarly, factors known to be associated with the production of filled pauses (e.g. difficulty during lexical access) have generally not been found to be associated with the production of repetitions. In our analysis of the production of disfluencies in the MTC (Corpus Analysis 1), the cognitive burden thought to be associated with taking the role of giver of instructions was found to increase the likelihood of producing repair and filled pause J-tokens, but not repetition J-tokens. Taking together work on the production of disfluencies, it would be fair to say that we know very little about why speakers come to produce repetitions.

The gap in our current knowledge is not, however, restricted to repetitions. One may argue that in one sense we generally know very little about why speakers produce any of the hesitations that they do. When introducing hesitations in Chapter 2, we saw that they may come in a variety of different forms. Hesitating speakers produce filled pauses (sometimes producing *uh* and other times producing *um*), repetitions, prolongations, and, of course, in some cases they simply remain silent. While some of the difficulties that lead to the production of certain hesitations may be known, what remains a mystery is how the difficulty leads to the production of particular types of hesitations.

One appeal of the hesitation-as-signal hypothesis is that it could explain why hesitations come in a variety of forms. If we wanted to understand why speakers sometimes produce a filled pause, while other times they produce a repetition, we could point to Clark and Fox Tree (2002), who suggest that filled pauses are produced as a signal of an upcoming delay, and Clark and Wasow (1998), who suggest that one reason for producing a repetition may be to make a preliminary commitment to an utterance (although, we would note that the goal of both hesitations may be similar: stopping an interlocutor from interpreting the disruption as the end of the speaker's turn). However, if we were to accept that hesitations are merely symptoms of difficulty, then it is not clear why these different types occur. For example, why does a speaker say *uh* rather than repeating the last word that they said. In some cases, the answer may lie in the position in the utterance at which the disruption occurs. For example, beginning an utterance by repeating the last word said in a previous utterance would seem unlikely (although the speaker could repeat the first word of the current utterance), and if a disruption occurred after a word had been articulated then it would, of course, not be possible to retrospectively produce a prolongation. There are many cases, though, where more than one type of hesitation is possible (e.g. between two words a speaker could produce either a silent or filled pause, among other alternatives).

It could well be the case that particular types of difficulty result in particular types of hesitations (although there are examples where one source of difficulty, e.g. low name agreement, has been shown to be responsible for the production of several types of hesitation). If so, then we ought to investigate what it is about the architecture of the production system that gives rise to these associations. One example of an account of such an association is Blackmer and Mitton's (1991) proposed autonomous restart capability (introduced in Chapter 4 and tested in Chapter 5), where it is argued that repetitions occur because an articulator that has no new material to produce resorts to reproducing old material. Of course, here we could reasonably ask why the articulator produces anything at all: If it has nothing to say then why say anything?

Much progress has been made in the past sixty years to reveal the types of difficulty which are associated with the production of hesitations. We would suggest that the next step that ought to be taken is to begin to investigate why it is that particular hesitations are produced. In light of the absence of evidence suggesting that hesitations are being designed as signals, such attempts should be concerned with explaining the production of different types of hesitations in terms of the properties of the language production system, rather than in proposed communicative functions that different types of hesitations could serve.

8.2 Rate of speech and the timing of turn-taking

Corpus Analysis 2 used data from the MTC to test three predictions derived from Wilson and Wilson's (2005) oscillator theory of turn-taking about the relationships between rate of speech and the durations of inter-turn intervals (ITIs; the interval between one turn ending and the subsequent turn beginning). The oscillator theory predicts that conversational partners should come to speak at similar rates as endogenous oscillators, representing their readiness to initiate a syllable, become entrained. Consistent with this, we found that the rate at which a speaker spoke in one turn was related to the rate at which their partner spoke in the previous turn.

As the oscillators that become entrained represent a speaker's readiness to initiate a syllable, we reasoned that people who speak faster (those whose oscillators would have a shorter periodicity) should begin a turn earlier, relative to the end of their partner's turn, than those who speak slower (as they would reach their maximal readiness to initiate a syllable sooner). Consistent with this, we found that turns that were delivered at faster rates were initiated earlier than turns that were delivered at slower rates.

As the entrainment of oscillators is the means by which speakers are able to make more precise predictions about when their partner will end their turn, we reasoned that speakers who were more entrained with their partner would produce more "precise" turn exchanges. Operationalising precision as the closeness of an ITI to zero, we found no evidence that pairs of speakers who were more closely entrained (i.e. those with smaller differences between their rates of speech) produced any more precise turn exchanges than those who were less entrained.

8.2.1 Turn-yielding cues or turn-yielding signals?

In Chapter 6 we raised the point that it is not entirely clear what the nature of turn-yielding cues is. They could be cues that a turn *will* end, or they could be cues that appear regularly at points at which turn exchanges could occur. If it is the former then a reasonable question to ask would be whether speakers are designing turn-yielding cues as an invitation for their partner to take a turn (rather than the cues occurring as some sort of "side effect" of a speaker reaching the end of their turn). If this was the case then they would clearly meet a Gricean definition of a signal: They would be being produced with the intention that an interlocutor respond by interpreting them as an invitation to begin a new turn.

If one wanted to test whether turn-yielding cues are, in fact, turn-yielding signals then they could adopt an approach similar to that taken in our testing of the hesitation-as-signal hypothesis. We would anticipate that it should be possible to make experimental manipulations which, if turn-yielding cues were being designed as signals, would influence their production. For example, conversational partners could be provided with an external means of regulating turn-taking, such as having partners converse using "walkie-talkies". In such a paradigm, an auditor would be unable to begin a turn until the current speaker had released the button to finish broadcasting (the auditor could be provided with visual feedback to make it clear when the button had been released). If speakers were designing acoustic cues to invite their interlocutor to speak then we might expect that, relative to a telephone conversation, they should produce fewer cues when talking over walkie-talkies.

8.3 Conclusion

Both of the theories tested in the studies presented in this thesis originate from the similar perspective that conversation is organised by following a series of rules and principles. The hesitation-as-signal hypothesis is based on the assumption that partners in conversation have an obligation to account for the ways in which they use each others' time (Clark, 1996, 2002). When a person cannot account for their use of time by speaking, because their production has been disrupted, they produce a hesitation to signal that they are experiencing a delay. Despite proposing a mechanism for the timing of turn-taking that is grounded in neurophysiology, the oscillator theory is heavily influenced by Sacks et al.'s (1974) ideas about turn-taking. In Sacks et al.'s account, conversational partners wait for speakers to finish and then attempt to begin a turn themselves as soon as possible. Knowledge about what a TCU could consist of, and that each speaker has the right to produce at least one of these per turn, allows people to project when a turn will end, with a series of selection rules allowing partners to coordinate who will speak next.

In the empirical studies presented in this thesis we saw little clear support provided for either the hesitation-as-signal hypothesis or the oscillator theory. Consequently, we argued for alternative accounts of why hesitations are produced, and of the timing of turn-taking. These accounts do not rely on assumptions that conversation is regulated in ways suggested by Clark and Sacks et al. In explaining why it might be that speakers produce hesitations, we suggested that they arise as symptoms of difficulty that a speaker experiences. They are, in effect, the sound of the production system "breaking down". In discussing the timing of turn-taking, we suggested that Gambi and Pickering's (2011) theory, where the timing of an utterance may be one of many aspects of an utterance that are predicted, may explain the frequency of very short ITIs. However, we also speculated that perhaps auditors may simply begin speaking at the point at which they either believe they know what the speaker is going to say (through predictive processes) or when they feel they have taken what they need from the speaker's utterance. While Gambi and Pickering's theory would not necessarily be incompatible with Sacks et al.'s account, the suggestion that the goal of people in conversation is really to begin their own turn as soon as possible would be a radical departure from Sacks et al.'s ideas of cooperation. On the basis of the work presented in this thesis, we would suggest that it may be difficult to support the view that conversation is as well organised as some authors have claimed.

Appendix A

Passages used in Experiment 1

No change, close change and distant change words, respectively, are given in bold.

1. The doctor checked to see how much longer he had to work. He saw that the patient with the **virus / infection / tissue** was at the front of the queue. A kind but strict-looking nurse brought the boy in.
2. We all wondered where the new employee was going. It was obvious the woman carrying the **rucksack / backpack / briefcase** was a bit lost. In such a big complex it's so easy to lose your way.
3. Tony heard all about the celebrities at the Oscar ceremony. Apparently the film about the **aliens / martians / dinosaurs** had been universally praised. Everybody thought it had been a wonderful ceremony.
4. Simon really needed to decide what to do with his life. He said that the job advertised in the **magazine / newspaper / church** had looked interesting. He really wanted something that would challenge him.
5. The police still didn't know how to proceed with investigations. They thought the boy caught with the **lighter / matches / gun** was a likely suspect. The witnesses had not been very helpful at all.
6. We found out what the neighbours had been up to. The tree that had blocked the **street / road / view** had been cut down. It should make a real difference to their garden.
7. The journalist wasn't sure what he should be doing. He knew that the story about the **burglary / robbery / budget** was long overdue. But his editor would be needing the front page picture.
8. The lawyer wondered how he could construct a solid case. Obviously the document for the **building / property / judges** would be useful. He couldn't afford to let the partners down.

9. The taxi driver didn't know where he was supposed to be. Somehow the apartments with the **truck / lorry / fence** in front seemed familiar. If he didn't find his way soon he would lose the customer.
10. The secretary checked to see what had to be done next. The letter to the **client / customer / board** was on the boss' desk. All the office chores had to be finished by five o'clock.
11. The theatre critic was certain about his latest recommendation. He thought the play about the pair of **policemen / detectives / pilots** would run for months. He knew the theatre business and was usually right.
12. The air traffic controller checked that everything was running smoothly. The important plane carrying the **packages / parcels / delegates** was approaching the runway. It could be quite a stressful job.
13. The advertising executive explained how to reach the target audience. He said the poster featuring the **kitten / puppy / model** was a safe bet. He had a lot of experience in the advertising industry.
14. The ramblers thought they were getting near to the village. It seemed that the path beside the **canal / stream / forest** was going in the right direction. But without a detailed map there was no way to be certain.
15. It became clear how attitudes in the city had started to change. Recent reports of the **killings / murders / crimes** had made the community more vigilant. But a heavy police presence would still be necessary.
16. The fireman asked us how the incident had started. We pointed out the woman wearing the **sweater / jumper / scarf** who had dialled 999. They wanted to get the full story.
17. The crime squad guessed the criminal was somewhere in the local area. Soon the area behind the **pond / lake / warehouse** was completely surrounded. But he was not found and the search continued for days.
18. I couldn't decide whether I liked the new cinema layout. I hoped the seat by the **exit / door / aisle** would give me a good view. It turned out to be a wonderful evening's entertainment.
19. He asked me if I had ever had a supernatural experience. I told him about the ghost in the **graveyard / cemetery / mansion** that had scared me. I don't think he believed me.
20. The vet wondered what all the noise was about The dog with injuries to his **legs / paws / mouth** would not stop barking. The owner was getting quite embarrassed.

21. The student asked the professor for advice about the course. He said that the historical book on **rituals / ceremonies / battles** would be essential. The student needed all the advice she could get.
22. The museum owner wanted to know about the preparations for the exhibit. It turned out the box containing the **drawing / painting / vase** was still in the van. There would be terrible trouble if anything went missing.
23. The student would have to choose very carefully this year. The course on **chemistry / biology / computers** would probably have to be avoided. It was important to have a timetable with no clashes.
24. The zookeeper knew he had some cleaning to do. He had noticed that the cage for the **tigers / lions / eagles** was beginning to smell. It was a big job and would probably take all day.
25. We found out what the commotion was about. The window of the **house / flat / car** had been broken. The act of vandalism was to be discussed at the next community meeting.
26. Everyone at the book launch wondered what had caused the delay. It turned out the bag belonging to the **author / writer / reporter** had been checked thoroughly. Security at events like this was always tight.
27. The girl wondered how easy her homework would be. It was in the bag lying in front of the **couch / sofa / table** in the living room. She hated doing homework for school.
28. The sailor was enjoying being on dry land again. The equipment for his **boat / ship / mast** would take a while to fix. He had a number of friends that he was planning on visiting while he could.
29. The girl was searching all over her room for the tickets. She thought she had left the envelope inside her **closet / wardrobe / handbag** along with the present and card. If she didn't find them soon, she would be very late.
30. The editor had sighed as she pulled into the driveway. The villa which sat beside the **coast / shore / mountain** was always a welcome sight. She had been very busy for the past month and was looking forward to a relaxing weekend.
31. The brewer was always experimenting with new concoctions. The barrel with the **wheat / grain / berries** had started to ferment. He was planning on selling the drink at the local market.
32. The firemen were busy searching through the remains. The old cottage in the **woods / forest / hills** had been abandoned for years. Almost everything had been destroyed in the fire.

33. The farmer had organised his finances more carefully this year. He was already planning for the **storms / rain / droughts** that often happened late in the year. A good harvest would mean he would be debt free by the end of the season.
34. The museum had previously been considered to be very secure. The footprints on the **lawn / grass / roof** showed where the thief had entered. The sculptures had been insured but would be impossible to replace.
35. The athlete was struggling to contain all his emotions. The crowd that had gathered at the **stadium / arena / airport** was like nothing he had experienced. Despite feeling very nervous, he was expecting to enjoy the competition.
36. The two generals met in private for a frank discussion. The conditions of their **agreement / arrangement / surrender** would still have to be negotiated. It seemed obvious to both of them that all sides were hoping for a quick end to the war.
37. Take-overs of organisations are increasingly common and require careful negotiation. The chairman who the consultant had previously interviewed about the **company / business / directors** was knowledgeable, but very resistant to changes in the structure of his company. It was not clear whether the take-over would be successful.
38. Learning a new language is easier if you hear it being spoken. The student who the family had willingly accommodated during her **holiday / vacation / studies** was friendly and her English really improved during her stay. She became much better than her schoolmates.
39. It is rare to find people who are really good at motivating others to learn. The teacher who the child had really admired after the **lesson / lecture / game** was talented, because she could explain very technical ideas in a simple way. This had a good effect on her students.
40. Not considering other people and vehicles when playing in the road can be dangerous. The policeman who the bicyclist had disobeyed on the **street / road / pavement** was friendly and only issued a warning instead of a fine. The bicyclist was fortunate that the punishment was not worse.
41. The quality of teaching at the college was legendary. The advisor who the students have always appreciated for her **kindness / compassion / humour** is excited because she recently won a teaching award. She was not the first at the college to achieve such recognition.
42. Working for Childline can be very rewarding work. The counselor who the teenager had previously called on the **helpline / telephone / mobile** was

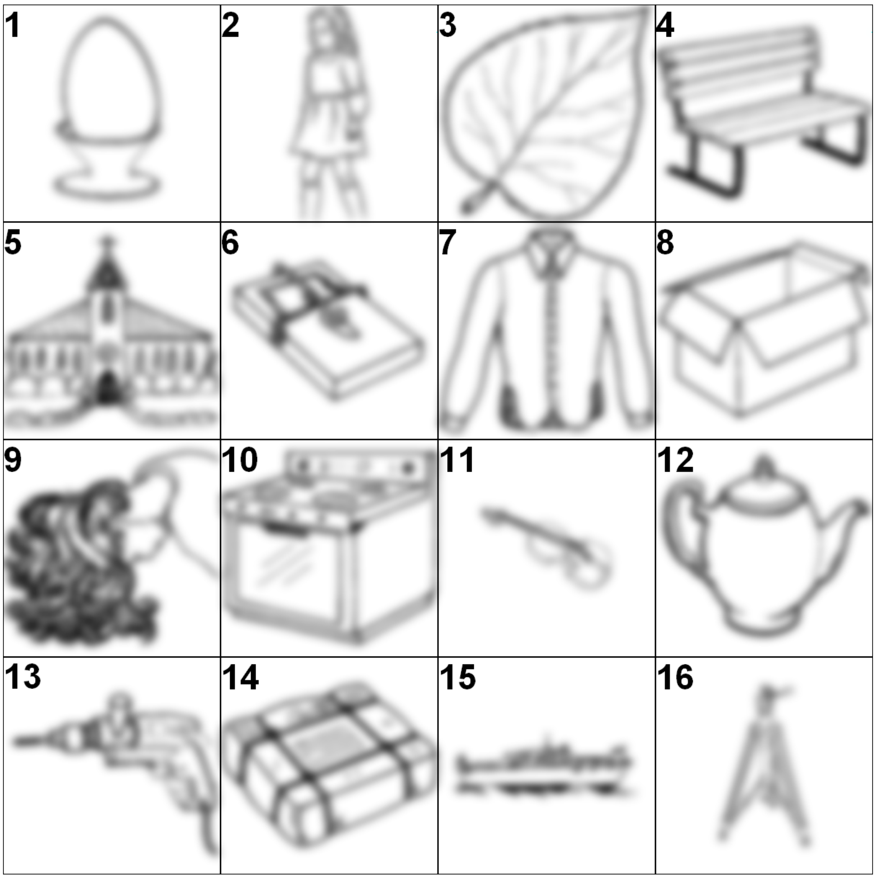
helpful since she really cared about his problems. She has always wanted to make a difference for people worse-off.

43. Growing old generally means an increase in dependency on others. The neighbor who the Girl Guide had regularly bought groceries for at the **shop / stores / market** was old and sick and needed help making her dinner. Her life would be much harder without this help.
44. Sometimes people have a great time when they expect not to. The visitor who the host had belatedly invited to the **disco / dance / concert** was shy but ended up having a fantastic time. Everyone else made them feel very welcome.
45. Getting used to going to nursery school can be difficult. The child who the play leader had repeatedly comforted in the **playground / schoolyard / classroom** eventually settled down and played in the sandpit. In time the child came to enjoy nursery school.

Appendix B

Names of images used in Experiment 2

Where only one name is given these were *disfluency* images (with hard-to-name in bold). Where two names are given, these are *alignment* images (preferred/dispreferred, with the name used by the confederate in italics).



1 Egg

2 *Girl*/Child

3 Leaf

4 Bench/*Seat*

5 Church/*Cathedral*

6 **Mousetrap**

7 Shirt/*Blouse*

8 Box

9 Hair

10 Cooker/*Oven*

11 Violin/*Viola*

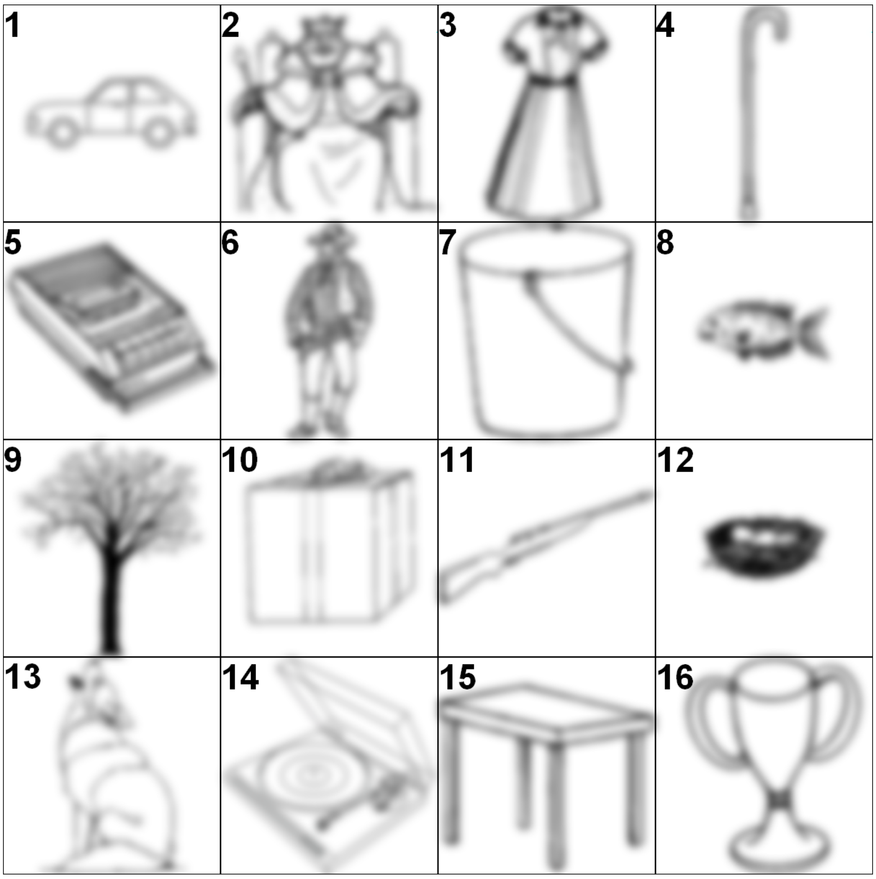
12 **Teapot**

13 **Drill**

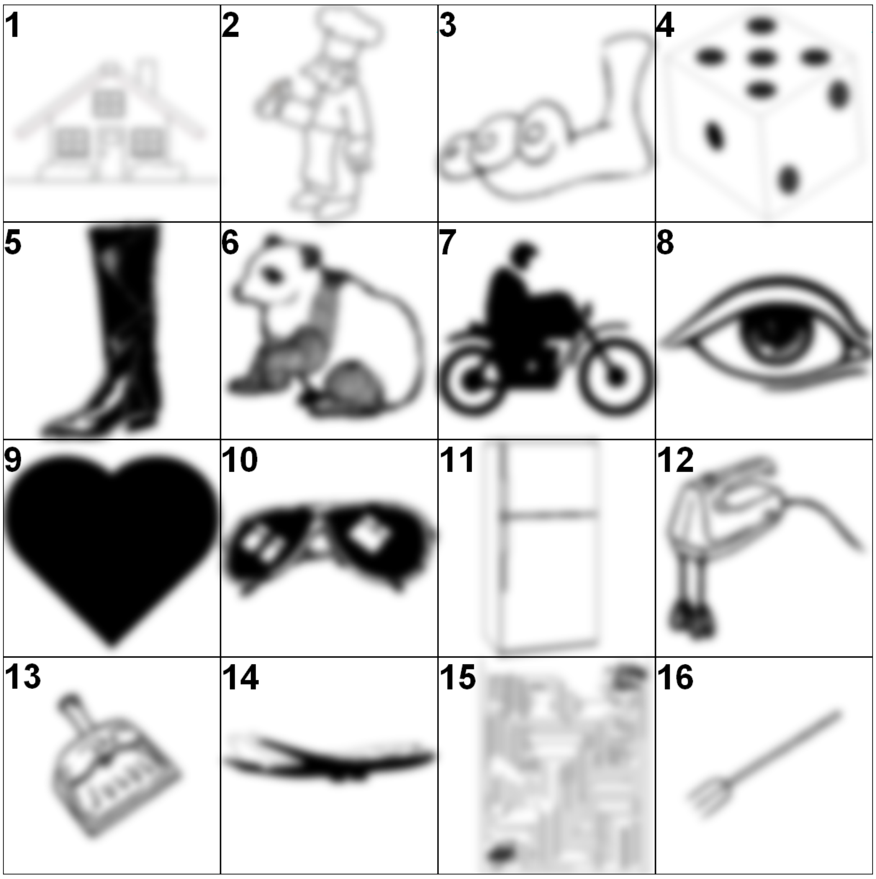
14 *Parcel*/Package

15 *Boat*/Yacht

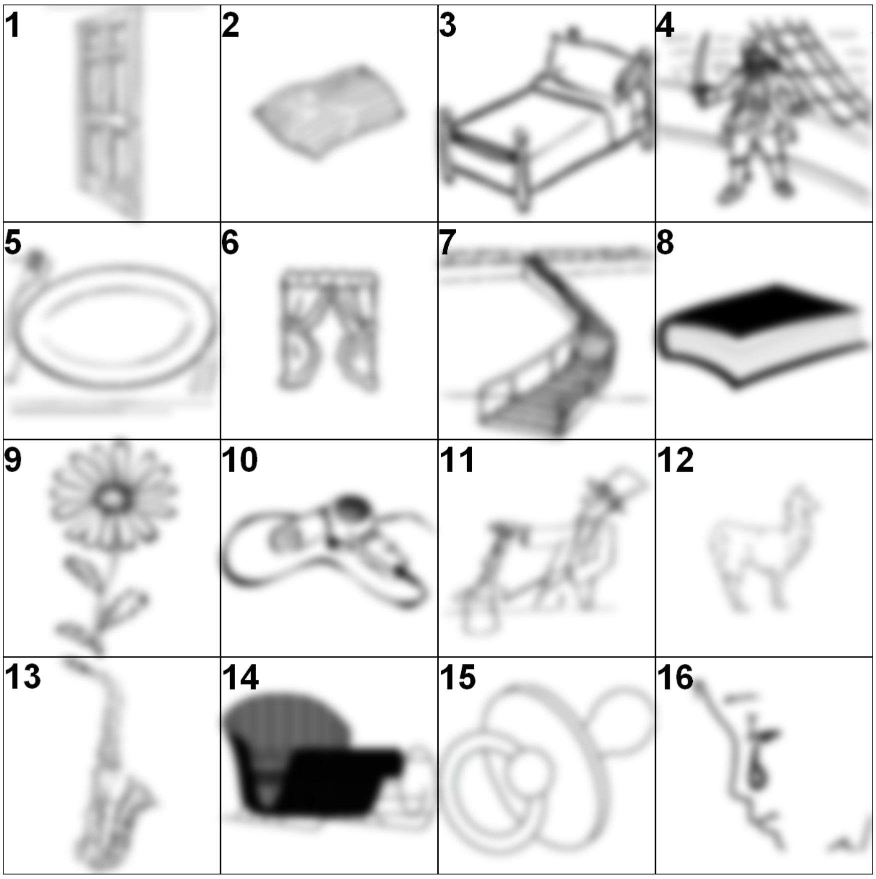
16 **Trypod**



1 Car
2 *King*/Sovereign
3 Dress
4 Walking stick/*Cane*
5 Tape recorder/*Cassette Recorder*
6 **Cowboy**
7 Bucket/*Pail*
8 Fish
9 Tree
10 Present/*Gift*
11 Rifle/*Gun*
12 **Nest**
13 **Wolf**
14 *Turntable*/Record Player
15 *Table*/Desk
16 **Trophy**



1 House
2 *Chef*/Cook
3 Foot
4 Dice/*Die*
5 Boot/*Shoe*
6 **Panda**
7 Motorbike/*Motorcycle*
8 Eye
9 Heart
10 Sunglasses/*Shades*
11 Fridge/*Refrigerator*
12 **Mixer**
13 **Dustpan**
14 *Aeroplane*/Aircraft
15 *Maze*/Labyrinth
16 **Pitchfork**



1 Door
2 *Pillow*/Cushion
3 Bed
4 Pirate/*Sailor*

5 Plate/*Dish*
6 **Curtains**
7 Stairs/*Staircase*
8 Book

9 Flower
10 *Hairdryer*/Oven
11 Magician/
Conjurer
12 **Llama**

13 **Saxophone**
14 *Sleigh*/Sled
15 *Dummy*/Pacifier
16 **Tear**

Appendix C

Finlayson, Lickley, and Corley (2010)

Proceedings of DiSS-LPSS Joint Workshop 2010,
25-26 September 2010, Tokyo, Japan, pp. 119-122

The influence of articulation rate, and the disfluency of others, on one's own speech

Ian R. Finlayson[1], *Robin J. Lickley*[1], *Martin Corley*[2]

[1]Speech Science Research Centre, Queen Margaret University, UK
[2]Department of Psychology, University of Edinburgh, UK
ifinlayson@qmu.ac.uk

Abstract

Disfluencies are a regular feature of spontaneous speech, and much has been learnt about the effects of various linguistic factors on their production. Speech usually occurs within dialogue, yet little is known about the influence of an interlocutor's speech on a speaker's own fluency. It has been shown that speakers tend to align on various levels, converging, for example, on lexical, and syntactic levels. But we know little about convergence in rate of speech or disfluency. Little is also known about the effects of speech rate on fluency in a speaker's own speech. In this paper, we examine these effects through analysis of speech rate, hesitation and error correction in a corpus of task-oriented dialogues (the HCRC Map Task Corpus). Our findings demonstrate that different types of disfluencies can be influenced in different ways by speech rate. Furthermore, the probability of an interlocutor being disfluent appears to affect the speaker's own likelihood, raising the possibility that interlocutors may "align" on disfluent, as well as fluent, speech.

Index Terms: articulation rate, alignment, accommodation theory, dialogue

1. Introduction

Spontaneous speech is rarely fully fluent. Natural dialogue is peppered with disfluencies such as repairs, repetitions and fillers (e.g. *uh* and *um*). While much has been learned about various linguistic factors which influence the production of disfluencies (e.g. [1, 2]), relatively little attention has been paid to how, on the one hand, rate of speech and the speech and, on the other, the speech of the interlocutor may affect fluency. This study explores these factors and their interaction.

Common in many theories of language production (e.g. [3]) is that the process involves a series of stages. In conceptualisation, a general plan for the utterance is formed; in formulation, the lexical forms are selected and ordered, in articulation, motor commands are sent to the articulatory mechanism. A delay at any of these stages, whether through indecision or error and repair, may result in a need for the speaker to pause, as the rate of motor execution exceeds the rate of planning. Various strategies are possible: The speaker could simply elect to remain silent until the next chunk of speech is ready (though this seems to be the least preferred option); they could maintain some form of vocalization, either by prolonging the previous sound (if the pause is mid-utterance), or by producing a filler (e.g., um/uh); or they may elect to restart the current phrase, thus producing a repetition (cf. [4]).

If restarts are the consequence of the articulator executing one plan before it is able to formulate the next then in the case of fast speech, where we may expect to see a greater incidence of the articulator "out-pacing" the components upon which it draws from, speakers may be more likely to produce repetitions. Repetitions may not, however, be the only strategy available, and we may expect to see other forms of disfluency produced as a consequence of the articulator stalling for time while it awaits new input. A possible alternative would be to produce a filler, a superficially meaningless sound which could occupy the delay.

Where articulation rate has been indirectly manipulated it has been shown that faster speech is more likely to contain repetitions [5], but not likely to contain more fillers. Participants were asked to describe the path taken by a dot around a network which contained images of everyday objects. In half of these trials, the speed of the dot's movement was increased and analysis showed a corresponding significant increase in the articulation rate. This faster condition also saw a higher frequency of errors and error repairs.

It should perhaps come as no surprise that when forced to accelerate our speech, which the dot's faster movements require, participants made more errors. However within this paradigm it is not clear where exactly causality lies. Do participants make more repairs, and repetitions, because they are speaking faster? Or is this increase in errors a consequence of the time pressures caused by having to keep up with the speed of the dot's movements?

In the absence of pressure we may expect there to be no difference in the volume of disfluencies produced by naturally faster and slower speakers. Why, after all, would a person consistently speak at a rate which impaired their ability to maintain fluency? The map task corpus [6] affords us the chance to explore the effects of a naturally occuring range of articulation rates on the production of disfluencies without the confound of having to place pressure upon speakers to produce differences.

The conversational nature of the map task also allows us to explore what effects, if any, an interlocutor may have on both articulation rate and disfluencies. Possible interactions between the speech rates of interlocutors have already received considerable theoretical attention [7]. Accommodation theory suggests that speakers' speech rates may be sensitive to those of their interlocutors: when speaking to an interlocutor who speaks slower, a speaker may tend to reduce their own speech rate in order to converge with their partner.

Recent years have seen a change in our ideas about communication, with the gap between production and comprehension continuing to narrow [8]. While it has been suggested that production and comprehension rely upon one another in order to ensure the alignment of representations between interlocutors in dialogue necessary for successful communication, little consideration has been given to the effects upon production of the disfluent speech that speakers regularly hear from their interlocutors.

In the present study we will investigate the influence of speaker's articulation rate upon their likelihood of producing substitutions, repetitions and fillers, before exploring if interlocutors show a tendency to align on articulation rate and their production of disfluencies.

2. Method

2.1. Corpus

Materials came from the HCRC Map Task corpus [6]; transcribed and annotated dialogues recorded between 64 University of Glasgow students (32 male, 32 female) taking part in a cooperative task. Participants took turns to direct their partner along a path which used labelled images of objects as landmarks. Each participant had their own map which their partner could not see. Participants were split up into quads, each consisting of two pairs of friends (although each pair was unfamiliar to the other). In addition, half of all quads performed the task with a screen preventing them from seeing each other at all. All coding has been converted to XML which can be queried using the NITE XML Toolkit [9].

2.2. Unit of analysis

The markup of the corpus is segmented into individual units, known as timed units, which correspond to individual words or silences. Each timed unit has a corresponding unit for its part of speech, one of which being fillers, which include: *eh*, *ehm*, *er*, *erm*, *uh* and *um*.

For disfluencies which did not correspond to individual units we used Lickley's [10] taxonomy of editing disfluencies, as this forms the basis for the disfluency coding which appears in the corpus. Editing disfluencies come in four forms but we will only focus on two of these, with examples of both given in (1): (1a) substitutions, which correspond to what are commonly called repairs; and (1b) repetitions, where one or more words are repeated immediately following their first mention. The structure of each disfluency follows that set out by Levelt [11], with each containing a reparandum, an interruption point and a repair.

(1a) I don't suppose you've got [the balloons] the baboons

(1b) Right [there's a] there's a line about quarter of the way down

All timed units were extracted from the corpus, with the exception of those silences where the participant was listening to their partner, and non-vocal noises, creating a data set of 174,049 units. Each of these units were then coded for whether or not they were disfluent (either a filler, or a reparandum word). Variables were subsequently created for individual types of disfluency (fillers, substitutions, repetitions, insertions and deletions) and each unit was coded appropriately.

2.3. Analysis

As our dependent variables of interest in the present study were binomial (whether or not a unit was disfluent), likelihood was modelled using logit mixed-effects models [12, 13]. Logit mixed-effects models provide an advantage over ordinary logistic regression in that they allow us to include random effects in models, removing the need for separate by-participants and by-items analyses. All analyses were performed in R [14], using the lme4 package [15].

Prior to analysing the effect of articulation rate and interlocutor disfluency on our dependant variables, "control models" were constructed containing participants and the map being described as random effects. Both linguistic and non-linguistic factors were tested for inclusion in these models. In order to control for the finding that disfluencies are associated with the production of longer utterances, a measure of utterance length was required. As it is difficult to get a measure of utterance length in a unstructured dialogue (see [16] for a discussion of these difficulties), the length of each move was used as an approximate analog.

As each individual word of a reparandum was considered, it was possible that speakers who tended to produce longer reparanda may produce spurious results. To control for this, the mean lengths of each speaker's reparanda within each conversation was included. In order to control for the fact that the more a participant said the greater opportunity they had to produce disfluencies, the word count for each conversation was first added. This figure included reparandum words and fillers, in addition to fluent words, however word fragments were excluded. Both word count and mean reparanda length were converted to z-scores.

The following non-linguistic factors were subsequently tested for inclusion in the model: speaker's role (giver, or follower, of instructions), interlocutors' ability to make eye-contact, interlocutors' familiarity, and finally speaker's gender.

These control models were then compared with "full models" containing potential predictors of interest: speaker's articulation rate, partner's articulation rate and the probability of the partner producing each type of disfluency. Articulation rate was measured in syllables per second. The number of syllables in the 100 most used words in the corpus (accounting for 75% of all speech) were counted. Considering only these 100 words, the total number of syllables produced by each speaker in each conversation, was divided by the summed duration of each of these words. Finally, probabilities were obtained by dividing the total number of each disfluency a speaker produced by the total number of words they produced in each conversation.

Once all predictors had been tested, the Wald statistic was used to assess if the estimated slope for each predictor was significantly different from zero. Where the addition of a predictor was found to improve the fit of a model its coefficient will be reported alongside the odds ratio (OR; e^{β}).

3. Results

3.1. Articulation Rate and Disfluency

All control models constructed include utterance length, mean reparandum length and word count where they significantly improve model fit. Where other predictors were included they are reported (see [16] for a full account of the construction of these models).

3.1.1. Substitutions

In order to examine the likelihood of words appearing in the reparanda of substitutions a control model was first constructed, which included the speakers role, giving log-likelihood -8390.8. The addition of the speaker's articulation rate did not improve upon the fit of the model, with a log-likelihood of -8390.8 ($\chi^2(1) = 0.012, p < 1$), suggesting that articulation rate has no effect on speaker's substitutions.

3.1.2. Repetitions

A control model was constructed containing interlocutors' ability to make eye-contact and an interaction between gender and role, with a log-likelihood of -13484. This model was found to be improved by the addition of articulation rate, giving log-likelihood -13481 ($\chi^2(1) = 4.8296, p < .05$), suggesting that faster speakers were more likely to produce repetitions ($\beta = 0.130, p < .05; \mathrm{OR} = 1.138$).

3.1.3. Fillers

In order to model the probability of speakers producing a filler, a control model was constructed which included speaker's role, and provided a log-likelihood of -9578.9. The addition of articulation rate significantly improved the fit of the model, giving a log-likelihood of -9575.0 ($\chi^2(1) = 7.778, p < .01$). This finding suggests that faster speakers are less likely to produce fillers ($\beta = -0.180, p < .01; \mathrm{OR} = 0.835$).

3.2. Alignment of Articulation Rate

To assess the effect of interlocutor's articulation rate we modelled speaker's articulation rate with linear mixed-effects models, with Markov chain Monte Carlo sampling over 10,000 simulations to estimate coefficients. A control model was constructed containing word count, speaker's role, familiarity, speaker and partner's genders; with log-likelihood of -215.83. The addition of the interlocutor's articulation rate was found to improve the fit of the model, with log-likelihood -213.22 ($\chi^2(1) = 5.216, p < .05$), suggesting that talking to faster speakers increases one's own articulation rate. However, we must be cautious in interpreting this finding as role may be confounded with utterance length.

Additionally, our model was further improved by the inclusion of an interaction between speaker's role and the articulation rate of their partner, with log-likelihood -211.02 ($\chi^2(1) = 4.420, p < .05$). This suggests that while all speakers are sensitive to their interlocutor's articulation rate, the effect is increased for instruction followers.

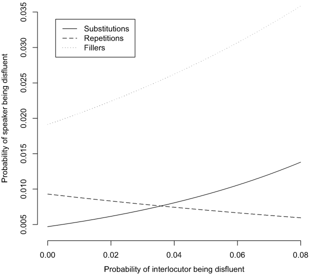
3.3. Alignment of Disfluency

The alignment of articulation rates suggest that aspects of a speaker's production appears sensitive to corresponding aspects of their interlocutors' speech, and we may rightly question if disfluencies behave similarly. However, as articulation rate may influence certain forms of disfluency, we will test the effect of interlocutor's articulation rate on our models before considering their own likelihood of being disfluent. For all types of disfluency, the best fitting models constructed in section 3.1 were used as controls. Estimated probabilities for each type of disfluency are shown in Figure 1.

3.3.1. Substitutions

The addition of interlocutor's articulation rate was not found to improve model fit, giving a log-likelihood of -8390.5 ($\chi^2(1) = 0.579, p < 1$); however including the probability of an interlocutor producing a substitution did significantly improve the model, with log-likelihood -8386.2 ($\chi^2(1) = 9.230, p < .1$). This suggests that speakers are more likely to produce substitutions when speaking to people who are themselves more likely to produce substitutions ($\beta = 13.597, p < .01; \mathrm{OR} = 803437$).

Figure 1: Relationship between speaker's and interlocutor's likelihood of production substitutions, repetitions and fillers. Where models include word count, mean reparandum length or articulation rates, mean values were used.



3.3.2. Repetitions

Similarly to substitutions, interlocutor's articulation rate was not found to improve our model of likelihood of producing repetitions, log-likelihood -13481 ($\chi^2(1) = 0.996, p < 1$). However, speakers were found to be less likely to repeat words when their interlocutor had a greater tendency to repeat ($\beta = -5.631, p < .05; \mathrm{OR} = 0.004$), with the addition of interlocutor probability improving model fit, giving a log-likelihood of -13478 ($\chi^2(1) = 5.809, p < .05$).

3.3.3. Fillers

Including interlocutor's articulation rate was found to improve upon the fit of our control model, giving a log-likelihood of -9571.3 ($\chi^2(1) = 7.280, p < .01$), suggesting that participants were less likely to produce fillers when talking to faster speakers ($\beta = -0.123, p < .01; \mathrm{OR} = 0.884$). Furthermore, the probability of a participant producing a filler increased when their partner was more likely to ($\beta = 8.073, p < .05; \mathrm{OR} = 3207.484$), log-likelihood -9568.7 ($\chi^2(1) = 5.294, p < .05$).

4. Discussion

Our results suggest that different types of disfluencies may be influenced in different ways by articulation rate, and the likelihood of one's interlocutor being disfluent:

- Faster speakers were more likely to produce repetitions; however they were less likely to produce fillers, and no effect was found with substitutions.
- Participants' articulation rate increased with faster speaking partners, particularly when the participant had the role of follower.
- Participants tended to be more likely to produce substitutions and fillers with interlocutors who were themselves

more likely to produce these two types of disfluency, while they produced fewer repetitions when speaking to a partners who were more prone to repeating.

Our finding that faster speakers were more likely to produce repetitions appears consistent with Blackmer & Mitton's [4] idea of an articulator possessing an autonomous restart capability. Increasing the rate of articulation may increase the likelihood of the articulator becoming "out of sync" with the components which precede it, and the prediction that in these situations speakers will tend to repeat the most recent plan appears to be borne out. While speakers may choose to stall through repetition, our finding that faster speakers are, in fact, less likely to produce fillers suggest that they are not used as a similar stalling mechanism.

While it appears that participants' articulation rate is variable, particularly in response to the articulation rate of their partner, it does not seem that they allow themselves to speak at such an accelerated rate that they are more prone to make the sorts of errors which require substitutions to resolve. This suggests that Oomen & Postma's [5] finding that errors occurred more frequently when a speaker was under time pressure may be due solely to the greater demand of having to keep up with a faster moving dot in order to describe its path, rather than being a consequence of speaking too quickly.

The alignment of articulation rates between partners provides evidence for the convergence of linguistic factors suggested by accommodation theory. However deeper understanding of this phenomena requires closer examination of variations in articulation rate across not only conversations, but also across conversational turns, which is beyond the scope of the present study.

It is not clear how the disfluencies of one's interlocutor may influence our own likelihood of being disfluent, with converging upon disfluencies seeming unlikely to bring a benefit. One possibility may be that hearing a conversational partner produce a greater number of errors makes one more comfortable in being error prone, reducing the care one takes in being fluent. This account would hold despite the same pattern not emerging with repetitions, as repetitions may reflect a different type of problem to those solved with substitutions. However it would offer no explanation for why the opposite pattern was found when we observed the effects of repetitions.

Similarly, an explanation is lacking for why one is less likely to produce fillers when speaking to a faster partner, even after one's own articulation rate has been taken into account. If fillers are used as signals for turn-taking, as has been suggested (e.g. [17, 18]), then we may speculate that engaging in a dialogue with someone who often uses fillers may lead one to employ them more often themselves in order to manage their side of the conversation, which may explain the link observed between our participant's likelihood of using fillers and their interlocutor's.

Future research may explore further how both linguistic and non-linguistic features of dialogues may influence speaker's likelihood to be disfluent, and how these factors may interact.

5. Acknowledgements

We thank Jean Carletta and Jonathan Kilgour for technical help.

6. References

[1] G. W. Beattie and B. Butterworth, "Contextual probability and word frequency as determinants of pauses in spontaneous speech," *Language and Speech*, vol. 22, pp. 201–211, 1979.

[2] H. Bortfeld, S. D. Leon, J. E. Bloom, M. F. Schober, and S. E. Brennan, "Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender," *Language and Speech*, vol. 44, no. 2, p. 123, 2001.

[3] W. J. M. Levelt, A. Roelofs, and A. S. Meyer, "A theory of lexical access in speech production," *Behavioral and brain sciences*, vol. 22, no. 01, pp. 1–38, 1999.

[4] E. R. Blackmer and J. L. Mitton, "Theories of monitoring and the timing of repairs in spontaneous speech," *Cognition*, vol. 39, no. 3, pp. 173–194, 1991.

[5] C. C. E. Oomen and A. Postma, "Effects of time pressure on mechanisms of speech production and self-monitoring," *Journal of Psycholinguistic Research*, vol. 30, pp. 163–184, 2001.

[6] A. Anderson, M. Bader, E. Bard, E. Boyle, G. M. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert, "The HCRC map task corpus," *Language and Speech*, vol. 34, pp. 351–366, 1991. [Online]. Available: http://www.hcrc.ed.ac.uk/maptask/

[7] H. Giles and P. Smith, *Language and Social Psychology*. Oxford: Basil Blackwell, 1979, ch. Accommodation Theory: Optimal Levels of Convergence.

[8] M. J. Pickering and S. Garrod, "Do people use language production to make predictions during comprehension?" *Trends in Cognitive Sciences*, vol. 11, no. 3, pp. 105–110, 2007.

[9] J. Carletta, S. Evert, U. Heid, and J. Kilgour, "The NITE XML Toolkit: Data model and query language," *Language Resources and Evaluation*, vol. 39, pp. 313–334, 2006. [Online]. Available: http://groups.inf.ed.ac.uk/nxt/index.shtml

[10] R. J. Lickley, "HCRC Disfluency Coding Manual," HCRC, Tech. Rep. 100, 1998. [Online]. Available: http://www.ling.ed.ac.uk/ robin/maptask/disfluency-coding.html

[11] W. J. M. Levelt, "Monitoring and self-repair in speech," *Cognition*, vol. 14, no. 1, pp. 41–104, 1983.

[12] N. E. Breslow and D. G. Clayton, "Approximate inference in generalized linear mixed models," *Journal of the American Statistical Society*, vol. 88, pp. 9–25, 1993.

[13] S. DebRoy and D. M. Bates, "Linear mixed models and penalized least squares," *Journal of Multivariate Analysis*, vol. 91, pp. 1–17, 2004.

[14] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008. [Online]. Available: http://www.R-project.org

[15] D. Bates, M. Maechler, and B. Dai, *lme4: Linear mixed-effects models using S4 classes*, 2009, r package version 0.999375-32. [Online]. Available: http://CRAN.R-project.org/package=lme4

[16] I. R. Finlayson, M. Corley, and R. J. Lickley, "Why use disfluency? The effect of situational factors on the production of disfluent speech in dialogue," Submitted.

[17] H. H. Clark and J. E. Fox Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, pp. 73–111, 2002.

[18] H. Maclay and C. E. Osgood, "Hesitation phenomena in spontaneous speech," *Word*, vol. 15, pp. 19–44, 1959.

Appendix D

Finlayson and Corley (2012)

Psychon Bull Rev
DOI 10.3758/s13423-012-0279-x

BRIEF REPORT

Disfluency in dialogue: an intentional signal from the speaker?

Ian R. Finlayson · Martin Corley

Abstract Disfluency is a characteristic feature of spontaneous human speech, commonly seen as a consequence of problems with production. However, the question remains open as to *why* speakers are disfluent: Is it a mechanical by-product of planning difficulty, or do speakers use disfluency in dialogue to manage listeners' expectations? To address this question, we present two experiments investigating the production of disfluency in monologue and dialogue situations. Dialogue affected the linguistic choices made by participants, who aligned on referring expressions by choosing less frequent names for ambiguous images where those names had previously been mentioned. However, participants were no more disfluent in dialogue than in monologue situations, and the distribution of types of disfluency used remained constant. Our evidence rules out at least a straightforward interpretation of the view that disfluencies are an intentional signal in dialogue.

Keywords Speech production · Social cognition

Around six per hundred spoken words are affected by *disfluencies*, including fillers such as *uh* and *um*, prolongations of both open and closed class words, repairs, and whole- or part-word repetitions (Bortfeld, Leon, Bloom, Schober, & Brennan, 2001; Fox Tree, 1995). Such disfluencies tend to occur when the topic of the speech is unfamiliar (Bortfeld et al., 2001; Merlo & Mansur, 2004) or is associated with a larger vocabulary (Schachter, Christenfeld, Ravina, & Bilous, 1991). They are often found at the beginnings of longer phrases (Oviatt, 1995; Shriberg, 1996) and before words with low contextual probability (Beattie & Butterworth, 1979).

These findings suggest that disfluencies reflect the difficulty that the speaker is having in retrieving the appropriate words to say. Open to question, however, is the issue of *why* difficulties in speech planning result in disfluency, rather than in some other accommodation. One possibility is that a disfluency is a mechanical by-product of the difficulty itself (e.g., Blackmer & Mitton, 1991). Alternatively, disfluencies may be used to communicate to the listener that the speaker is in difficulty (Clark & Fox Tree, 2002). Given that speech occurs most often in the form of dialogue, the resolution of this question is important in exploring the ways in which interlocutors communicate with each other. In the present article, we address the issue with two experiments that compare the situational effects of dialogue versus monologue on the production of disfluencies and of words.

According to Clark and Fox Tree (2002), speakers utter particular disfluencies in order to inform the listener, for example, about the length of an anticipated interruption to speech (Clark & Fox Tree, 2002; Fox Tree & Clark, 1997). In line with this view, investigations based on corpora of transcribed speech show that *thee* is followed by silence more often than *thuh* (Fox Tree & Clark, 1997) and that longer silences follow *um* than *uh* (Clark & Fox Tree, 2002), consistent with earlier speech comprehension findings that suggest that *uh* and *um* have different effects on listeners (Fox Tree, 2001). Although this view has been challenged (O'Connell & Kowal, 2005), evidence from recorded speech that is consistent with Clark and Fox Tree's findings has been reported elsewhere (e.g., Barr, 2001; Fox Tree, 2001).

I. R. Finlayson
CASL, Queen Margaret University,
Musselburgh, Scotland, UK

M. Corley (✉)
Psychology, Philosophy, Psychology and Language Sciences,
University of Edinburgh,
7 George Square,
Edinburgh EH8 9JZ, UK
e-mail: Martin.Corley@ed.ac.uk

Published online: 13 June 2012

Further support for the view that disfluencies are used communicatively appears to come from a study of patterns of disfluency in the speech of adults with autistic spectrum disorders (Lake, Humphreys, & Cardy, 2011). Lake et al. suggested that speakers with autism would be less likely to produce disfluencies that were specifically listener oriented. Accordingly, participants with autism produced fewer fillers than did matched controls but appeared to trade these off against disfluent repetitions and silent pauses. It should be noted that the reported findings are not directly compatible with Clark and Wasow's (1998) suggestion that fillers and repetitions serve functionally similar communicative purposes; nor do they match evidence showing that listeners are similarly affected by *uh* and silent pauses, (Corley, MacGregor, & Donaldson, 2007; MacGregor, Corley, & Donaldson, 2010), but not by repetitions (MacGregor, Corley, & Donaldson, 2009). But the study supports a general suggestion that different disfluencies may be produced for different reasons.

However, the facts that fillers tend to precede silence or that different people produce different patterns of disfluency do not lead to the conclusion that disfluencies are intentionally chosen to serve as signals to the listener, any more than the smoke that accompanies fire (or not) is "chosen." Moreover, although disfluencies affect listeners, both immediately and in the longer term (e.g., Arnold, Tanenhaus, Altmann, & Fagnano, 2004; Corley et al., 2007; Fox Tree, 2001; Swerts & Krahmer, 2005), one cannot conclude from this that speakers use them to communicate, any more than the fact that a hand is withdrawn from the flame proves that the fire uses pain to affect behavior. Although evidence is consistent with the view that disfluencies are uttered with communicative intent, it remains possible that they are simply a consequence of delays to the speech plan, co-occurring with them automatically in ways that listeners can stochastically exploit.

In contrast to disfluencies, there is little room for doubt that the words that constitute an utterance (and convey its primary message) are chosen by the speaker. According to Pickering and Garrod's (2004) *interactive alignment* model, alignment at all levels of dialogue (from the choice of individual words to that of syntactic structure) is at the root of successful communication. Because alignment is fundamental, "the production of a word or utterance in dialogue is only distantly related to the production of a word or utterance in isolation" (Pickering & Garrod, 2004, p. 183). Speakers in dialogue are highly likely to refer to things using the same words that their interlocutors have just used.

Whereas word choice in dialogue is well understood, there has to date been no direct experimental investigation of the role that disfluency plays in dialogue. In this article, we present a study designed to investigate whether disfluencies are used communicatively or whether they are an automatic consequence of difficulty in the formulation of speakers' utterances, by comparing the production of disfluencies across monologue and dialogue situations. By manipulating the ease with which pictures can be named (see Hartsuiker & Notebaert, 2010; Schnadt & Corley, 2006) in a card-sorting task, we ensure that there will be difficulties in lexical selection: Of interest is whether these difficulties automatically result in disfluency or whether disfluencies are found only in dialogues, where they would be informative to the listener.

The monologue/dialogue manipulation is similar to that used by Bavelas, Gerwing, Sutton, and Prevost (2008) in their investigation of the production of nonverbal gestures. In that study, face-to-face dialogues were compared with telephone dialogues and monologue production. While gestures were produced in all three settings, they occurred in greater frequency in the two dialogue conditions than in the monologue condition. If we assume that disfluencies serve a communicative purpose, then, as for gestures, we may reasonably expect fewer disfluencies to be produced in monologue.

To show that participants in the present study are affected by the monologue/dialogue manipulation, a subset of the pictures used have more than one name. By manipulating the name that one (confederate) party in the dialogue has just used for each of these pictures, we should be able to show that the participants align in dialogue, by tending to choose the same names. This manipulation serves as a demonstration that, in common with other confederate-dialogue tasks (e.g., Cleland & Pickering, 2003), the participants are sensitive to their interlocutors and their word choices are governed by the principles of alignment. If word choice is affected by the presence of an interlocutor but the production of disfluency is not, it will be harder to argue that disfluency is produced with communicative intent.

Experiment 1

Participants were asked to perform two tasks. In one, they were provided with grids containing pictures of objects and were instructed to name them in sequence (monologue condition). In the other, they used similar grids to play a picture-matching task with a confederate of the experimenter (dialogue condition). In each condition, half of the images the participant named were *disfluency images*, used to establish how disfluent the speaker was, and half were *alignment images*, used to measure alignment. Disfluency images were selected on the basis of the difficulty with which they could be named. Other things being equal, images that were difficult to name were expected to elicit more disfluencies than were those that were easy. Alignment images each corresponded to pairs of names that were used either frequently (preferred) or infrequently (dispreferred) in pretests.

We predicted that participants would be unlikely to use the dispreferred names, except in cases where they had previously been used by the confederate.

Method

Participants Twenty native British-English-speaking undergraduate students from the University of Edinburgh volunteered to take part in the experiment.

Materials Images were chosen from the International Picture Naming Project (IPNP: Szekely et al., 2004), which provides information about the naming of 520 black-and-white line drawings of common objects, some of which are freely downloadable. Where images could not be obtained directly from the IPNP, suitable images were selected from a commercial clip art package.

Thirty-two disfluency images were classified as either difficult or easy (16 of each), on the basis of the findings of Schnadt and Corley (2006). Difficult images had low codability (they corresponded to several possible names), with *H* values (Snodgrass & Vanderwart, 1980) of above 0.85 ($M = 1.60$, $SD = 0.39$) in the IPNP; CELEX frequencies (Baayen, Piepenbrock, & van Rijn, 1993) of the dominant names were kept below 25 counts per million (cpm: $M = 4.00$, $SD = 4.75$). Easy images had high codability and high frequency, with *H* below 0.15 ($M = 0.06$, $SD = 0.07$) and CELEX frequencies of the dominant names above 75 cpm ($M = 255$, $SD = 167$). Example images are given in Fig. 1.

Ten raters were asked to name each of an additional 40 images and to rate alternative image names for appropriateness. The alternative names were infrequently used names for each image taken from the Beckman Spoken Picture Naming Norms (Griffin & Huitema, 1999). Eight images were discarded because the most common name was used by fewer than 80 % of the participants or the selected alternative name had a mean appropriateness rating of less than 2.5 out of 5. The remaining 32 images constituted the alignment images, each associated with a commonly used (*preferred*) name and an alternative (*dispreferred*) name (see Fig. 1).

Finally, 32 filler images were selected. These were not subject to any constraint other than that they would be easily recognized as depicting objects named by the confederate.

Four 4×4 grids were created, and the images were randomly assigned to each, with the constraint that each grid included eight disfluency images and eight alignment images. An additional four grids containing printed names in lieu of images (and therefore, serving as scripts) were created for use by the confederate. Eight of the names corresponded to alignment items on the relevant picture grid (five were dispreferred names, to increase the opportunity for alignment). In lieu of the disfluency items, each of the confederate's grids included the names of eight filler items. For the matching tasks, participants and the confederate were each given four blank 4×4 grids on which to arrange cards depicting the images named by their interlocutors. All grids were numbered 1–16, starting in the top left corner.

Procedure In order to prevent participants from realizing that their performance in monologue and dialogue would later be compared, a cover story was created that they would be performing two separate experiments for two different experimenters (only one of whom was able to be present). To reinforce this, they were given two different instruction sheets and signed two different consent forms. When performing in the monologue condition, participants were told that the researcher needed recordings of phonemes obtained from arbitrary natural speech for use in a further

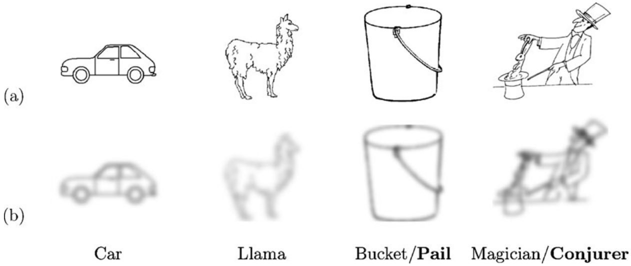


Fig. 1 Examples of easy-to-name (car) and a hard-to-name (llama) images and two alignment images (for each image, the preferred name is given, followed by the dispreferred name used by the confederate in bold), as used in **a** Experiment 1 and **b** Experiment 2

project. These instructions were designed to minimize the communicative aspect of the task. In the dialogue condition, participants were told that they were involved in a study investigating the ways in which people work together to perform a task. The order of conditions was counterbalanced across participants, and upon completion of both, they were informed as to the true nature of the experiment.

Each of the four grids was used equally often in the monologue and dialogue conditions. In the monologue condition, participants were shown each of two grids in turn and were asked to name the pictures in sequence. In order to imitate spontaneous speech, it was suggested that participants name each image in a sentence, although no guidance was given about the structure of the sentence. If participants asked, they were simply instructed to ensure that they stated the number of the square and its contents.

In the dialogue condition, the confederate acted like a second naïve participant. The experimental participant was introduced to the confederate, and both were seated at a table with a partition separating them. This prevented the participant and confederate from seeing each other or the other's grids but did not restrict them from hearing each other. Both were given grids and were instructed that they should take turns to name in sequence each item and its position in their grids and were provided with an example of what they might say: "In box one I have a dog." Upon hearing the partner naming an image in the grid, each had to place the matching individual image on to the appropriate square of a blank grid. The confederate always went first, reading from the appropriate "script" grid. This ensured that the participant heard the preferred or dispreferred name for a given item before it was his or her turn to name the relevant picture. However the confederate never produced a "name" immediately before the participant named the same image, ensuring that the participant could not simply echo what the confederate said at any stage of the experiment. Once all of the images in a grid had been named, the procedure was repeated with a second grid.

Each participant's speech was recorded throughout the experiment, using an iRiver H120 digital recorder.

Transcription and coding Transcription and coding were carried out by the first author. Due to experimenter error, recordings of a single grid were missing for each of 2 participants. Thus, the analysis was based on recordings of 78 grid descriptions.

Each grid description was first divided into 16 utterances describing the location of each picture, which tended to consist of two parts: a description of the numeric location, followed by an image description. Example transcriptions of fluent and disfluent utterances locating pictures are given in (1).

(1a) On five there is a leaf.
(1b) In the fifth box there is a: [pause] um [pause] tape recording device.

The 1,248 resulting utterances were then coded as follows. For the 624 alignment images, we recorded whether each image was given the preferred or dispreferred name (23 utterances used other names and were discounted from further analysis). Where participants used more than one name, the first name used was recorded.

Coding for the 624 disfluency images was restricted to the image description part of each relevant utterance, which included the image name and preceding function, but not content, words (e.g., "there is a . . . device" in 1b). A data-driven approach was taken to generating categories of disfluency. Disfluencies in the first 10 sets of transcriptions were used to generate categories. Each utterance was scored as fluent (no discernible disfluency) or as disfluent, and numbers of disfluencies in each category (prolongation, *uh*, *um*, hesitation, or repetition) were additionally noted.

Results

We conducted two independent analyses. The first, focusing on the alignment images, established whether the names that participants chose for these images were affected by the names a confederate used. The second, using the disfluency images, investigated whether the disfluencies participants produced were influenced by the presence of a confederate.

Because our dependent variables were binomial (whether or not the dispreferred name had been used; whether or not there was a disfluency), we modeled outcome likelihood, using logit mixed effects models (Breslow & Clayton, 1993; DebRoy & Bates, 2004). All analyses were carried out in R (R Development Core Team, 2011) using the lme4 package (Bates, Maechler, & Bolker, 2011). All predictors were sum coded, with values of −.5 and .5 chosen as levels (confederate absent/present, preferred/dispreferred name scripted, easy/difficult, respectively), allowing odds ratios to be readily calculated without additional manipulation of model coefficients. For each analysis, we constructed a full model (with maximal random effect structure) and report the coefficients for each fixed effect, together with the likelihood that each coefficient equals zero, derived from Wald's *Z*.

Influence of confederate on naming

Table 1 shows the proportions of trials on which participants chose dispreferred names for the alignment images. In conditions where a confederate was present,

Table 1 Proportions of trials on which participants used the dispreferred name to refer to alignment images for Experiments 1 and 2 (with standard errors in parentheses). Where the confederate was present, a preferred or dispreferred name was scripted and would previously have been heard by the participant; where the confederate was absent, the scripted name was nominal only, in that no name was actually heard before the participant named each item

		Confederate Absent	Confederate Present
Exp. 1	Preferred name scripted	.10 (.03)	.03 (.02)
	Dispreferred name scripted	.11 (.02)	.47 (.04)
Exp. 2	Preferred name scripted	.18 (.04)	.04 (.02)
	Dispreferred name scripted	.12 (.02)	.59 (.03)

63 % of the alignment images would have been previously referred to using a dispreferred name (since, for each grid, the confederate's script included a dispreferred name for five out of eight alignment images). In cases where there was no confederate, these images are still referred to as being in a *dispreferred* condition; since the experiment was fully counterbalanced, we can compare cases where the dispreferred name was previously mentioned (in dialogue) with cases where there was no confederate present to mention it.

Participants were found to be over six and a half times more likely to use a dispreferred name when the confederate was present ($p < .001$), $\beta = 1.884$, $SE = 0.448$ ($e^{1.884} = 6.578$), and over 17 times as likely when a dispreferred name was scripted ($p < .01$), $\beta = 2.866$, $SE = 0.974$. These two factors were found to interact ($p < .001$), $\beta = 3.876$, $SE = 1.137$, showing that participants were sensitive to the name previously used by their partner when it was their turn to name the image.

Influence of confederate on disfluency

Because of different views on the communicative function of silent pauses, we analyzed the disfluencies produced first including and then without including the *silence* category. The proportions of trials on which participants used a disfluency in naming disfluency images are shown in Table 2.

Table 2 Proportions of trials on which participants referred disfluently to disfluency images for Experiments 1 and 2 (with standard errors in parentheses)

		Confederate Absent	Confederate Present
Exp. 1	Easy images	.09 (.02)	.09 (.02)
	Hard images	.18 (.03)	.23 (.03)
Exp. 2	Easy images	.10 (.02)	.11 (.02)
	Hard images	.31 (.03)	.35 (.03)

Including silences, images classified as difficult were 3 times as likely to be associated with disfluency as were easy images ($p < .01$), $\beta = 1.125$, $SE = 0.409$ ($e^{1.125} = 3.080$). However, no effect was found for the presence of a confederate ($p < 1$), $\beta = 0.239$, $SE = 0.420$, suggesting that participants were no more (or less) likely to be disfluent when a partner was present. There was no evidence of any interaction between these factors ($p < 1$), $\beta = 0.414$, $SE = 0.540$. Disfluencies other than silences were over 2 times as likely to be produced when difficult images were named ($p = .02$), $\beta = 0.858$, $SE = 0.353$. Without silences, no other effect reached significance ($ps > .89$).

To test whether the distributions of participants' disfluencies were affected by the presence of a confederate, we tabulated the total numbers of disfluencies in five categories observed across the experiment. Table 3 shows the totals observed in the presence and absence of a confederate. As can be clearly seen, the distribution of disfluencies was not affected by the presence of a confederate, a fact confirmed by Fisher's exact test ($p = .95$).

Discussion

Experiment 1 showed that, while word choice differed between monologue and dialogue, the use of disfluency did not. However, given that participants named items disfluently on fewer than 10 % of occasions, it is possible that the lack of disfluency effect reflects a scarcity of observations. To address this issue and to ensure that the null effect obtained in Experiment 1 could be replicated, we ran an additional experiment, which was identical to Experiment 1 except that the images used were digitally manipulated to make them harder to recognize and, therefore, more likely to result in disfluent descriptions.

Experiment 2

For Experiment 2, the 96 images used for Experiment 1 were blurred using a Gaussian algorithm ($\sigma = 6$ pixels). In

Table 3 Total numbers of disfluencies observed in each of five categories across the experiment

		Prolongation	*Uh*	*Um*	Repetition	Silence
Exp. 1	Confederate absent	41	7	8	4	54
	Confederate present	35	7	10	5	51
Exp. 2	Confederate absent	79	3	21	12	141
	Confederate present	66	8	18	8	112

all other respects, the experiment was identical to Experiment 1; participants were 24 native British-English-speaking undergraduate students from the University of Edinburgh, who participated in return for course credit. Speech was recorded using a ZOOM H4n digital recorder.

Transcription and coding

Two raters each transcribed and coded the recordings for 15 participants. Raters were instructed to count the occurrences of each of the five categories of disfluency identified in Experiment 1. For the 6 participants who were rated by both raters, there was 86.4 % agreement on disfluencies. For each of these 6 participants, one rater's coding was selected at random for analysis.

Results

Influence of confederate on naming

Table 1 shows the proportions of trials on which participants chose dispreferred names for the alignment images. Participants were over 6 times more likely to use dispreferred names when a confederate was present ($p < .001$), $\beta = 1.840$, $SE = 0.553$ ($e^{1.840} = 6.297$). When a dispreferred name had been scripted, participants were over 20 times more likely to use it themselves ($p < .01$), $\beta = 3.019$, $SE = 1.072$; as in Experiment 1, the two factors interacted ($p < .001$), $\beta = 6.699$, $SE = 1.372$.

Influence of confederate on disfluency

Table 2 shows the proportions of disfluent trials. Including silences, difficult images were almost 7 times as likely to be associated with disfluency as easy images ($p < .001$), $\beta = 1.917$, $SE = 0.459$ ($e^{1.917} = 6.800$). No effect was found for the presence of a confederate ($p < 1$), $\beta = 0.216$, $SE = 0.247$, and these two factors did not interact ($p < 1$), $\beta = -0.044$, $SE = 0.528$. Disfluencies other than silences were almost five and a half times as likely to be associated with disfluency ($p < .001$), $\beta = 1.700$, $SE = 0.414$. Without silences, no other effect reached significance ($ps > .43$).

Counts for each category in the presence and absence of a confederate are shown in Table 3. A Fisher's exact test showed that the presence of a confederate did not influence the distribution of disfluencies ($p = .47$).

A final analysis combined the data from both experiments. A regression model was constructed that included a fixed effect for experiment, which was allowed to interact with all other fixed effects, and an experiment-by-items random slope. Speakers were no more likely to be disfluent in the presence of a confederate (p < 1), $\beta = 0.166$, $SE = 0.171$. A main effect of experiment showed that using blurred images made participants over one and a half times more likely to be disfluent ($p < .05$), $\beta = 0.466$, $SE = 0.207$. A marginal interaction between experiment and difficulty suggested that the effect of blurring on disfluency was larger for difficult images ($p = .09$), $\beta = 0.693$, $SE = 0.410$. No other interactions with experiment were significant (all $ps < 1$). An analysis excluding silences confirmed this pattern of results, although the effect of experiment became marginal ($p = .06$), $\beta = 0.450$, $SE = 0.242$.

General discussion

The present study was designed to investigate whether or not disfluencies are used by speakers to signal difficulty to their interlocutors. We manipulated whether a task was performed communicatively (in a dialogue condition) or noncommunicatively (in a monologue condition) and investigated the effects of this manipulation on the production of disfluency. As a precondition to being able to interpret our findings, we had to show that in the dialogue condition, speakers were in fact producing language that took their listeners into account. Results from the alignment items show unequivocally that this was true. In line with previous, similar work (Clark & Wilkes-Gibbs, 1986; Cleland & Pickering, 2003), when the experimental confederate referred to a picture using a dispreferred name, participants were many times more likely to choose that name to refer to the same picture than they were in cases where the more common, preferred, name had previously been used.

Having established that participants' language choices were affected by the presence of an interlocutor, the question remains of what factors caused them to be disfluent. Participants were much more likely to refer disfluently to images when those images corresponded to several names (cf. Hartsuiker & Notebaert, 2010) and the most commonly used name was low frequency. These effects were exacerbated when the images were blurred (cf. Schnadt & Corley, 2006). This suggests that disfluencies reflect cognitive difficulty, either in selecting a particular name (cf. Vitkovich & Tyrrell, 1995) or in retrieving a low-frequency name (cf. Caramazza, Costa, Miozzo, & Bi, 2001; Jescheniak & Levelt, 1994). However, there was no evidence at all that the presence of an interlocutor in the dialogue condition affected the likelihood of being disfluent. Moreover, this finding is not the consequence of conflating different types of disfluency. If particular disfluencies are viewed as communicative signals of upcoming difficulty (cf. Fox Tree, 2001), we might expect participants to use them more with a listener present. But there was no suggestion that nonsilent disfluencies were used more often or that the distributions of disfluency types used differed between monologue and dialogue conditions.

There are three potential interpretations of these findings. First, participants might not have had awareness of the confederate in any significantly communicative sense and might, instead, have viewed each condition as a monologue. According to this view, lexical alignment with the confederate would be attributed to straightforward priming, and disfluency levels across conditions would remain constant because the conditions were communicatively equivalent.

We would not wish to contest that priming has a role to play, given Pickering and Garrod's (2004) view that priming mechanisms are fundamental to alignment in dialogue. However, evidence suggests that, at least at the lexical level, the names chosen for images are influenced by beliefs about one's interlocutor (Branigan, Pickering, Pearson, MacLean, & Brown, 2011), as part of a general tendency for speakers to take into account what they believe their listeners to know (Isaacs & Clark, 1987), and we see no reason to believe that our participants were not sensitive to these factors. Moreover, if both conditions were perceived as monologues, proponents of the "disfluency as signal" view would need to account for the fact that disfluencies were uttered throughout both experiments (and 12 % of these where either *um* or *uh*).

A second interpretation of the present findings relies on the observation that dialogue is the most common form of speech, while monologue is a special case (Garrod & Pickering, 2004; Pickering & Garrod, 2004). It is possible that participants continued to use disfluency as a signal in the monologue conditions either out of habit or, perhaps, even because they lacked a special set of communicative strategies that were more suitable for monologue. Anecdotally, there do appear to be occasions where disfluency rates are adapted for monologue—in public speaking, for example—but if one accepts the view that the use of disfluencies as signals is a habit that is hard to break, then testing Clark and Fox Tree's (2002) suggestion that disfluencies are used as communicative signals is likely to be difficult. One possibility may be to explore the developmental evidence: Whereas there is evidence that children as young as 2 can infer that an adult is likely to refer to a novel object after a filler (Kidd, White, & Aslin, 2011), the distinction reported by Clark and Fox Tree (2002) between pauses following *um* and *uh* does not appear in the speech of 3- to 4-year-olds (Hudson Kam & Edwards, 2008).

For Clark and Fox Tree (2002), the use of particular disfluencies is clearly seen as intentional. But determining whether speakers are doing something intentionally is difficult, particularly when they are not consciously aware of doing it. Thus, the claim that speakers use disfluencies to communicate remains uncontested, not because it is right (or wrong), but because it is difficult to verify. In the absence of a direct solution to this problem, the present article provides the first example of experimental disfluency research focusing specifically on the case of dialogue. We replicated previous findings on lexical alignment and showed that the production of disfluencies was affected by the ease with which words in the intended message could be selected. However, we found no evidence to suggest that the disfluencies a speaker produces are influenced by the presence of a listener. Whereas this finding does not rule out the possibility that disfluencies are created intentionally, it does not provide evidence to support this claim. The third, and simplest, account of the existing evidence is, therefore, that disfluencies do not serve a communicative purpose, other than in the sense that listeners are able to exploit their occurrence in predictable circumstances. Instead, they are by-products of difficulty in speech, whether there is someone present to whom the difficulty can be communicated or not.

Author Note The authors wish to thank Ewan Keith and Genevieve Warriner-Gallyer for help with Experiment 2.

References

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

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX Lexical Database.* Retrieved from http://celex.mpi.nl/

Barr, D. J. (2001). Trouble in mind: Paralinguistic indices of effort and uncertainty in communication. In: S. Santi, I. Guatella, C. Cave, & G. Konopczynski (Eds.), Oralité et gestualité: Interactions et comportements multimodaux dans la communication (pp 595–600). Paris: L'Harmattan.

Bates, D., Maechler, M., & Bolker, B. (2011). lme4: Linear mixed-effects models using S4 classes [Computer software manual]. Available from http://cran.r-project.org/package=lme4 (R package version 0.999375-39).

Bavelas, J., Gerwing, J., Sutton, C., & Prevost, D. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language, 58,* 495–520.

Beattie, G. W., & Butterworth, B. (1979). Contextual probability and word frequency as determinants of pauses in spontaneous speech. *Language and Speech, 22,* 201–211.

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

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech, 44,* 123.

Branigan, H. P., Pickering, M. J., Pearson, J., MacLean, J. F., & Brown, A. (2011). The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition, 121,* 41–57.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Society, 88,* 9–25.

Caramazza, A., Costa, A., Miozzo, M., & Bi, Y. (2001). The specific-word frequency effect: Implications for the representation of homophones in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1430–1450.

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

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.

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

Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition, 105,* 658–668.

DebRoy, S., & Bates, D. M. (2004). Linear Mixed models and penalized least squares. *Journal of Multivariate Analysis, 91,* 1–17.

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

Fox Tree, J. E. (2001). Listeners' uses of um and uh in speech comprehension. *Memory & Cognition, 29,* 320–326.

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

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences, 8,* 8–11.

Griffin, Z. M., & Huitema, J. (1999). *Beckman Spoken Picture Naming Norms.* Retrieved from http://langprod.cogsci.uiuc.edu/~norms/

Hartsuiker, R. J., & Notebaert, L. (2010). Lexical access problems lead to disfluencies in speech. *Experimental Psychology, 57,* 169–177.

Hudson Kam, C. L., & Edwards, N. A. (2008). The use of uh and um by 3-and 4-year-old native English-speaking children: Not quite right but not completely wrong. *First Language, 28,* 313.

Isaacs, E. A., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology. General, 116,* 26–37.

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 824–843.

Kidd, C., White, K. S., & Aslin, R. N. (2011). Toddlers use speech disfluencies to predict speakers' referential intentions. *Developmental Science, 14,* 925–934.

Lake, J. K., Humphreys, K. R., & Cardy, S. (2011). Listener vs. speaker-oriented aspects of speech: Studying the disfluencies of individuals with autism spectrum disorders. *Psychonomic Bulletin & Review, 18,* 135–140.

MacGregor, L. J., Corley, M., & Donaldson, D. I. (2009). Not all disfluencies are are equal: The effects of disfluent repetitions on language comprehension. *Brain and Language, 111,* 36–45.

MacGregor, L. J., Corley, M., & Donaldson, D. I. (2010). Listening to the sound of silence: Investigating the consequences of disfluent silent pauses in speech for listeners. *Neuropsychologia, 48,* 3982–3992.

Merlo, S., & Mansur, L. L. (2004). Descriptive discourse: Topic familiarity and disfluencies. *Journal of Communication Disorders, 37,* 489–503.

O'Connell, D. C., & Kowal, S. (2005). Uh and um revisited: Are they interjections for signaling delay? *Journal of Psycholinguistic Research, 34,* 555–576.

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

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *The Behavioral and Brain Sciences, 27,* 169–190.

R Development Core Team. (2011). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from http://www.r-project.org (ISBN 3-900051-07-0).

Schachter, S., Christenfeld, N., Ravina, B., & Bilous, F. (1991). Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology, 60,* 362–367.

Schnadt, M. J., & Corley, M. (2006). The influence of lexical, conceptual and planning based factors on disfluency production. In: *Proceedings of the Twenty-Eighth Meeting of the Cognitive Science Society* (pp. 750–755). Vancouver, Canada.

Shriberg, E. E. (1996). Disfluencies in switchboard. In: *Proceedings of the International Conference on Spoken Language Processing, Addendum* (pp. 11–14). Philadelphia, PA.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 174–215.

Swerts, M., & Krahmer, E. (2005). Audiovisual prosody and feeling of knowing. *Journal of Memory and Language, 53,* 81–94.

Szekely, A., Jacobsen, T., D'Amico, S., Devescovi, A., Andonova, E., Herron, D.,… Bates, E. (2004). A new on-line resource for psycholinguistic studies. *Journal of Memory and Language*, *51*, 247–250.

Vitkovich, M., & Tyrrell, L. (1995). Sources of disagreement in object naming. *Quarterly Journal of Experimental Psychology, 48A,* 822–848.

References

Agresti, A. (2003). *Categorical data analysis.* Hoboken, NJ: Wiley-Interscience.

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419–439.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G. M., Garrod, S., ... Weinert, R. (1991). The HCRC Map Task Corpus. *Language and Speech*, *34*, 351–366.

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. *Attention, Perception, & Psychophysics*, *59*(4), 580–592.

Anderson, K. J., & Leaper, C. (1998). Meta-analyses of gender effects on conversational interruption: Who, what, when, where, and how. *Sex Roles*, *39*(3-4), 225–252.

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

Arnold, J. E., Hudson Kam, C. L., & Tanenhaus, M. K. (2007). If you say *thee uh* you are describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 914–930.

Arnold, J. E., Wasow, T., Ginstrom, R., & Losongco, T. (2000). Heaviness vs newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, *76*, 28–55.

Auer, P., Couper-Kuhlen, E., & Müller, F. (1999). *Language in time: The rhythm and tempo of spoken interaction.* New York, NY: Oxford University Press.

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

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* New York, NY: Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX Lexical Database.* http://celex.mpi.nl/.

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

Ball, P. (1975). Listeners' responses to filled pauses in relation to floor apportionment. *British Journal of Social and Clinical Psychology*.

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), 1–22.

Bard, E. G., & Lickley, R. J. (1998). Disfluency deafness: Graceful failure in the recognition of running speech. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 108–113).

Bard, E. G., Lickley, R. J., & Aylett, M. P. (2001). Is disfluency just difficulty? In *Disfluency in Spontaneous Speech (DiSS'01), ISCA Tutorial and Research Workshop (ITRW)* (pp. 97–100).

Bard, E. G., Shillcock, R. C., & Altmann, G. T. M. (1988). The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. *Perception & Psychophysics*, *44*(5), 395–408.

Barr, D. J. (2001). Trouble in mind: Paralinguistic indices of effort and uncertainty in communication. In S. Santi, I. Guãtella, C. Cave, & G. Konopczynski (Eds.), *Oralité et gestualité: Interactions et comportements multimodaux dans la communication* (pp. 597–600). L'Harmattan.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure in mixed-effects models: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Barr, D. J., & Seyfeddinipur, M. (2010). The role of fillers in listener attributions for speaker disfluency. *Language and Cognitive Processes*, *25*(4), 441–455.

Bastiaansen, M., & Hagoort, P. (2006). Oscillatory neuronal dynamics during language comprehension. *Progress in Brain Research*, *159*, 179–196.

Bates, D., Maechler, M., & Bolker, B. (2013). [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=lme4 (R package version 0.999999-2)

Bates, D. M. (2006, May). *lmer, p-values and all that.* Retrieved from `https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html`

Bavelas, J., Gerwing, J., Sutton, C., & Prevost, D. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, *58*(2), 495–520.

Beattie, G. W. (1977). The dynamics of interruption and the filled pause. *British Journal of Social and Clinical Psychology*, *16*(3), 283–284.

Beattie, G. W. (1978). Floor apportionment and gaze in conversational dyads. *British Journal of Social and Clinical Psychology*, *17*(1), 7–15.

Beattie, G. W. (1981). The regulation of speaker turns in face-to-face conversation: Some implications for conversation in sound-only communication channels. *Semiotica*, *34*(1-2), 55–70.

Beattie, G. W., & Butterworth, B. (1979). Contextual probability and word frequency as determinants of pauses in spontaneous speech. *Language and Speech*, *22*, 201–211.

Belin, P., Zilbovicius, M., Crozier, S., Thivard, L., Fontaine, A., Masure, M.-C., & Samson, Y. (1998). Lateralization of speech and auditory temporal processing. *Journal of Cognitive Neuroscience*, *10*(4), 536–540.

Bell, A. (1984). Language style as audience design. *Language in Society*, *13*(2), 145–204.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational english. *Journal of Memory and Language*, *60*(1), 92–111.

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

Beňuš, Š. (2009). Are we "in sync": Turn-taking in collaborative dialogues. In *INTERSPEECH 2009 - 10th Annual Conference of the International Speech Communication Association* (pp. 2167–2170).

Berkum, J. J. A. V., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443–467.

Binnenpoorte, D., Bael, C. V., Os, E. D., & Boves, L. (2005). Gender in everyday speech and language: a corpus-based study. In *Interspeech'2005 - Eurospeech, 9th European Conference on Speech Communication and Technology.* Lisbon, Portugal.

Birch, S., & Rayner, K. (1997). Linguistic focus affects eye movements during reading. *Memory & Cognition*, *25*(5), 653–660.

Blackmer, E. R., & Mitton, J. L. (1991). Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition*, *39*(3), 173–194.

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

Boomer, D. S. (1965). Hesitation and grammatical encoding. *Language and Speech*, *8*(3), 148–158.

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech*, *44*(2), 123–147.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*, 211–252.

Boyle, E. A., Anderson, A. H., & Newlands, A. (1994). The effects of visibility on dialogue and performance in a cooperative problem solving task. *Language and Speech*, *37*(1), 1–20.

Bradley, R. A., & Srivastava, S. S. (1979). Correlation in polynomial regression. *The American Statistician*, *33*(1), 11–14.

Branigan, H. P., Lickley, R. J., & McKelvie, D. (1999). Non-linguistic influences on rates of disfluency in spontaneous speech. In *Proceedings of the XIVth International Congress of Phonetic Sciences.* San Francisco.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic coordination in dialogue. *Cognition*, *75*(2), 13–25.

Bredart, S., & Modolo, K. (1988). Moses strikes again: Focalization effect on a semantic illusion. *Acta Psychologica*, *67*(2), 135–144.

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

Brennan, S. E., & Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*, *34*(3), 383–398.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Society*, *88*, 9–25.

British National Corpus. (1995). Retrieved from `http://www.natcorp.ox.ac.uk/`

Broen, P. A., & Siegel, G. M. (1972). Variations in normal speech disfluencies. *Language and Speech*, *15*(3), 219–231.

Brown, P. M., & Dell, G. S. (1987). Adapting production to comprehension: The explicit mention of instruments. *Cognitive Psychology*, *19*(4), 441–472.

Bull, M. (1996). An analysis of between-speaker intervals. In *Proceedings of the Edinburgh Postgraduate Conference in Linguistics and Applied Linguistics '96* (pp. 18–27).

Bull, M., & Aylett, M. (1998). An analysis of the timing of turn-taking in a corpus of goal-oriented dialogue. In *Proceedings of the 5th International Conference on Spoken Language Processing.*

Buzsáki, G., & Draguhn, A. (2004). Neuronal oscillations in cortical networks. *Science*, *304*(5679), 1926–1929.

Byrne, D., Griffitt, W., & Stefaniak, D. (1967). Attraction and similarity of personality characteristics. *Journal of Personality and Social Psychology*, *5*(1), 82–90.

Call, J., & Tomasello, M. (1994). Production and comprehension of referential pointing by orangutans (*Pongo pygmaeus*). *Journal of Comparative Psychology*, *108*(4), 307–317.

Campione, E., & Véronis, J. (2002). A large-scale multilingual study of silent pause duration. In *Speech prosody 2002, international conference.*

Cappella, J. N., & Planalp, S. (1981). Talk and silence sequences in informal conversations III: Interspeaker influence. *Human Communication Research*, *7*(2), 117–132.

Carletta, J., Evert, S., Heid, U., & Kilgour, J. (2006). The NITE XML Toolkit: Data model and query language. *Language Resources and Evaluation*, *39*, 313–334.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., & Anderson, A. (1996). *HCRC Dialogue Structure Coding Manual* (Tech. Rep. No. 82). HCRC.

Carletta, J., & Mellish, C. S. (1996). Risk-taking and recovery in task-oriented dialogue. *Journal of Pragmatics*, *26*(1), 71–107.

Caspers, J. (2003). Local speech melody as a limiting factor in the turn-taking system in dutch. *Journal of Phonetics*, *31*(2), 251–276.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*(7), e1000436.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology*, *76*(6), 893–910.

Christenfeld, N. (1995). Does it hurt to say um? *Journal of Nonverbal Behavior*, *19*, 171–186.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359.

Clark, H. H. (1994). Managing problems in speaking. *Speech Communication*, *15*, 243–250.

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

Clark, H. H. (2002). Speaking in time. *Speech Communication*, *36*(1), 5–13.

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

Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. K. Joshi, I. A. Sag, & B. L. Webber (Eds.), *Elements of discourse understanding.* 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 Behavior*, *22*, 245–258.

Clark, H. H., & Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive Psychology*, *37*(3), 201–242.

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

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

Collard, P. (2009). *Disfluency and listeners' attention: An investigation of the immediate and lasting effects of hesitations in speech.* Unpublished doctoral dissertation, University of Edinburgh.

Collard, P., Corley, M., MacGregor, L. J., & Donaldson, D. I. (2008). Attention orienting effects of hesitations in speech: Evidence from ERPs. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 696–702.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, *33*(4), 497–505.

Cook, M. (1969). Transition probabilities and the incidence of filled pauses. *Psychonomic Science*, *16*, 191–192.

Cook, M., & Lallgee, M. (1970). The interpretation of pauses by the listener. *British Journal of Social and Clinical Psychology*, *9*(4), 375–376.

Cook, M., Smith, J., & Lalljee, M. G. (1974). Filled pauses and syntactic complexity. *Language and Speech*, *17*(1), 11–16.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, *6*(1), 84–107.

Corley, M., & Hartsuiker, R. J. (2011). Why um helps auditory word recognition: The temporal delay hypothesis. *PloS One*, *6*(5), e19792.

Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition*, *105*, 658–668.

Corley, M., & Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, *2*(4), 589–602.

Coulston, R., Oviatt, S., & Darves, C. (2002). Amplitude convergence in children's conversational speech with animated personas. In *7th International Conference on Spoken Language Processing (ICSLP2002 - INTERSPEECH 2002)* (Vol. 4, pp. 2689–2692).

Coupland, N., & Giles, H. (1988). Introduction the communicative contexts of accommodation. *Language & Communication*, *8*(3), 175–182.

Cowan, J. M., & Bloch, B. (1948). An experimental study of pause in english grammar. *American Speech*, 89–99.

Cutler, A., & Pearson, M. (1986). On the analysis of prosodic turn-taking cues. In C. Johns-Lewis (Ed.), *Intonation in discourse* (pp. 139–156). London, UK: Croom Helm.

DebRoy, S., & Bates, D. M. (2004). Linear mixed models and penalized least squares. *Journal of Multivariate Analysis*, *91*, 1–17.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word preactivation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117–1121.

De Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 515–535.

Duez, D. (1985). Perception of silent pauses in continuous speech. *Language and Speech*, *28*(4), 377–389.

Dumas, G., Nadel, J., Soussignan, R., Martinerie, J., & Garnero, L. (2010). Interbrain synchronization during social interaction. *PloS One*, *5*(8), e12166.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, *23*(2), 283–292.

Duncan, S. (1974). On the structure of speaker-auditor interaction during speaking turns. *Language in Society*, *3*(2), 161–180.

Duncan, S., Brunner, L. J., & Fiske, D. W. (1979). Strategy signals in face-to-face interaction. *Journal of Personality and Social Psychology*, *37*(2), 301–313.

Duncan, S., & Niederehe, G. (1974). On signalling that it's your turn to speak. *Journal of Experimental Social Psychology*, *10*(3), 234–247.

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, *24*(6), 409–436.

Eklund, R. (1999). A comparative study of disfluencies in four swedish travel dialogue corpora. In *Proceedings of the icphs satellite workshop on disfluency in spontaneous speech (diss)* (pp. 3–6).

Eklund, R. (2001). Prolongations: A dark horse in the disfluency stable. In *Disfluency in Spontaneous Speech (DiSS'01), ISCA Tutorial and Research Workshop (ITRW)* (pp. 5–8).

Eklund, R., & Shriberg, E. E. (1998). Crosslinguistic disfluency modeling: A comparative analysis of Swedish and American English human–human and human–machine dialogs. In *The 5th International Conference on Spoken Language Processing* (pp. 2631–2634).

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, *20*(5), 540–551.

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: A TMS study. *European Journal of Neuroscience*, *15*, 399–402.

Ferreira, F. (1991). Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language*, *30*(2), 210–233.

Ferreira, F. (1993). Creation of prosody during sentence production. *Psychological Review*, *100*(2), 233–253.

Ferreira, F. (2007). Prosody and performance in language production. *Language and Cognitive Processes*, *22*(8), 1151–1177.

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, *11*(1), 11–15.

Ferreira, F., & Henderson, J. M. (1991). Recovery from misanalyses of garden-path sentences. *Journal of Memory and Language*, *30*(6), 725–745.

Ferreira, F., & Patson, N. (2007). The “good enough” approach to language comprehension. *Language and Linguistics Compass*, *1*(1-2), 71–83.

Finlayson, I. R., & Corley, M. (2012). Disfluency in dialogue: An intentional signal from the speaker? *Psychonomic Bulletin & Review*, *19*(5), 921–928.

Finlayson, I. R., Lickley, R. J., & Corley, M. (2010). The influence of articulation rate, and the disfluency of others, on one’s own speech. In *Proceedings of DiSS-LPSS Joint Workshop* (pp. 119–122).

Forsyth, R., Clarke, D., & Lam, P. (2008). Timelines, talk and transcription: A chronometric approach to simultaneous speech. *International Journal of Corpus Linguistics*, *13*(2), 225–250.

Fowler, C. A., & Housum, J. (1987). Talkers’ signaling of “new” and “old” words in speech and listeners’ perception and use of the distinction. *Journal of Memory and Language*, *26*(5), 489–504.

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

Fox Tree, J. E. (2001). Listeners’ uses of um and uh in speech comprehension. *Memory & Cognition*, *29*(2), 320–326.

Fox Tree, J. E. (2002). Interpreting pauses and ums at turn exchanges. *Discourse Processes*, *34*(1), 37–55.

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

Fraundorf, S. H., & Watson, D. G. (2008). Dimensions of variation in disfluency production in discourse. In *Proceedings of the 12th Workshop on the Semantics and Pragmatics of Dialogue (Londial 2008)* (pp. 131–138).

Fraundorf, S. H., & Watson, D. G. (2011). The disfluent discourse: Effects of filled pauses on recall. *Journal of Memory and Language*, *65*(2), 161–175.

Fries, P. (2005). A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends in Cognitive Sciences*, *9*(10), 474–480.

Fry, D. (1975). Simple reaction-times to speech and non-speech stimuli. *Cortex*, *11*(4), 355–360.

Gambi, C., & Pickering, M. J. (2011). A cognitive architecture for the coordination of utterances. *Frontiers in Psychology*, *2*.

Gann, T. M., & Barr, D. J. (2012). Speaking from experience: Audience design as expert performance. *Language and Cognitive Processes*, 1–23.

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

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, *8*, 8–11.

Gelman, A. (2005). Analysis of variance–why it is more important than ever. *The Annals of Statistics*, *33*(1), 1–53.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* New York, NY: Cambridge University Press.

Ghitza, O. (2011). Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. *Frontiers in Psychology*, *2*.

Ghitza, O., & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, *66*(1-2), 113–126.

Giles, H. (1973). Accent mobility: A model and some data. *Anthropological Linguistics*, *15*(2), 87–105.

Giles, H., Coupland, J., & Coupland, N. (1991). *Contexts of accommodation: Developments in applied sociolinguistics.* New York, NY: Cambridge University Press.

Giles, H., & Powesland, P. F. (1975). A social psychological model of speech diversity. *Speech Style and Social Evaluation*, 154–70.

Giles, H., & Smith, P. (1979). Accommodation theory: Optimal levels of convergence. In H. Giles & R. St. Clair (Eds.), *Language and social psychology.* Oxford: Basil Blackwell.

Giraud, A.-L., Kleinschmidt, A., Poeppel, D., Lund, T. E., Frackowiak, R. S., & Laufs, H. (2007). Endogenous cortical rhythms determine cerebral specialization for speech perception and production. *Neuron*, *56*(6), 1127–1134.

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–517.

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 1992.* (Vol. 1, pp. 517–520).

Goldinger, S. D. (1998). Echoes of echoes? an episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279.

Goldman-Eisler, F. (1958). The predictability of words in context and the length of pauses in speech. *Language and Speech*, *1*(3), 226–231.

Goldman-Eisler, F. (1954). On the variability of the speed of talking and on its relation to the length of utterances in conversations. *British Journal of Psychology. General Section*, *45*(2), 94–107.

Gravano, A. (2009). *Turn-taking and affirmative cue words in task-oriented dialogue.* Unpublished doctoral dissertation, Columbia University.

Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, *25*(3), 601–634.

Gregory, S. W. (1990). Analysis of fundamental frequency reveals covariation in interview partners' speech. *Journal of Nonverbal Behavior*, *14*(4), 237–251.

Gregory, S. W., & Webster, S. (1996). A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions. *Journal of Personality and Social Psychology*, *70*, 1231–1240.

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

Grice, H. P. (1969). Utterer's meaning and intention. *The Philosophical Review*, *78*(2), 147–177.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41–58). New York: Seminar Press.

Griffin, Z. M. (2003). A reversed word length effect in coordinating the preparation and articulation of words in speaking. *Psychonomic Bulletin & Review*, *10*(3), 603–609.

Griffin, Z. M., & Huitema, J. (1999). *Beckman Spoken Picture Naming Norms.* `http://langprod.cogsci.uiuc.edu/∼norms/`.

Gross, J., & Ligges, U. (2012). nortest: Tests for normality [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=nortest` (R package version 1.0-2)

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, *69*(2), 274–307.

Harris, H. D. (2002). Holographic reduced representations for oscillator recall: A model of phonological production. In W. D. Gray & S. C. D (Eds.), *The proceedings of the 24th annual meeting of the cognitive science society* (pp. 423–428).

Hartsuiker, R. J., & Notebaert, L. (2010). Lexical access problems lead to disfluencies in speech. *Experimental Psychology*, *57*(3), 169–177.

Hawkins, P. R. (1971). The syntactic location of hesitation pauses. *Language and Speech*, *14*(3), 277–288.

Heim, S., Opitz, B., Müller, K., & Friederici, A. (2003). Phonological processing during language production: fmri evidence for a shared production-comprehension network. *Cognitive Brain Research*, *16*(2), 285–296.

Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, *38*(4), 555–568.

Heldner, M., Edlund, J., & Hirschberg, J. (2010). Pitch similarity in the vicinity of backchannels. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association* (pp. 3054–3057).

Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences*, *4*(4), 131–138.

Hieke, A. E. (1981). A content-processing view of hesitation phenomena. *Language and Speech*, *24*(2), 147–160.

Hieke, A. E., Kowal, S., & O'Connell, D. C. (1983). The trouble with "articulatory" pauses. *Language and Speech*, *26*(3), 203–214.

Hindle, D. (1983). Deterministic parsing of syntactic non-fluencies. In *Proceedings of the 21st annual meeting on Association for Computational Linguistics* (pp. 123–128).

Hjalmarsson, A. (2011). The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication*, *53*(1), 23–35.

Hollingworth, A., & Henderson, J. M. (2002). Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human, Perception, and Performance*, *28*(1), 113–136.

Horton, B. W. (2007). *Predicting common ground sequences from prosody, timing, friendship, and experience.* Unpublished doctoral dissertation, The Ohio State University.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, *59*(1), 91–117.

Hostetter, A. B., Cantero, M., & Hopkins, W. D. (2001). Differential use of vocal and gestural communication by chimpanzees (*Pan troglodytes*) in response to the attentional status of a human (*Homo sapiens*). *Journal of Comparative Psychology*, *115*(4), 337–343.

Howell, P., & Au-Yeung, J. (2001). Application of EXPLAN theory to spontaneous speech control. In *Disfluency in Spontaneous Speech (DiSS'01), ISCA Tutorial and Research Workshop (ITRW)* (pp. 9–12).

Howell, P., & Au-Yeung, J. (2002). The EXPLAN theory of fluency control applied to the diagnosis of stuttering. In E. Fava (Ed.), *Clinical linguistics: Theory and applications in speech pathology and therapy* (pp. 75–94). Amsterdam: John Benjamins Publishing.

Hudson Kam, C. L., & Edwards, N. A. (2008). The use of uh and um by 3- and 4-year-old native english-speaking children: Not quite right but not completely wrong. *First Language*, *28*(3), 313–327.

Indefrey, P., & Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*(1), 101–144.

Isaacs, E. A., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology. General*, *116*, 26–37.

Isard, A. (2001). An XML Architecture for the HCRC Map Task Corpus. In *Proceedings of the 5th International Workshop on Formal Semantics and Pragmatics of Dialogue (BI-DIALOG 2001).* Bielefeld, Germany.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446.

Jarvella, R. J. (1971). Syntactic processing of connected speech. *Journal of Verbal Learning and Verbal Behavior*, *10*(4), 409–416.

Jefferson, G. (1989). Preliminary notes on a possible metric which provides for a "standard maximum" silence of approximately one second in conversation. In D. Roger & P. Bull (Eds.), *Conversation: an interdisciplinary perspective.* Multilingual Matters.

Jensen, O., Kaiser, J., & Lachaux, J.-P. (2007). Human gamma-frequency oscillations associated with attention and memory. *Trends in Neurosciences*, *30*(7), 317–324.

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*, 824–843.

Johnsrude, I., Zatorre, R., Milner, B., & Evans, A. (1997). Left-hemisphere specialization for the processing of acoustic transients. *NeuroReport*, *8*(7), 1761–1765.

Jungers, M., & Hupp, J. (2009). Speech priming: Evidence for rate persistence in unscripted speech. *Language and Cognitive Processes*, *24*(4), 611–624.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*(1), 133–156.

Kamide, Y., Scheepers, C., & Altmann, G. T. M. (2003). Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from German and English. *Journal of Psycholinguistic Research*, *32*(1), 37–55.

Kasl, S. V., & Mahl, G. F. (1965). Relationship of disturbances and hesitations in spontaneous speech to anxiety. *Journal of Personality and Social Psychology*, *1*(5), 425–433.

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

Kendon, A. (1978). Looking in conversation and the regulation of turns at talk: A comment on the papers of G Beattie and D R Rutter et al. *British Journal of Social and Clinical Psychology*, *17*(1), 23–24.

Kidd, C., White, K. S., & Aslin, R. N. (2011). Toddlers use speech disfluencies to predict speakers' referential intentions. *Developmental Science*, *14*(4), 925–934.

Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Research Reviews*, *29*(2), 169–195.

Knoeferle, P., Crocker, M. W., Scheepers, C., & Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: evidence from eye-movements in depicted events. *Cognition*, *95*(1), 95–127.

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., & Den, Y. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. *Language and Speech*, *41*(3-4), 295–321.

Kraljic, T., & Brennan, S. E. (2005). Prosodic disambiguation of syntactic structure: For the speaker or for the addressee? *Cognitive Psychology*, *50*(2), 194–231.

Kuriki, S., Mori, T., & Hirata, Y. (1999). Motor planning center for speech articulation in the normal human brain. *NeuroReport*, *10*(4), 765–769.

Kutas, M., Delong, K. A., & Smith, N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. In M. Bar (Ed.), *Predictions in the brain: Using our past to generate a future* (pp. 190–207). New York, NY: Oxford Univ Press.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, *62*, 621–647.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163.

Lallgee, M. G., & Cook, M. (1969). An experimental investigation of the function of filled pauses in speech. *Language and Speech*, *12*(1), 24–28.

Lau, E. F., & Ferreira, F. (2005). Lingering effects of disfluent material on comprehension of garden path sentences. *Language and Cognitive Processes*, *20*(5), 633–666.

Lay, C. H., & Paivio, A. (1969). The effects of task difficulty and anxiety on hesitations in speech. *Canadian Journal of Behavioural Science*, *1*(1), 25–37.

Leavens, D. A., Hopkins, W. D., & Bard, K. A. (1996). Indexical and referential pointing in chimpanzees (*Pan troglodytes*). *Journal of Comparative Psychology*, *110*(4), 346–353.

Leavens, D. A., Hopkins, W. D., & Thomas, R. K. (2004). Referential communication by chimpanzees (*Pan troglodytes*). *Journal of Comparative Psychology*, *118*(1), 48–57.

Leavens, D. A., Russell, J. L., & Hopkins, W. D. (2005). Intentionality as measured in the persistence and elaboration of communication by chimpanzees (*Pan troglodytes*). *Child Development*, *76*(1), 291–306.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition*, *14*(1), 41–104.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* London, England: The MIT press.

Levelt, W. J. M., & Kelter, S. (1982). Surface form and memory in question answering. *Cognitive Psychology*, *14*(1), 78–106.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(01), 1–38.

Levin, H., & Lin, T. (1988). An accommodating witness. *Language & Communication*, *8*(3), 195–197.

Levin, H., Silverman, I., & Ford, B. L. (1967). Hesitations in children's speech during explanation and description. *Journal of Verbal Learning and Verbal Behavior*, *6*(4), 560–564.

Levitan, R., & Hirschberg, J. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association* (pp. 3081–3084).

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

Lickley, R. J. (1995). Missing disfluencies. In *Proceedings of International Congress of Phonetic Sciences* (Vol. 4, pp. 192–195).

Lickley, R. J. (1998). *HCRC disfluency coding manual* (Tech. Rep. No. 100). HCRC. Retrieved from `http://www.ling.ed.ac.uk/ robin/maptask/disfluency-coding.html`

Lickley, R. J. (2001). Dialogue moves and disfluency rates. In *Disfluency in Spontaneous Speech (DiSS'01), ISCA Tutorial and Research Workshop (ITRW).*

Lickley, R. J., & Bard, E. G. (1996). On not recognizing disfluencies in dialog. In *Proceedings of the International Conference on Spoken Language Processing* (pp. 1876–1879). Philadelphia, PA.

Lickley, R. J., & Bard, E. G. (1998). When can listeners detect disfluency in spontaneous speech? *Language and Speech*, *41*(2), 203–226.

Local, J., & Kelly, J. (1986). Projection and "silences": Notes on phonetic and conversational structure. *Human Studies*, *9*(2), 185–204.

MacGregor, L. J. (2008). *Disfluency affect language comprehension: evidence from event-related potentials and recognition memory.* Unpublished doctoral dissertation, University of Edinburgh.

MacGregor, L. J., Corley, M., & Donaldson, D. I. (2009). Not all disfluencies are are equal: The effects of disfluent repetitions on language comprehension. *Brain and Language*, *111*(1), 36–45.

MacGregor, L. J., Corley, M., & Donaldson, D. I. (2010). Listening to the sound of silence: Investigating the consequences of disfluent silent pauses in speech for listeners. *Neuropsychologia*, *48*, 3982–3992.

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

MacNeilage, P. F. (1998). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, *21*(4), 499–511.

MacNeilage, P. F., & Davis, B. L. (2001). Motor mechanisms in speech ontogeny: phylogenetic, neurobiological and linguistic implications. *Current Opinion in Neurobiology*, *11*(6), 696–700.

Magyari, L., & De Ruiter, J. P. (2008). Timing in conversation: the anticipation of turn endings. In *Proceedings of the 12th Workshop on the Semantics and Pragmatics of Dialogue (Londial 2008)* (pp. 139–146).

Mar, R. A. (2004). The neuropsychology of narrative: story comprehension, story production and their interrelation. *Neuropsychologia*, *42*(10), 1414–1434.

Marquardt, D. W., & Snee, R. D. (1975). Ridge regression in practice. *The American Statistician*, *29*(1), pp. 3–20.

Martin, J. G. (1970). On judging pauses in spontaneous speech. *Journal of Verbal Learning and Verbal Behavior*, *9*(1), 75–78.

Martin, J. G., & Strange, W. (1968a). Determinants of hesitations in spontaneous speech. *Journal of Experimental Psychology*, *76*(3, Pt.1), 474–479.

Martin, J. G., & Strange, W. (1968b). The perception of hesitation in spontaneous speech. *Attention, Perception, & Psychophysics*, *3*(6), 427–438.

Matarazzo, J. D., Weitman, M., Saslow, G., & Wiens, A. N. (1963). Interviewer influence on durations of interviewee speech. *Journal of Verbal Learning and Verbal Behavior*, *1*(6), 451–458.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models.* London: Chapman & Hall/CRC.

McFarland, D. H. (2001). Respiratory markers of conversational interaction. *Journal of Speech, Language and Hearing Research*, *44*(1), 128–143.

McKelvie, D. (1998). *SDP – Spoken Dialogue Parser* (Tech. Rep. No. 96). University of Edinburgh: HCRC. Retrieved from `http://www.hcrc.ed.ac.uk/publications/`

Menenti, L., Gierhan, S. M. E., Segaert, K., & Hagoort, P. (2011). Shared language overlap and segregation of the neuronal infrastructure for speaking and listening revealed by functional mri. *Psychological Science*, *22*(9), 1173–1182.

Meyer, A. S., Belke, E., Häcker, C., & Mortensen, L. (2007). Use of word length information in utterance planning. *Journal of Memory and Language*, *57*(2), 210–231.

Miall, R. C., & Wolpert, D. M. (1996). Forward models for physiological motor control. *Neural Networks*, *9*(8), 1265–1279.

Miller, R. M., Sanchez, K., & Rosenblum, L. D. (2010). Alignment to visual speech information. *Attention, Perception, & Psychophysics*, *72*(6), 1614–1625.

Morris, R. K., & Folk, J. R. (1998). Focus as a contextual priming mechanism in reading. *Memory & Cognition*, *26*(6), 1313–1322.

Nakatani, C., & Traum, D. (1999). *Coding Discourse Structure in Dialogue* (Tech. Rep. No. 99-03). University of Maryland Institute for Advanced Computer Studies.

Natale, M. (1975a). Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology*, *32*(5), 790–804.

Natale, M. (1975b). Social desirability as related to convergence of temporal speech patterns. *Perceptual and Motor Skills*, *40*(3), 827–830.

Nicholson, H. B. M. (2007). *Disfluency in dialogue: attention, structure and function.* Unpublished doctoral dissertation, University of Edinburgh.

Nicholson, H. B. M., Bard, E. G., Lickley, R., Anderson, A. H., Mullin, J., Kenicer, D., & Smallwood, L. (2003). The intentionality of disfluency: Findings from feedback and timing. In *Disfluency in Spontaneous Speech (DiSS'03), ISCA Tutorial and Research Workshop* (pp. 17–20).

O'Connell, D. C., & Kowal, S. (2005). Uh and um revisited: Are they interjections for signaling delay? *Journal of Psycholinguistic Research*, *34*, 555–576.

O'Connell, D. C., Kowal, S., & Kaltenbacher, E. (1990). Turn-taking: A critical analysis of the research tradition. *Journal of Psycholinguistic Research*, *19*(6), 345–373.

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

O'Regan, J. K., Deubel, H., Clark, J. J., & Rensink, R. A. (2000). Picture changes during blinks: Looking without seeing and seeing without looking. *Visual Cognition*, *7*(1-3), 191–211.

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

Oviatt, S. L., & Cohen, P. R. (1991). Discourse structure and performance efficiency in interactive and non-interactive spoken modalities. *Computer Speech and Language*, *5*(4), 297–326.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, *119*, 2382-2393.

Pardo, J. S., Gibbons, R., Suppes, A., & Krauss, R. M. (2012). Phonetic convergence in college roommates. *Journal of Phonetics*, *40*(1), 190–197.

Pardo, J. S., Jay, I. C., Hoshino, R., Hasbun, S. M., Sowemimo-Coker, C., & Krauss, R. M. (2013). Influence of role-switching on phonetic convergence in conversation. *Discourse Processes*, *50*(4), 276–300.

Pardo, J. S., Jay, I. C., & Krauss, R. M. (2010). Conversational role influences speech imitation. *Attention, Perception, & Psychophysics*, *72*(8), 2254–2264.

Pardo, J. S., Jordan, K., Mallari, R., Scanlon, C., & Lewandowski, E. (2013). Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. *Journal of Memory and Language*, *69*(3), 183–195.

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*(3), 545–554.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*, 169–190.

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, *11*(3), 105–110.

Pickering, M. J., & Garrod, S. C. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), 329–47.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York, NY: Springer.

Plauché, M., & Shriberg, E. E. (1999). Data-driven subclassification of disfluent repetitions based on prosodic features. In *Proceedings of International Congress of Phonetic Sciences* (Vol. 2, pp. 1513–1516).

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as "asymmetric sampling in time". *Speech Communication*, *41*(1), 245–255.

Pulvermüller, F. (2010). Brain-language research: Where is the progress? *Biolinguistics*, *4*(2-3), 255–288.

Pulvermüller, F., & Fadiga, L. (2010). Active perception: sensorimotor circuits as a cortical basis for language. *Nature Reviews Neuroscience*, *11*(5), 351–360.

Pulvermüller, F., Huss, M., Kherif, F., del Prado Martin, F. M., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences*, *103*(20), 7865–7870.

Putman, W. B., & Street, R. L. (1984). The conception and perception of noncontent speech performance: Implications for speech-accommodation theory. *International Journal of the Sociology of Language*, *1984*(46), 97–114.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Rayner, K., Well, A. D., & Pollatsek, A. (1980). Asymmetry of the effective visual field in reading. *Attention, Perception, & Psychophysics*, *27*(6), 537–544.

Rayner, K., Well, A. D., Pollatsek, A., & Bertera, J. H. (1982). The availability of useful information to the right of fixation in reading. *Attention, Perception, & Psychophysics*, *31*(6), 537–550.

Reitter, D., Moore, J. D., & Keller, F. (2006). Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society.*

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, *8*(5), 368–373.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (2000). On the failure to detect changes in scenes across brief interruptions. *Visual Cognition*, *7*(1-3), 127–145.

Reynolds, A., & Paivio, A. (1968). Cognitive and emotional determinants of speech. *Canadian Journal of Psychology*, *22*(3), 164–175.

Ribeiro Jr., P. J., & Diggle, P. J. (2001). geoR: a package for geostatistical analysis. *R-NEWS*, *1*(2), 15–18. Retrieved from `http://cran.r-project.org/doc/Rnews`

Ross, A. S. C. (1954). Linguistic class-indicators in present-day english. *Neuphilologische Mitteilungen*, *55*, 113–149.

Rutter, D. R., Stephenson, G., Ayling, K., & White, P. (1978). The timing of looks in dyadic conversation. *British Journal of Social and Clinical Psychology*, *17*(1), 17–21.

Sachs, J. S. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Attention, Perception, & Psychophysics*, *2*(9), 437–442.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, *50*(4), 696–735.

Sanford, A. J. S., Sanford, A. J., Filik, R., & Molle, J. (2005). Depth of lexical-semantic processing and sentential load. *Journal of Memory and Language*, *53*(3), 378–396.

Sanford, A. J. S., Sanford, A. J., Molle, J., & Emmott, C. (2006). Shallow processing and attention capture in written and spoken discourse. *Discourse Processes*, *42*(2), 109–130.

Schachter, S., Christenfeld, N., Ravina, B., & Bilous, F. (1991). Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology*, *60*, 362–367.

Schachter, S., Rauscher, F., Christenfeld, N., & Tyson Crone, K. (1994). The vocabularies of academia. *Psychological Science*, *5*, 37–41.

Schnadt, M. J. (2009). *Lexical influences on disfluency production.* Unpublished doctoral dissertation, University of Edinburgh.

Schnadt, M. J., & Corley, M. (2006). The influence of lexical, conceptual and planning based factors on disfluency production. In *Proceedings of the twenty-eighth meeting of the Cognitive Science Society.* Vancouver, Canada.

Schober, M. F., & Brennan, S. E. (2003). Processes of interactive spoken discourse: The role of the partner. In A. C. Graesser, M. A. Gernsbacher, & S. R. Goldman (Eds.), *Handbook of discourse processes* (pp. 123–164). Mahwah, NJ: Lawrence Erlbaum Associates.

Schroeder, C. E., & Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. *Trends in Neurosciences*, *32*(1), 9–18.

Scott, S. K., McGettigan, C., & Eisner, F. (2009). A little more conversation, a little less action-candidate roles for the motor cortex in speech perception. *Nature Reviews Neuroscience*, *10*(4), 295–302.

Sereno, S. C., & Rayner, K. (1992). Fast priming during eye fixations in reading. *Journal of Experimental Psychology: Human, Perception, and Performance*, *18*(1), 173–184.

Seyfeddinipur, M., Kita, S., & Indefrey, P. (2008). How speakers interrupt themselves in managing problems in speaking: Evidence from self-repairs. *Cognition*, *108*(3), 837–842.

Shannon, C. E. (1951). Prediction and entropy of printed english. *Bell System Technical Journal*, *30*(1), 50–64.

Shockley, K., Baker, A. A., Richardson, M. J., & Fowler, C. A. (2007). Articulatory constraints on interpersonal postural coordination. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(1), 201–208.

Shockley, K., Sabadini, L., & Fowler, C. A. (2004). Imitation in shadowing words. *Perception & Psychophysics*, *66*(3), 422–429.

Shockley, K., Santana, M.-V., & Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human, Perception, and Performance*, *29*(2), 326–332.

Shriberg, E. E. (1994). *Preliminaries to a theory of speech disfluencies.* Unpublished doctoral dissertation, University of California at Berkley.

Shriberg, E. E. (1996). Disfluencies in Switchboard. In *Proceedings of the International Conference on Spoken Language Processing, Addendum* (pp. 11–14). Philadelphia, PA.

Shriberg, E. E. (1999). Phonetic consequences of speech disfluency. In *Proceedings of the International Congress of Phonetic Sciences* (Vol. 1, pp. 619–622). San Francisco.

Shriberg, E. E., & Stolcke, A. (1996). Word predictability after hesitations: A corpus-based study. In *Proceedings of the Fourth International Conference on Spoken Language Processing* (Vol. 3, pp. 1868–1871).

Siegel, G. M., Lenske, J., & Broen, P. (1969). Suppression of normal speech disfluencies through response cost. *Journal of Applied Behavior Analysis*, *2*(4), 265–276.

Siegman, A. W., & Pope, B. (1965). Effects of question specificity and anxiety-producing messages on verbal fluency in the initial interview. *Journal of Personality and Social Psychology*, *2*(4), 522–530.

Smith, V. L., & Clark, H. H. (1993). On the course of answering questions. *Journal of Memory and Language*, *32*, 25–38.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human, Learning, and Memory*, *6*, 174–215.

Stephens, G. J., Silbert, L. J., & Hasson, U. (2010). Speaker–listener neural coupling underlies successful communication. *Proceedings of the National Academy of Sciences*, *107*(32), 14425–14430.

Stephens, J., & Beattie, G. (1986). Turn-taking on the telephone: Textual features which distinguish turn-final and turn-medial utterances. *Journal of Language and Social Psychology*, *5*(3), 211–222.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., . . . Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*(26), 10587–10592.

Street, R. L. (1982). Evaluation of noncontent speech accommodation. *Language & Communication*, *2*(1), 13–31.

Street, R. L. (1984). Speech convergence and speech evaluation in fact-finding interviews. *Human Communication Research*, *11*(2), 139–169.

Street, R. L., Brady, R. M., & Putman, W. B. (1983). The influence of speech rate stereotypes and rate similarity or listeners' evaluations of speakers. *Journal of Language and Social Psychology*, *2*(1), 37–56.

Sturt, P., Sanford, A. J., Stewart, A., & Dawydiak, E. (2004). Linguistic focus and good-enough representations: An application of the change-detection paradigm. *Psychonomic Bulletin & Review*, *11*(5), 882–888.

Svartik, K., & Quirk, R. (1980). *A corpus of english conversation.* Lund, Sweden: Gleerup.

Swerts, M. (1997). Prosodic features at discourse boundaries of different strength. *The Journal of the Acoustical Society of America*, *101*, 514–521.

Swerts, M. (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics*, *30*(4), 485–496.

Szekely, A., Jacobsen, T., D'Amico, S., Devescovi, A., Andonova, E., Herron, D., ... others (2004). A new on-line resource for psycholinguistic studies. *Journal of Memory and Language*, *51*, 247–250.

Taboada, M. (2010). Spontaneous and non-spontaneous turn-taking. *Pragmatics*, *16*(2), 329–360.

Tauroza, S., & Allison, D. (1990). Speech Rates in British English. *Applied Linguistics*, *11*(1), 90–105.

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, *30*, 415–433.

Ten Bosch, L., Oostdijk, N., & Boves, L. (2005). On temporal aspects of turn taking in conversational dialogues. *Speech Communication*, *47*(1), 80–86.

Tognoli, E., Lagarde, J., DeGuzman, G. C., & Kelso, J. a. S. (2007). The phi complex as a neuromarker of human social coordination. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(19), 8190–8195.

Trueswell, J. C., & Kim, A. E. (1998). How to prune a garden path by nipping it in the bud: Fast priming of verb argument structure. *Journal of Memory and Language*, *39*, 102–123.

Verhoeven, J., De Pauw, G., & Kloots, H. (2004). Speech rate in a pluricentric language: A comparison between Dutch in Belgium and the Netherlands. *Language and Speech*, *47*(3), 297–308.

Vousden, J. I., Brown, G. D. A., & Harley, T. A. (2000). Serial control of phonology in speech production: A hierarchical model. *Cognitive Psychology*, *41*(2), 101–175.

Walker, M. B., & Trimboli, C. (1982). Smooth transitions in conversational interactions. *The Journal of Social Psychology*, *117*(2), 305–306.

Walker, M. B., & Trimboli, C. (1984). The role of nonverbal signals in coordinating speaking turns. *Journal of Language and Social Psychology*, *3*(4), 257–272.

Ward, A., & Litman, D. (2007). Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora. In *Speech and Language Technology in Education (SLaTE 2007)* (pp. 57–60).

Ward, L. M. (2003). Synchronous neural oscillations and cognitive processes. *Trends in Cognitive Sciences*, *7*(12), 553–559.

Ward, P., & Sturt, P. (2007). Linguistic focus and memory: An eye movement study. *Memory & Cognition*, *35*(1), 73–86.

Warren, T., & Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition*, *85*(1), 79–112.

Watanabe, M., Den, Y., Hirose, K., & Minematsu, N. (2004). Clause types and filled pauses in japanese spontaneous monologues. In *Proceedings of the 8th International Conference on Spoken Language Processing* (pp. 905–908).

Watkins, K. E., Strafella, A. P., & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia*, *41*(8), 989–994.

Webb, J. T. (1969). Subject speech rates as a function of interviewer behaviour. *Language and Speech*, *12*(1), 54–67.

Wennerstrom, A., & Siegel, A. F. (2003). Keeping the floor in multiparty conversations: Intonation, syntax, and pause. *Discourse Processes*, *36*(2), 77–107.

Whiteside, S. P. (1996). Temporal-based acoustic-phonetic patterns in read speech: Some evidence for speaker sex differences. *Journal of the International Phonetic Association*, *26*(1), 23–40.

Wicha, N. Y. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in spanish sentence reading. *Journal of Cognitive Neuroscience*, *16*(7), 1272–1288.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America*, *91*, 1707–1717.

Wilson, M., & Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review*, *12*(6), 957–968.

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience*, *7*(7), 701–702.

Wilson, T. P., Wiemann, J. M., & Zimmerman, D. H. (1984). Models of turn taking in conversational interaction. *Journal of Language and Social Psychology*, *3*(3), 159–183.

Wilson, T. P., & Zimmerman, D. H. (1986). The structure of silence between turns in two-party conversation. *Discourse Processes*, *9*(4), 375–390.

Włodarczak, M., Juraj, S., & Wagner, P. (2012). Syllable-boundary effect: temporal entrainment in overlapped speech. In *Proceedings of Speech Prosody 2012* (p. 611-614).

Wolpert, D. M., Doya, K., & Kawato, M. (2003, March). A unifying computational framework for motor control and social interaction. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *358*(1431), 593–602.

Yngve, V. H. (1970). On getting a word in edgewise. In *Chicago Linguistics Society, 6th Meeting* (pp. 567–578).

Yuan, J., Liberman, M., & Cieri, C. (2006). Towards an integrated understanding of speaking rate in conversation. In *INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing* (pp. 541–544).

Yuan, J., Liberman, M., & Cieri, C. (2007). Towards an integrated understanding of speech overlaps in conversation. *The 16th International Congress of Phonetic Sciences*, 1337–1340.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162–185.