DISFLUENCY:

INTERRUPTING SPEECH AND GESTURE

Cover design: Linda van den Akker & Inge Doehring
Cover illustration: Claudia Renzler
Printed and bound by Ponsen & Looijen bv, Wageningen
ISBN 90-76203-25-3

DISFLUENCY:
INTERRUPTING SPEECH AND GESTURE

een wetenschappelijke proeve
op het gebied van Letteren

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de Rector Magnificus prof. dr. C.W.P.M. Blom,
volgens besluit van het College van Decanen
in het openbaar te verdedigen op maandag 6 juni 2006
des namiddags om 3.30 uur precies

door
Mandana Seyfeddinipur
geboren op 31 mei 1967
te Offenbach, Duitsland

Promotor:	Prof. dr. S. C. Levinson
Co-promotores:	Dr. P. Indefrey (MPI)
	Dr. S. Kita (University of Bristol, UK)
Manuscriptcommissie:	Prof. dr. W. Vonk
	Dr. R. J. Hartsuiker (University of Gent, Belgium)
	Dr. A. Özyürek (MPI)

The research reported in this thesis was supported by a grant from the Max-Planck-Gesellschaft zur Förderung der Wissenschaften, München, Germany.

Preface

I would like to thank my promotor Stephen C. Levinson, and my supervisors Sotaro Kita and Peter Indefrey for their guidance. Stephen C. Levinson gave me the opportunity to work in an extraordinary scientific environment. Sotaro Kita introduced me with incredible politeness to the world of quantitative research and statistics. Peter Indefrey, with his clarity and logical thinking, was invaluable in guiding this project to completion.

My special thank goes to Adam Kendon and Oscar Gatto. Oscar Gatto enabled many joyful Tete a Tete's, and provided me with cheerful funny poetry that kept me going. In many hours of discussions of data and theory, many dinners at India Gate and elsewhere, Adam Kendon always supported and encouraged me not to lose sight of the importance of basic research and the need for a gesture phonology.

I am indebted to Marianne Gullberg, Alissa Melinger and Simone Sprenger who have been amazing friends and great colleagues. Marianne Gullberg not only discussed data, read, commented and dry ran, she also provided many cathartic sessions, and the opportunity to live in the Echo. Thanks to Alissa Melinger who was always there to help and to rally my spirits in the most funny way even when discussing stats. I am grateful to Simone Sprenger for so many things like reading, writing, and cooking just to name a few. Thank you for your friendship and for your support.

Herb Clark, David McNeill, Cornelia Müller, and Asli Özyürek, Manny Schegloff and David Wilkins helped shape my thinking about gesture and disfluency through many stimulating discussions. I am grateful for the many insights that these discussions have brought me.

My thanks also go to the great staff at the MPI. Angela Heuts, Perry Janssen, Paul Lommen, and Norbert Nielen were always courteous and never tired of answering all my questions about German and Dutch forms. I received great technical support from

Reiner Dirksmeier, Gerd Klaas, Paul Trilsbek, Ad Verbunt, Rick van Viersen, and Jethro van Zevenbergen, who patiently made sense out of my requests for help. Agnes and José, Pim, Jan, and Hans were always there with a friendly word and often, much needed comic relief.

I am very grateful to my friends Claudia Renzler, who designed the beautiful cover, Matthias Plettenberg for making pictures and Lutz Kaulfuss who unexpectedly found himself helping with the final touches of the manuscript. Thanks to Lea Hald and Leah Roberts for proofreading the manuscript and to Simone Sprenger again, this time for the Dutch translation.

I feel lucky to have made so many friends along the way, among them: Birgit Hellwig, Anna and Andrew Margetts, Sonja Eisenbeiss, Friederike Luepke, Juergen Bohnemeyer, David Wilkins, Frank Wiersma, Pamela Perniss, Frank Seifart and Federico Rossano, Edith Sjoerdsma, Melissa Bowerman, Gunter Senft and the Burenhults. Working on weekends and in the evenings at the Institute was stimulating and enjoyable because of them. Throughout the years everyone in his/her own special way gave me moral and practical support, in the form of hospital care packages and phone calls, many dinners and coffee breaks, shopping sprees and great parties, and white board sessions and dry runs, just to name a few. Special thanks to my friends and paranimfen Pamela Perniss and Federico Rossano who are masters in making things fun. Thank you all for all of that and for much much more!

Oliver Müller, Simone Sprenger and Hedderik van Rijn have been there from the beginning of my time at MPI, and I cherish the warm and close friendship that has developed between us through many dinners, movies, cups of coffee at de Blonde Pater, and commiseration sessions. Especially during the last year, they helped keep me grounded by distracting me, making me laugh, and giving me support when I needed it. See you in the sun among the orange trees!

I also want to thank my second family members Kati Buchwald, Lutz Kaulfuss, Dominik Peter, Claudia Renzler, Heiko Müller, Schwester Helga und Horst Schlemmer. Thank you for always being near, although far away.

And of course I thank the Herr's...Ricarda, and my siblings Carmen, Julia and Jonas, who gave me a second home and with whom it was always such a joy to play with and to forget about the rest.

Finally, my wonderful and amazing mother carried me through with her unconditional loving support and her deep trust in me. Thank you for being there all the time.

And Dale ... awesome!

Contents

1. General Introduction

Chapter 1

1.1 Introduction

In everyday conversation speech disfluencies are ubiquitous and come in various forms. Speakers hesitate in the form of silent pauses or elements like 'uh' and 'um'. They search for words, utter wrong or inappropriate words and expressions or sentences that are ungrammatical. Disfluencies are pervasive in conversation because putting ideas into words is a complex process that must proceed through stages of conceptual, syntactic, morphological, and phonological encoding before articulation can start. At every stage of encoding, things can and do go wrong. If an error is detected, speakers can interrupt processing and speaking to correct the error. In order to do so speakers must have an effective system of self-monitoring.

The ability to self-monitor speech prior to articulation and to anticipate problems in understanding is a prerequisite of intersubjectivity (Schegloff, 1992), the achievement and maintenance of common understanding throughout talk in interaction. Speakers design their talk within a given context for their recipients. Nevertheless, mutual understanding is not a given, but is negotiated. Speakers display their understanding sequentially during the course of interaction. The ability to self-monitor what is going to be said enables the speaker to correct a problematic expression prior to its articulation and thereby avoid possible misunderstanding or misinterpretation by the recipient. When misunderstandings or non-understandings arise, they are quite costly to fix, requiring an exchange of turns inserted into the main business at hand. It is much more efficient to avoid such distractions, hence the premium on immediate self-repair, made possible by self-monitoring.

Given the importance of self-monitoring for successful conversation, a long-standing goal of psycholinguistic research has been to understand the cognitive mechanisms by which self-monitoring is realized. This includes the processes involved in the detection of errors and mechanisms for the suspension and resumption of speech (see Figure 1.1 below for an illustration of the time points in disfluent utterances labeled with these terms). Efforts in this direction have led to the emergence of well-specified, empirically testable theories of self-monitoring (see Chapter 2). These theories assume that the 'self-monitor' functions to enable speakers to accurately express their intentions through speech. However, interactional approaches to language use suggest that in face-to-face conversation the speaker must deal with broader concerns than simply avoiding or correcting errors in speech. The general aim of the current work is to elaborate these cognitive theories of self-monitoring by taking a broader view on the tasks that the self-monitor is charged with in conversation.

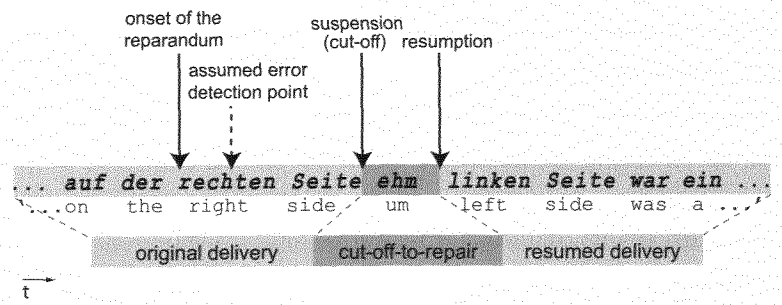

Figure 1.1. Schematic representation of the structure of a disfluent utterance. In the given example an erroneous word, the so-called reparandum, *rechten* ('right') was uttered. A possible point of error detection might have been after the first syllable *rech* ('ri') of the reparandum. The original delivery is cut off after the word *Seite* ('side') (suspension). After suspension, a time interval (cut-off-to-repair) might follow during which speakers might, for example, pause or utter 'uh/um'. Sometimes there is no such interval but speakers resume their delivery immediately. The moment the delivery is resumed is called resumption.

The way talk unfolds in time is constrained by the resources and the limitations of speech production and the striving for accurate expression. However, speakers must balance this demand for accuracy against other demands brought about by the conversational situation, such as producing utterances in a timely fashion (Clark 1996; 2002). If speakers suspend speech and pause, they risk losing the floor and appearing

ineloquent. This demands a self-monitoring system that can operate in parallel to processes of utterance planning and articulation and that can balance the demand for fluency with the demand for accuracy. When speakers detect trouble, they must decide: 1) whether to correct; 2) whether to interrupt their speech; 3) when to interrupt; 4) how to interrupt; and 5) how to correct. How do speakers balance the demands of accuracy and fluency in dealing with errors in their speech production? Do speakers interrupt speech immediately in order to avoid uttering erroneous information that could mislead the listener? Or is the interruption of speech delayed in order to maintain fluency and to repair as fast as possible? This one set of questions addressed in this thesis.

A second demand that arises in face-to-face conversation derives from the multimodal nature of the speaker's communicative performance. In face-to-face interaction interlocutors can see as well as hear one another, and therefore can communicate via the visual as well as the auditory channel. A primary way in which this visual communication is effected is through gesture, movements of the arms and hands that typically accompany speech. Gesture is temporally and semantically closely coordinated with speech, (Kendon, 1983; McNeill, 1992) and has been shown to have an impact on listeners' interpretation of, and memory for, speech (Beattie & Shovelton, 1999; Cassell, McNeill, & McCullough, 1999; Graham & Heywood, 1975; Kelly, Barr, Breckinridge Church, & Lynch, 1999). It has also been shown that gestures are communicatively intended by the speaker (Cohen & Harrison, 1973; Melinger & Levelt, 2004) and that speakers design their gestures for their addressees (Bavelas, Kenwood, Johnson, & Phillips, 2002; Gerwing & Bavelas, 2005; Özyürek, 2002). Indeed, Kendon (2004, p. 127) argues, "In creating an utterance that uses both modes of expression the speaker creates an ensemble in which gesture and speech are employed together as partners in a single rhetorical enterprise." Clark (1996) offers a similar view, namely that gesture and speech are but two channels of a single composite signaling system, wherein the information that is to be communicated is distributed across modalities. For these reasons, speakers who experience fluency problems in the speech modality must not only make decisions about what to do about their speech but also about their gesture; namely, whether and when to stop it. However, very little is currently known about what happens to gesture during speech disfluency.

A further reason for investigating gestural behavior during speech disfluency is that it has the potential to provide additional insights into the processes underlying speech suspension. Gestural movements precede specific events in speech (Chui, 2005; Kendon, 2004; McNeill, 1992; Morrel-Samuels & Krauss, 1992). This temporal relationship might potentially be informative with regard to the questions as to when speakers interrupt their speech and what happens to the associated gesture.

In sum, when speakers detect an error in their speech, they are faced with a complex set of decisions that they must make in order to balance various demands of communicative performance in face-to-face conversation. These demands include not only the need for accuracy, but also the need to maintain fluency and to maintain temporal and semantic coordination with gesture. As an approach to understanding how this balance is achieved, the present work examines the following two sets of questions:

Do speakers interrupt their speech immediately upon error detection in order to avoid uttering erroneous information, or is the interruption a planned process that is based on strategic decisions with respect to manner and timing of interruption and correction, such that fluency can be maintained?

What happens to a speaker's gesture when speech is suspended? What does this imply for gesture-speech coordination, and can it provide insight into the processes underlying self-monitoring and the interruption of speech?

These questions are addressed in this dissertation through a collection of a corpus of semi-natural conversational data and analyses of speech disfluencies[1] and co-occurring gestures within it.

[1] "Disfluency" has been referred to in the literature in many different ways (see Lickley, 1994). Classification schemes are abundant and vary greatly (e.g., Bear, Dowding, & Shriberg, 1992; Blackmer & Mitton, 1991; Hieke, 1981; Levelt, 1983; Maclay & Osgood, 1959; Postma & Kolk, 1992). I will use the term disfluency following Shriberg (1994, p.1) who considers "disfluencies as cases in which a contiguous stretch of linguistic material must be deleted to arrive at the sequence a speaker 'intended', likely the one that would be uttered upon a request for repetition." This also includes so-called filled pauses like *uh* and *uhm*, which are seen as linguistic elements (Clark & Fox Tree, 2002; Shriberg, 1994).

The dissertation is divided into two main sections, the first of which addresses how speakers trade off accuracy and fluency by relating the structure of a speech disfluency to the complexity of the processing that must be undertaken to correct the error. The second study addresses the multimodality of the speaker's performance by examining the gestural response to speech disfluency. In the remainder of the current chapter, I will provide further motivation for these two studies, introduce the data set on which the analyses are based, and conclude with an overview of the structure of the dissertation.

1.1.1 Interrupting speech

As noted above, there is a potential tension between the demands of accuracy and speed in speech production. As a consequence, a crucial question for monitoring theories is what happens when the monitor detects an erroneous expression that is about to be articulated or that is already under articulation. Two competing hypotheses have been suggested, which we term: The Main-Interruption-Rule hypothesis (MIR hypothesis) and the Delayed-Interruption-for-Planning hypothesis (DIP hypothesis).

Levelt (1983) suggested the Main Interruption Rule (MIR). According to the MIR hypothesis, an interruption process of the entire speech production system is initiated immediately upon error detection, with a constant latency of 200 ms between sending a stop signal and the suspension of speech (Levelt, 1983, p. 56). Because the entire speech production system is halted, replanning or planning of the repair can only start after speech suspension. Note that I will refer to the observable time point of the stopping of the overt speech stream as *suspension*. I will refer to the internal process eventually leading to suspension as *interruption*. Interruption covers the time interval from stop signal until suspension referred to as *interruption latency*. The verb *to interrupt* is not used as a technical term.

Levelt (1983) sought evidence for the MIR hypothesis in a study in which Dutch participants described paths through a network of colored circles. A prediction of the MIR hypothesis is that linguistic boundaries (e.g., syllable, word, phrase) should not be respected in speech suspension. Although the majority of suspensions occurred at word boundaries, Levelt found that erroneous words (e.g., *red circle* vs. *blue circle*) were

more likely to be suspended within-word than words that were merely inappropriate. Examples of inappropriate words are words that are not specific enough, not the best choice, or words that have not been used consistently throughout the discourse; for example, *dot* vs. *circle*. Based on this finding, Levelt modified the MIR by suggesting that only 'true' errors are subject to the MIR. Inappropriate expressions and words that are not wrong following erroneous words (so-called *neutral words* (Levelt, 1983, p. 62) are not suspended within-word but after-word. Thus, Levelt suggests that the interruption of speech is a planned process based on the evaluation of the erroneousness of the reparandum and the moment of error detection. Levelt proposes that speakers signal the erroneousness of a word by suspending within-word. Likewise, by suspending after-word, speakers signal the correctness of that word. Levelt assumes that this is due to differences in the communicative status of erroneous expressions as compared to inappropriate expressions, which are merely not specific enough. An inappropriate expression is correct but may need further qualification, while an erroneous expression "has to be undone as soon as possible" (Levelt, 1983, p. 63). Hence, the motivation for the MIR hypothesis is pragmatic: speakers attempt to avoid uttering erroneous information.

The MIR hypothesis has not remained unchallenged. Studies testing temporal predictions of the MIR hypothesis have provided evidence that the erroneousness of the reparandum is not the only factor determining the moment of interruption. In this thesis, an alternative account of the process of speech interruption is proposed, the Delayed-Interruption-For-Planning hypothesis. This hypothesis is based on a suggestion by Blackmer and Mitton (1991) that the availability of the repair, rather than the erroneousness or the correctness of the suspension word, can be a determining factor for the initiation of speech interruption. Under the DIP hypothesis, as soon as an error is detected, speakers start planning the repair while they go on speaking as long as prepared material is in the formulator and the articulatory buffer. This way, speakers minimize the pause between speech suspension and repair. In other words, the demand for fluency overrides the demand for accuracy. If the repair processing is completed

while there is still material available in the articulatory buffer, speech interruption is initiated, regardless of whether the interruption results in a within-word or after-word suspension.

Support for this view comes from disfluency studies using interactional approaches. Interactional approaches to language use focus on how interlocutors handle errors and disfluencies in conversation. In this view, speaking is constrained by the affordances of the primary mode of speaking, namely social interaction (Clark, 1996; Sacks, 1992). The moment-by-moment organization of the interaction is taken as a primary determinant of how problematic speech is handled and accounted for.

Interactional approaches point out that in everyday conversation time is of the essence. Because conversation is a cooperative joint action, speakers should strive to make their contributions in a timely manner (Clark 1996). In addition, speakers also need to respect the rules of turn taking determining who speaks for how long and when (Sacks, Schegloff, & Jefferson, 1974). Studies in the tradition of Conversation Analysis have shown the orderliness of repair operations in talk-in-interaction (Schegloff, 1979, 1992; Schegloff, Jefferson, & Sacks, 1977). The preferred format of repairs in interaction is self-initiated self-repair within the turn in which the trouble source occurred (Schegloff et al., 1977). Repairs initiated by interlocutors are dispreferred. When speech is interrupted immediately upon error detection, the speaker has to pause to process the repair. During that time the interlocutor could potentially initiate a repair and even execute it, and thus the speaker risks both losing the floor and being corrected by the interlocutor. It is therefore in a speaker's interest to repair as fast as possible, so that the distance between the troublesome expression and the repair does not become too long and the repair is seen as a correction of that expression. By repairing as fast as possible, the speaker can also ensure that misinterpretations by the interlocutor do not happen and that a correction or clarification by the recipient will not become a relevant move. In this way the progression of the turn itself and of the sequential organization of turns within conversation will not be delayed (Schegloff, 1979).

Another explanation for why speakers may not interrupt immediately is motivated by a signaling account of disfluency. This signaling view has also provided an account

for the observed tendency of speakers to minimize pauses. Clark and Wasow (1998) argue that speakers often initiate speaking while expecting not to be able to complete the constituent they have started out with. They then suspend speech after the first word of the constituent and as they resume, they repeat the first word of the constituent, which is now produced fluently. Clark and Wasow (1998) call this a commit-and-repair strategy. Since speakers are under interactional time pressure, they have to justify the time they take speaking or pausing. Not starting to speak for too long might make them appear as: "...opting out, as confused or distracted, as uncertain about what they want to say, or as having nothing immediately to contribute" (Clark & Wasow, 1998, p. 238). By beginning the first word of a constituent, they show that they are in the process of planning and forestall these attributions.

Hence, speakers have different strategies at their disposal to handle problems in speaking and the demands of the interactional situation at the same time. Maintaining fluency and striving to repair covertly can be seen as one such strategy and as the pragmatic motivation of the DIP hypothesis.

Taken together, both accounts, the MIR hypothesis and the DIP hypothesis assume that speakers decide when to interrupt upon error detection. The MIR hypothesis assumes that speakers interrupt speech immediately upon error detection or decide to delay till the end of the word if the trouble word is merely inappropriate or if the suspension word is neutral (i.e., a correct word following the trouble word). They do so in order to signal the erroneousness or correctness of the suspension word. In contrast, the DIP hypothesis assumes that speakers interrupt only after the repair processing has been completed, in order to minimize the pause following suspension and to repair as fast as possible.

1.1.2 Gesture and speech disfluency

What happens to gesture during disfluency? A commonsensical view is that gesture is used to communicate when speech fails, when we stop speaking. For example, it is often assumed that second language (L2) speakers, are likely to attempt to communicate information through gesture instead of speech when they cannot find the appropriate words. However, studies in second language acquisition have shown that although L2

speakers gesture more than L1 speakers due to insufficient mastery of the second language, they do not gesture more during silences (Gullberg, 1998). More generally, research has shown that gestures are more prevalent during speaking than during hesitations (e.g., Beattie & Aboudan, 1994; McNeill, 1992; Nobe, 2000).

Speech and gesture are semantically and temporally intricately intertwined in that the meaningful movement part of a gesture is synchronized with the verbally co-expressive element in speech (Chui, 2005; Kendon, 1980, 1983; McNeill, 1985, 1992). For example, a gesture depicting the act of throwing might be synchronized with the co-expressive speech 'threw it' in 'he grabbed the ball and threw it into the window.' Observations such as these suggest that gesture and speech are cognitively linked in their production (De Ruiter, 1998; Feyereisen, 1997; Hadar & Butterworth, 1997; Kita & Özyürek, 2003; Krauss, Chen, & Chawla, 1996; McNeill, 1985). Further evidence for a critical link between speech and gesture can be derived from neuropsychological studies. These studies have shown that in aphasia gesture breaks down in parallel with language. Patients with Broca's aphasia, who suffer from grammatical deficits, mainly produce representational gestures depicting semantic content, while those suffering from Wernicke's aphasia, who display semantic deficits, mainly produce semantically empty beat gestures (Cicone, Wapner, Foldi, Zurif, & Gardner, 1979; Feyereisen & Lannoy, 1991; Pedelty, 1987). This kind of evidence has led to the view, that gestures are visual manifestations of the imagistic aspects of cognition, and thus effectively provide a window onto the mind of the speaker (McNeill, 1985, 1992).

In light of the close temporal and semantic coordination of gesture and speech, it is conceivable that a disfluency in speech is also reflected in gesture. Although studying the gestural response to speech disfluency is worthwhile in its own right, an additional motivation is that gesture might provide information about the moment of error detection and thus help resolve the debate about whether speakers interrupt their speech immediately or delay for planning the repair.

One reason why the controversy concerning the moment of speech interruption has remained unresolved is the inherent problem of observing the covert process of error detection. Because error detection cannot be inferred unambiguously from the temporal intervals on the surface structure of the disfluency like the time between

suspension and resumption (cut-off-to-repair interval) or the time between onset of the error to the moment of suspensions (error-to-cut-off interval), the moment of error detection has effectively served as a free parameter in the discussion as to when speech is interrupted (see Figure 1.2 below). For instance, it can be argued that, in cases where an erroneous word is completed and not interrupted within-word, the detection happened so late that the completion of the erroneous word could not be avoided. Conversely, it can also be argued that detection was early but interruption was delayed to complete the word under articulation. Moreover, even if an erroneous word was suspended within-word this does not have to be a result of immediate interruption upon error detection. It is possible that the interruption was delayed in order to minimize the resulting pause as assumed by the DIP hypothesis.

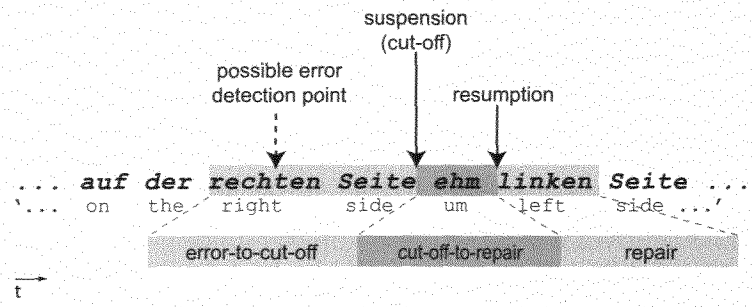

Figure 1.2. Schematic representation of the surface structure of a disfluent utterance. The interval from the onset of the erroneous word to the suspension (cut-off) is called the error-to-cut-off interval. The interval between suspension (cut-off) and resumption is called the cut-off-to-repair interval.

An index that is not based on events in speech and that is closer in time to the moment of error detection could provide further evidence as to whether speech interruption is initiated immediately or not. The gestural response to the detection of an error could function as such an index. One reason to expect an earlier response in gesture is that in fluent speech, gestures tend to temporally precede the semantically co-expressive elements in speech. In order to use gesture as an indicator of covert error detection, an operationalization of gestural responses to speech disfluencies will be developed that utilizes the temporal relationship of speech and gesture.

1.2 The data set

Many cognitive studies of speech disfluencies rely upon data that have been elicited in experimentally controlled laboratory tasks (e.g., Levelt, 1983; Oomen, 2001). On the one hand, experimentally controlled speech production data allow us to determine the erroneousness or inappropriateness of a given reparandum (the decisive factor for immediate or delayed interruption following the MIR hypothesis). On the other hand, laboratory tasks are often produced in social isolation (e.g., speakers describe patterns to a tape recorder), such that it is not clear if the full burden of interactional affordances comes into play in the same way as in everyday conversation. Also, eliciting gestures in a controlled task often limits the amount and the types of gestures that are deployed. For example, a task like the network descriptions used in Levelt (1983) does not yield high gestural activity and elicits mostly only one type of gesture, namely gestures tracing the paths through the network (see, e.g., Melinger & Levelt, 2005). A controlled task often used in gesture studies is cartoon retellings. Speakers retell a cartoon to an interlocutor, who has not seen it (McNeill, 1992). For this task the amount of co-occurrences of disfluencies and gestures seems to be limited (Kita, 1993).

Since the present study aimed to investigate the temporal characteristics of speech disfluencies as well as the temporal co-ordination of both gesture and speech disfluencies, it was necessary to elicit a considerable amount of gestures and speech disfluencies that closely approximate those observed in naturalistic dialogue. Given these considerations, we decided to base the investigation on semi-naturalistic data elicited in an interactional setting. Speakers described houses and apartments they live in or have lived in to an interlocutor. The disadvantage of such a semi-naturalistic data set is that speech production is not controlled, thereby limiting the kinds of inferences that may be drawn about the type of problem causing a given disfluency. For example, it is often not possible to determine if the cause of a repair was an erroneous or an inappropriate expression. However, we believe that the advantage of a naturalistic data set outweighs the resulting ambiguity of repairs.

A second type of data was collected in an experimental study made necessary by the results of the corpus study. In order to obtain data comparable to the corpus, the

same conversational task was employed. Speakers described houses and apartments to an interlocutor who was not familiar with the places described. Participants were presented with an auditory stop signal upon which they were supposed to stop speaking and gesturing. This procedure permitted the interruption of gesture and speech at different points during their execution and the collection of the corresponding suspension latencies.

1.3 Overview of the thesis

Chapter 2 provides a theoretical background for those issues concerning architecture and processing mechanisms in speech production and self-monitoring that will be raised in the following chapters. The fundamentals of an explicit model of speech production (Levelt, 1989; Levelt, Roelofs, & Meyer, 1999) will be outlined. Cognitive theories of self-monitoring, which have sought an understanding of the processes underlying error detection and correction, will be described.

Chapters 3 and 4 present results from two studies in which analyses were conducted on the data set described above.

In Chapter 3, evidence is sought which distinguishes between the Main Interruption Rule (MIR) and Delayed-Interruption-For-Planning (DIP) hypotheses by determining whether interruption of speech is determined by the moment of error detection or by repair readiness. Many studies have reported timing characteristics that fail to confirm the predictions of the MIR hypothesis (Blackmer & Mitton, 1991; Kormos, 2000; Oomen & Postma, 2001; Oomen & Postma, 2002; Van Hest, 1996). Yet it remains a controversy as to whether or not the MIR hypothesis can in fact account for these findings (Hartsuiker & Kolk, 2001). A primary reason why these studies have failed to resolve the issue is because they have not taken into account the varying degrees of complexity of the replanning process. If interruption follows immediately upon error detection as the MIR hypothesis states, the complexity of the replanning should be reflected in the duration of the interval between speech suspension and resumption (cut-off-to-repair interval, Blackmer & Mitton, 1991). Typologies of repairs,

such as those provided by Clark (1996) or Levelt, (1983) suggest a wide variety of complexity of repairs. They range from minor repairs, such as simple phoneme substitutions, to major repairs, such as the generation of an entirely new utterance. Depending on how many levels of the production system are involved in the generation of the repair, the replanning time should differ. According to the MIR hypothesis, such variations in the complexity of the replanning process should be reflected in the length of the cut-off-to-repair interval, with major repairs requiring more time than minor ones. According to the DIP hypothesis, in contrast, the replanning process can be partly masked by continued speech production. Thus, the cut-off-to-repair interval does not necessarily reflect the entire replanning time. The study in Chapter 3 investigates this relationship of repair complexity and the duration of the cut-off-to-repair interval.

Chapter 4 investigates gestures in disfluent utterances. The goal of this study is twofold: (1) to examine whether gesture is sensitive to speech disfluency; and (2) to obtain evidence for the moment of error detection in speech monitoring. A few previous studies, which will be reviewed in detail in Chapter 4, have provided evidence that gestures may be sensitive to speech disfluencies (Christenfeld, Schachter & Bilous, 1991; De Ruiter, 1998; Kita, 1993; Mayberry & Jaques, 2000; McNeill, 1992; Ragsdale & Silvia, 1982). However, the scope of such studies is limited in that either the number of observations was low or the temporal assessment of gestural responses to speech disfluencies was not very fine-grained. Thus, a more systematic investigation into the gestural response to disfluency is needed. To this end, we will use three measures in order to systematically explore whether the gestures produced during disfluent speech differ from those produced during fluent speech. These measures are: frequency of gesture suspension, timing of gesture suspension in relation to speech suspension, and location of the gesture suspension within the gestural movement.

If it can be shown that gesture is sensitive to speech disfluency, the second question whether gesture can provide evidence for the process of speech interruption can be addressed. To this end, we will investigate whether gestural movements can be used to constrain the time interval containing error detection.

The results of the corpus study and their interpretation hinge crucially on the question of what triggers gesture suspension and what triggers speech suspension. Any

observed asynchrony between speech and gesture suspension, however, might be due to modality-specific suspension latencies rather than to factors related to self-monitoring in language production. Therefore, an experiment was conducted that assesses the suspension latencies of gesture and speech in sustained discourse, and is reported in Chapter 4.

The final chapter, Chapter 5, provides a summary and discussion of the findings.

2. Self-monitoring theories and language production

Chapter 2

2.1 Introduction

Theories of self-monitoring and self-repair in language production have been developed to account for the cognitive processes underlying error detection, speech interruption, and repair. These models are all based on evidence suggesting that the speech production process proceeds through different stages, and they describe how the monitor must interface with these stages. Although different assumptions are made about how monitoring processes are integrated into the underlying production architecture, they all share one critical assumption: namely that speakers can monitor their internal speech (Levelt, 1989, p. 9) prior to articulation.

This assumption has been supported by findings from several studies (e.g., Dell & Repka, 1992). One line of research has shown that erroneous words can be suspended before they can be heard and recognized. Blackmer & Mitton (1991) and Levelt (1983) found instances in which erroneous words were suspended 100-150 ms after onset of articulation (e.g., after the first phoneme or syllable). Considering that word recognition by listening to overt speech is estimated to be possible about 200 ms after word onset (Marslen-Wilson & Tyler, 1981) and that the interruption of speech takes some 150-200 ms (Logan, 1982; Logan & Cowan, 1984), it is unlikely that detection and interruption of the erroneous word can be due to listening to one's own speech (Blackmer & Mitton, 1991; Hartsuiker & Kolk, 2001; Levelt, 1983).

In another line of research, speakers have been shown to be able to monitor their internal speech and to detect and correct errors even when they cannot hear their own

overt speech. In several studies using a white noise paradigm, which diminishes auditory feedback, speakers have been found to be able to monitor their internal speech production and to detect errors (Lackner & Tuller, 1979; Postma & Kolk, 1992; Postma & Noordanus, 1996).

In yet another line of research, Motley, Camden, and Baars (1982) provide evidence for prearticulatory monitoring by measuring psychophysiological correlates of emotional arousal, the so-called *galvanic skin response*. Speakers had to read aloud two-word phrases, some of which were designed to lead to taboo words when the initial phonemes were exchanged (*tool kits* resulting in *cool tits*). Other word pairs would result in neutral expressions (*darn bore* resulting in *barn door*). Speakers were less likely to produce sound exchanges that would result in taboo words than exchanges that would result in neutral word pairs. As an index of the actual prearticulatory production and editing of the taboo word sequence, the speakers' galvanic skin response was measured. The galvanic skin response was larger in the taboo word condition than in the neutral word condition, suggesting that speakers had actually encoded the taboo word sequence. The larger galvanic skin response in this condition indicates that speakers had become aware of the taboo word sequence via prearticulatory monitoring and were able to edit it out prior to articulation.

2.1.1 Cognitive theories of self-monitoring

Although there is general agreement that speakers monitor internal speech production prior to articulation, there are different views on how this is achieved. In the past several decades, three basic kinds of self-monitoring models have been distinguished in the literature: production-based monitoring, node structure theory, and perceptual loop theory (for a detailed review, see Postma, 2000). These will be laid out in the following section.

2.1.1.1 Production-based monitoring

Production-based monitoring theories assume a single monitor (De Smedt & Kempen, 1987; Van Wijk & Kempen, 1987) or multiple monitors (Laver, 1980) as part of the

production system. Furthermore, it is assumed that intermediate and end products of the different stages of speech production are monitored. Laver (1980) distinguishes four speech production stages: ideation, abstract linguistic programming, abstract motor programming, and conversion of motor program to neuromuscular commands. Monitors are operative at the stages of linguistic programming and motor programming. As an error is detected, the repair is carried out immediately at the respective level before processing can proceed further. Blackmer & Mitton (1991) call this type of monitor a hold-up monitor, since further processing is blocked until the output is corrected. In case an error is not detected by one of the prearticulatory monitors, postarticulatory monitoring operates on the basis of the sensory information in order to detect errors in overt speech.

A limitation of Laver's theory is that it can only account for cases in which an erroneous word is suspended at the earliest 330 ms post onset. This 330 ms interval consists of the time needed to detect the error (detection latency) plus the time needed to interrupt speech (suspension latency). Laver estimates that postarticulatory monitoring is able to detect an error 180 ms after the onset of its articulation (Laver, 1980, p. 303). At the minimum, an additional 150 ms (Hartsuiker & Kolk, 2001; Logan & Cowan, 1984) for interrupting speech have to be added to this postarticulatory detection latency of 180 ms, which amounts to 330 ms. Thus, Laver's (1980) postarticulatory monitor cannot account for very short latencies of 100-150 ms observed for the suspension of an erroneous word under articulation by Blackmer and Mitton (1991) and Levelt (1983).

In cases of very short latencies, error detection must have taken place via prearticulatory monitoring. However, in prearticulatory monitoring, production is blocked at the stage at which an error was detected, so that articulation of the erroneous word would not start in the first place. That is, prearticulatory monitoring cannot account for early suspensions of erroneous words under articulation.

With regard to the question when speakers interrupt speech, Laver suggests that upon post-utterance error detection the speaker assesses the "degree of degradation the communication is likely to suffer as a result of the registered error" (Laver, 1980, p. 303). On the basis of this assessment it is decided whether speech is interrupted and a

correction will be performed, or whether speech is continued and the slip will pass without any correction. If the decision is made to correct, the "articulation-halting mechanism" (Laver, 1980, p. 303) sets in by blocking further transformation and execution of neuromuscular commands of articulation. Thus, the impact of the error on the communication in a given conversation drives the decision whether articulation is halted or continued with the articulation of the erroneous expression. Laver does not further specify the relevant interactional parameters on which such a decision is based.

2.1.1.2 Node structure theory

A connectionist account of monitoring is given by MacKay's node structure theory (MacKay, 1982; 1987; 1990; 1992a; 1992b). In this theory, production and comprehension share a single, layered network of hierarchically organized nodes (e.g., propositional nodes, lexical nodes, syllable nodes, phonological compound nodes, phonological nodes), which are activated and selected through spreading activation. As a node spreads its activation top-down to subordinate nodes, activation spreads back bottom-up to the superordinate node. In order to avoid repeated reactivation via bottom-up priming, superordinate nodes become self-inhibited after activation and commitment. Bottom-up priming, which is transmitted from a subordinate node to its respective superordinate node, serves as the mechanism of error detection. For example, when a word such as *dog* is erroneously selected for *cat*, the activation of the *dog* node spreads back to the concept node for *dog,* which was not part of the original plan. That node accumulates activation in contrast to the node of the originally intended concept *cat*, which underwent self-inhibition after selection. The prolonged activation of the erroneous concept node automatically draws attention and causes error detection. In this way, errors are detected prior to articulation, almost as soon as they occur. If an error is not detected before articulation onset, it can be detected via sensory analysis nodes, which perform the auditory input processing.

Considering the question when speech is interrupted, MacKay does not make very specific suggestions. Nevertheless, according to MacKay, error detection does not necessarily lead to error correction: "Because *listeners* often fail to detect errors (e.g., Marslen-Wilson & Tyler, 1980), *speakers* can adopt a fairly liberal criterion in deciding

whether or not to correct their own errors" (MacKay, 1987, p. 169, citation and italics in the original). This notion entails a strategic component in deciding whether or not to correct upon error detection and thus whether or not to interrupt speech.

The suggested error-detection mechanism implies that every error should be detected almost immediately, since bottom-up priming is automatic and immediate (Levelt, 1989, p. 477). MacKay (1990) suggests that the distance between the level of nodes producing the error and the level of nodes at which the error is detected determines the speed and the likelihood of detection. For example (from MacKay, 1990), a phonological error like *crawl srace* (for *crawl space*) is detected at the same level at which it occurs because the initial consonant group *sr* does not exist in English and thus no node in the network is committed to this combination. This is different for a phonological error like *cool tarts* instead of *tool carts* in the sentence *they were moving tool carts down the assembly line* (example from MacKay, 1990). Nodes exist for the units of the erroneous words; namely, for consonant clusters, syllables, words, and noun phrases. Priming percolates backwards through the levels of consonant cluster nodes, syllable nodes, word nodes, and noun phrase nodes. Only as bottom-up activation reaches the propositional level will the error be detected, since no node exists for *cool tarts* in the context of *moving down the assembly line*. Thus, speed and likelihood of error detection depend on how far up in the hierarchy the error can be detected. The further up in the hierarchy error detection is possible, the less likely an error is to be detected and the longer it will take to detect it. Hence, an error can be detected as quickly as relay times between levels allow. The precise temporal characteristics of activation and relay times are not specified.

Both the production-based monitoring and the node structure theory lack detail with regard to the characteristics of the time course of error detection, interruption and repair. Concerning the question when speakers interrupt, Laver (1980) and MacKay (1987) suggest that speakers base their decision whether or not to interrupt on interactional considerations. The types of considerations that may lead to the interruption of speech are not specified in detail. Temporal details as well as strategic accounts of disfluent speech are, however, provided by Levelt's perceptual loop theory (Levelt, 1983; 1989), which will be introduced below.

2.1.1.3 Perceptual loop theory

The perceptual loop theory is a theory about self-monitoring and error detection based on Levelt's (1989) speech production model. In order to provide a comprehensive account of the perceptual loop theory I will first lay out Levelt's speech production model in some detail.

2.1.1.3.1 Levelt's speech production model

Following Levelt (1989), the process of speech production proceeds through stages of conceptual preparation, formulation, and articulatory encoding. Accordingly, the speech production model is divided into three distinct processing components: the conceptualizer, the formulator, and the articulator (see Figure 2.1). In the conceptualizer the speaker generates a preverbal message, which represents the idea that the speaker intends to communicate in propositional format. The preverbal message is the input for the formulator, where it is translated into linguistic structure through the stages of grammatical and phonological encoding. During grammatical encoding, as chunks of the preverbal message enter the formulator, the corresponding so-called lemmas are then retrieved from the mental lexicon.

Lemmas contain semantic and syntactic information of the respective lexical items. Upon lemma selection syntactic building procedures are activated, which generate the surface structure of the respective phrase. Interim results of the grammatical encoding process can be stored in a syntactic buffer. In the second stage of the formulation process, phonological encoding takes place. Phonological encoding proceeds through three levels: morphological/metrical spellout, segmental spellout and phonetic spellout. At each level frames with slots are generated, which are then filled with the respective elements. Based on the diacritical features of the selected lemma (e.g., number, tense) the morphological and metrical composition of a word is made available during morphological/metrical spellout. At the following level of segmental spell-out frames for the respective syllables are created into which the segments are assembled. At the phonetic spell-out level the phonetic plan is translated into articulatory motor execution resulting in overt speech.

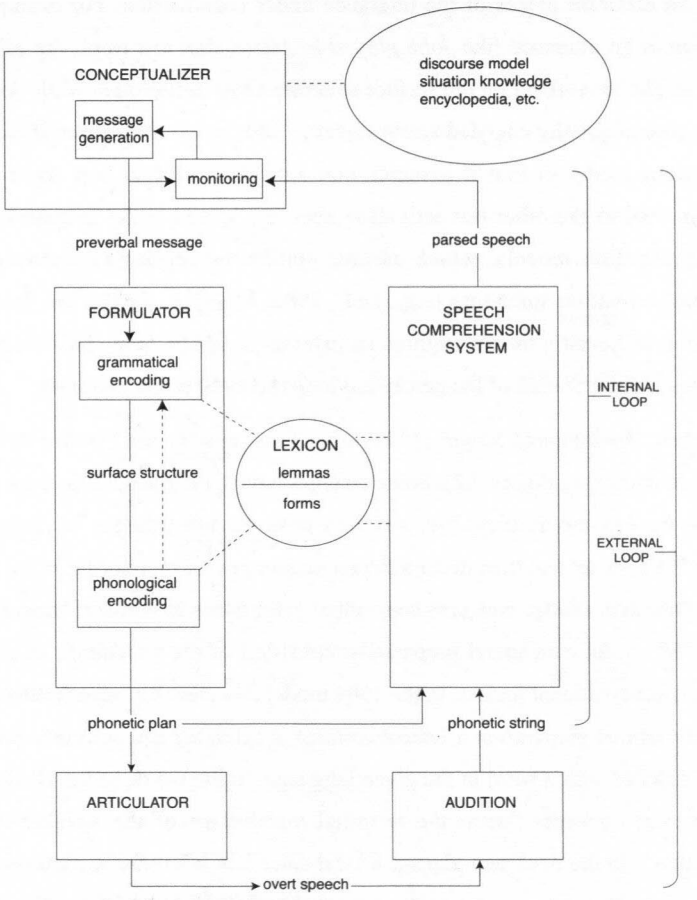

Figure 2.1. Levelt's model (1989) of speech production and self-monitoring.

The phonetic plan can be temporarily stored in the articulatory buffer. This buffer compensates for asynchronies between the encoding processes in the formulator and in the articulator. Finally, on the basis of the phonetic plan, the articulator executes the articulatory motor programs resulting in overt speech.

In Levelt's model speech production proceeds incrementally (Kempen & Hoenkamp, 1987) in that all components of the speech production system work in

parallel on different pieces of the utterance under construction. For example, in the generation of an utterance like *John played in Amsterdam last week*, the grammatical encoder might be working on the surface structure of *in Amsterdam*, while *John played* is being phonologically encoded (from Levelt, 1989, p. 25). Moreover, the model is a feed-forward model in that it assumes that activation spreads top down from one encoding level to the other but activation does not spread back. In contrast to other speech production models, which assume similar processing components but bi-directional activation spreading (e.g., Dell, 1986; MacKay, 1992a) the feed forward mechanism of Levelt's model requires an external feedback loop. This feedback loop constitutes a major feature of the perceptual loop theory of self-monitoring.

Levelt, Roelofs, and Meyer (1999) amend and modify the Levelt 1989 model in multiple respects (see Figure 2.2). Some changes have consequences for how the model accounts for self-monitoring. We will first describe the changes in comparison to Levelt's 1989 model and then discuss the consequences for monitoring in the following section. One such change concerns conceptual preparation in terms of lexical concepts. In the 1989 model conceptual preparation consisted of the generation of a preverbal message in propositional format. In the 1999 model this view has been further specified: during conceptual preparation a lexical concept is activated and selected. Each lexical concept matches with a word in the given language. In the words of Levelt et al. (1999, p. 8), lexical concepts "form the terminal vocabulary of the speaker's message construction." In the next step, during lexical selection, a lemma is retrieved from the mental lexicon that corresponds to the activated lexical concept. Due to the distinction between lexical concepts and lemmas, the role of lemmas has changed, too. While in the 1989 model the lemma specified the semantic and the syntactic properties, the lemma now only specifies the syntactic properties of the to-be-encoded word. The two versions of the theory also differ with respect to the specification of the relations between the comprehension system and the production system. In Levelt et al. (1999), the lemma stratum and the conceptual stratum are shared between the production and the comprehension system. Activation flow between lemma level and conceptual level is bi-directional, while in the 1989 model activation flow from conceptual to lemma level was feed forward only.

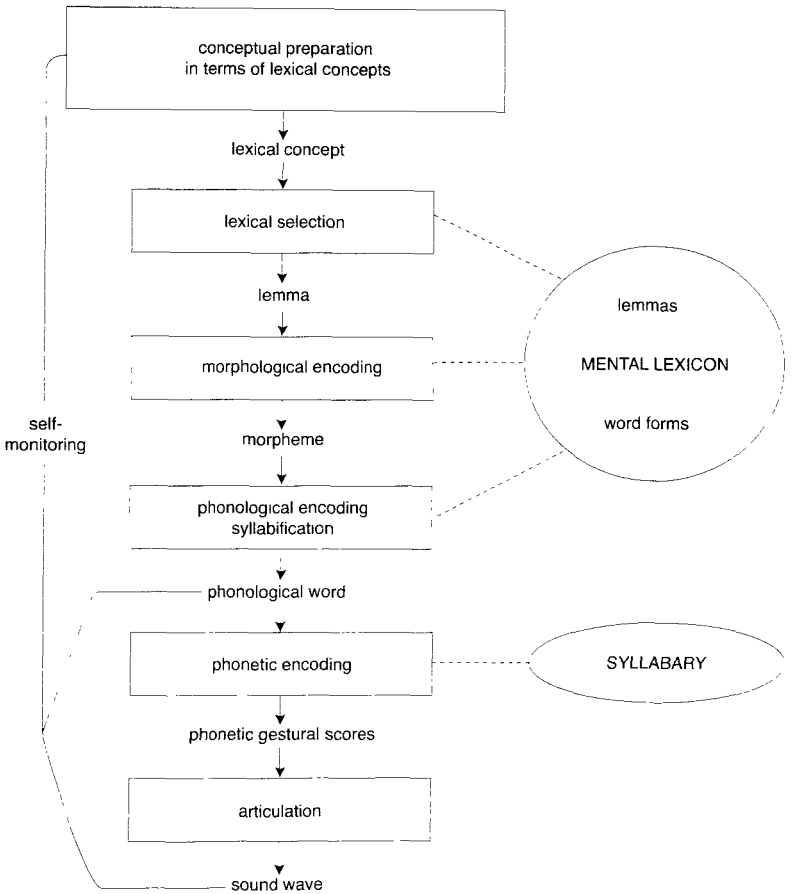

Figure 2.2. Levelt, Roelofs, and Meyer's (1999) model of speech production and self-monitoring.

After a lemma is selected and the syntactic properties have become available, the generation of the respective syntactic frame can proceed. Thereafter, morpho-phonological encoding begins. First the codes for the morphemes of the to-be-encoded word are accessed. Then the metrical and segmental properties are "spelled out" (Levelt et al., 1999, p. 5). A metrical template is generated into which the respective segments are inserted in a syllabified fashion. Syllabification is a context-dependent process that

proceeds online and incrementally. First the initial segments of the to-be-encoded word are combined into the first syllable. Then the next segments for the second syllable are filled in. As the syllables of a *phonological word* become available incrementally, phonetic encoding begins. Depending on the phonological syllables the respective articulatory gestural scores are retrieved from the mental syllabary, which constitutes a repository of the oral gestural scores of the most frequent syllables in a given language. Finally the oral gestural scores are executed by the articulatory system resulting in overt speech.

2.1.1.3.2 Self-monitoring and error detection

The account of self-monitoring and error detection in Levelt's (1989) speech production model is the perceptual loop theory. In the perceptual loop theory, Levelt (1983, 1989) assumes that prearticulatory as well as postarticulatory monitoring proceed through the same perceptual system that is used to monitor the speech of others; namely, the language comprehension system. Monitoring is a centrally governed process located in the conceptualizer. At this stage, speakers can monitor the preverbal message for its appropriateness (Levelt, 1989). The monitor has, furthermore, access to the output of the formulator and the articulator and compares the intended message with the actually encoded message. The output of the formulator and the articulator is fed into the speech comprehension system via two monitoring loops (see Figure 2.1 and 2.2). The phonetic plan, the output of the formulator, is accessed via the internal loop. Thus, speakers are able to monitor their inner speech, which in turn enables them to detect and correct errors before they are articulated. The output of the articulator, overt speech, is monitored via the external (auditory) loop. Here monitoring proceeds by listening to self-produced overt speech.

How is it possible for Levelt's proposed architecture to detect and intercept an error before it is articulated? Levelt (1989) offers the following account. As soon as the phonetic plan becomes available, monitoring starts via the inner loop. At the same time, the articulator starts preparing the motor execution of the phonetic plan. Levelt (1989) estimates the duration of this articulatory encoding—the time between reception of the phonetic plan and the start of articulation—to be 200-250 ms (based on Klapp, Anderson, & Berrian, 1973 and Klapp & Erwin, 1976). This means that in order to

prevent the articulation of an erroneous word, the inner loop monitoring has up to 250 ms to parse the phonetic plan, detect the error, and stop articulation. Levelt (1989) estimates that the internal parsing of the phonetic plan requires 150 ms (based on a 200 ms estimate for word recognition in overt speech by Marslen-Wilson & Tyler, 1981 and Tyler & Marslen-Wilson, 1986). Levelt subtracts 50 ms, because inner speech requires no auditory analysis. Thus, the earliest that error detection can take place is 150 ms after the phonetic plan has become available. This leaves up to 100 ms to interrupt speaking[2] before the erroneous word is articulated. According to Levelt (1989, p. 473), this is enough time under optimal attentional circumstances to interrupt articulation to avoid uttering the erroneous expression.

These estimates hold for cases of fast running speech where the phonetic plan is not stored in the articulatory buffer (Levelt, 1989). Given that the interval between the availability of the phonetic plan and the motor execution is longer when the phonetic plan is buffered, the speaker has more opportunity to detect and repair an error prearticulatorily (Blackmer & Mitton, 1991; Postma & Kolk, 1993; Oomen & Postma, 2001). The buffered material allows a certain look-ahead range for monitoring, and gives more time for inner loop monitoring, error detection, and repair before an erroneous element becomes ready for articulation. The more material is buffered, the more time the monitor has to detect an error. Because monitoring and articulatory encoding of the phonetic plan proceed in parallel, very early suspensions of erroneous words after, for example, the first phoneme can be accounted for by the perceptual loop theory.

The perceptual loop theory as described above was amended by Levelt et al. (1999) based on findings by Wheeldon and Levelt (1995). In the earlier version of the theory (Levelt, 1983; 1989) it is assumed that it is the output of the formulator, the phonetic plan, that is monitored. Wheeldon and Levelt (1995), however, provided evidence that the monitoring system accesses the phonological word, which is the

[2] Note that in Levelt 1983, 200 ms are assumed for interrupting articulation. The estimates of this latency vary from study to study between 100-200 ms (see Levelt, 1983, 1989; Hartsuiker & Kolk 2001).

output of phonological encoding (see Figure 2.2 above). Given that phonological encoding precedes phonetic encoding, the interval between onset of monitoring and motor execution is longer. Thus, more time is available for prearticulatory error detection and repair.

According to the 1989 version of the model, monitoring can start as soon as the phonetic plan becomes available, which is estimated by Levelt (1989) to be 200 ms prior to onset of articulation. A slightly revised estimate is given by Indefrey and Levelt (2004) for the picture naming task. They estimate phonetic encoding to begin about 145 ms before onset of articulation at about 600 ms. Based on these numbers, it can be estimated that monitoring of the phonetic plan could start 400-455 ms (600-200 ms or 145 ms) post picture onset.

In Levelt et al. (1999), the assembly of the phonological word begins approximately 330 ms post picture onset. Syllabification has an estimated duration of 25 ms per segment, which means that monitoring of the first syllable can begin 355 ms post picture onset. Hence, in Levelt et al. (1999), monitoring can start 45-100 ms earlier than in the Levelt (1989) version of the perceptual loop theory. Moreover, since the phonological representation is fed into the speech comprehension system, decoding of the phonetic representation is unnecessary; hence parsing should take less time than the estimated 150 ms (Levelt 1989) for word recognition based on the phonetic plan. Thus, in a model in which it is assumed that the internal loop accesses the phonological representation, the monitoring loop can be faster in prearticulatory monitoring, error detection, and correction.

If an error is not detected and corrected before it is articulated, it becomes overt. Accordingly, Levelt (1983; 1989) distinguishes between overt and covert repairs. The process underlying overt repairs is that the troublesome item is detected so late that it is either under articulation or already articulated, and the suspension of speech happens during or after articulation of the erroneous word. From the disfluent utterance one can mostly infer *that* something went wrong and *what* went wrong. Consider the following example from Levelt (1989, p. 479, italics in the original). In (1) the speaker, who was about to produce *yellow*, interrupted mid-word and corrected after the filled pause *er* by resuming with *orange*.

(1) we can go straight on to the *ye-*, er orange

By contrast, in covert repairs (see also Postma & Kolk, 1993) an error is detected and edited out prearticulatorily, such that in speech only a hesitation remains in the form of a filled or silent pause or a repetition of a lexical item. In these cases, no morpheme is changed, added or deleted in the resumed delivery, as the following Dutch examples illustrate (from Levelt, 1983, p. 55).

(2) *En aan de rechterkant een oranje stip, oranje stip*
'And at the right side an orange dot, orange dot'

(3) *Dan rechtsaf, uh grijs*
'then right, uh grey'

In example (2) the original utterance is suspended at *stip* ('dot') and the delivery is resumed with the repetition of *oranje stip* ('orange dot'). In the Dutch example (3), speech is suspended after *rechtsaf* ('right') and is resumed after the filled pause *uh*. Here the speaker resumes his utterance with *grijs* ('grey'), which is a correct description of the picture to be described (see p. 5 for a task description in the Levelt study).

Although the term covert repair suggests that error detection and correction have taken place covertly, the utterance itself does not contain an indication for an actual repair. It is possible that the speaker was dealing with fluency problems due to a delay in lexical retrieval and bridged the prolonged retrieval time with a filled pause. The reason for the disfluency might not be speech production itself, but might be rooted in the interactional situation. The speaker might have paused since the interlocutor did not pay attention or did not give feedback. Repetitions could be a result of a restart mechanism located at the articulator, which sets in when input from higher production levels is lacking (Blackmer & Mitton, 1991). Hence, there are multiple reasons for a speech suspension in a covert repair. Most importantly, such speech suspensions are not necessarily caused by error detection. This is different for overt repairs, in which in the resumption of some element or even the entire utterance is altered with respect to the original utterance. It can be assumed that the speaker has monitored speech, detected an error and decided to repair.

In sum, theories of self-monitoring and repair differ with respect to the stages of speech production that are monitored and the mechanisms of monitoring. The most explicit theory is the perceptual loop theory. We will therefore adopt the architecture suggested in this theory as a background to test and discuss the two hypotheses providing different accounts of the exact mechanism underlying self-interruption, the MIR hypothesis and the DIP hypothesis (see p. 5). In the next chapter we present a quantitative analysis of a corpus of speech disfluencies conducted to obtain evidence for or against the two hypotheses.

3. Speech suspension: error detection or repair readiness?

Chapter 3

3.1 Introduction

This chapter addresses the question: Do speakers initiate speech interruption upon error detection or upon repair readiness? The study aims to shed further light on the process of speech interruption and self-repair by investigating the relationships between the ways in which speakers suspend their speech, the amount of time needed to process the repair, and the complexity of the repair process. This will be done with respect to the two competing hypotheses described in the General Introduction that give differing accounts of the process of error detection, speech interruption, and repair. The Main-Interruption-Rule hypothesis (MIR hypothesis) claims that as soon as an error is detected, speech interruption is initiated. In contrast, according to the Delayed-Interruption-For-Planning hypothesis (DIP hypothesis), speakers do not initiate interruption immediately upon error detection but delay interruption until they have the repair ready. In the following we will review in detail the hypotheses and the evidence supporting either.

3.1.1 The Main-Interruption-Rule hypothesis

The notion of an immediate interruption upon the detection of an error goes back to Nooteboom. Nooteboom (1980) investigated whether speakers halt speech immediately upon error detection or whether they go on speaking to complete the linguistic unit under articulation. He analyzed repairs in the Meringer Corpus (Meringer, 1908) and found that speakers completed the linguistic unit under articulation (the erroneous word) in 90% of lexical errors (N = 163) and in about 70% of phonological errors (N = 252).

Furthermore, speakers suspended their speech predominantly after the first word boundary of the erroneous item suggesting that they did not go on speaking after error detection more than up to the next word boundary. Nooteboom concluded that the point of suspension is determined by the moment of error detection and by a preference to complete words under articulation. As mentioned above, Levelt (1983) argues for a general rule of immediate interruption termed the Main Interruption Rule (MIR) in an influential study of self-monitoring and self-repairs in Dutch (see p. 5). The MIR states: "Stop the flow of speech immediately upon detecting the occasion of repair" (Levelt, 1983, p. 56). Stopping the flow of speech upon detecting the occasion of repair means that as soon as an error is detected, all processing in the subcomponents of the speech production system is simultaneously interrupted (Levelt, 1983, p. 56). Levelt (1983) assumes a constant latency of about 200 ms for the interruption. Thereafter, the replanning of the repair proper starts.

Levelt (1983) sought evidence for the MIR hypothesis by investigating one consequence of the assumption of immediate interruption, namely that speech suspensions should occur in the vicinity of the reparandum, the to-be-corrected word, but not much later. Levelt found evidence for this assumption in the observation that for overt repairs (N = 689), in which an element in the resumption is altered in comparison to the original utterance, speakers tend to suspend speech more often within or right after the reparandum than within or right after words following the reparandum (67% vs. 33%).

Another consequence of the MIR hypothesis is that linguistic boundaries should not be respected in speech interruption: speakers should disrupt their speech at any given point within a constituent, word, syllable, or phoneme and not delay the interruption in order to complete a linguistic unit. Hence, under the MIR hypothesis, the occurrence of linguistic boundaries at speech suspensions points should not be above chance level. Levelt (1983) computed the proportion of constituent boundaries at suspension points for repairs in which speech was suspended one or more words after the reparandum (excluding within-word suspensions), assuming that in these repairs the speaker may have delayed interruption to complete the constituent. In the next step, the words from the beginning of the disfluent utterance were counted up to the position of

the suspension (e.g., after the 4th word). To provide a chance baseline, he then determined how often this position corresponded to a constituent boundary in a fluent utterance. The results revealed that in repairs the proportion of constituent boundaries at suspension points was higher than the proportion of constituent boundaries at corresponding locations in fluent baseline utterances. This suggests that speakers do respect linguistic boundaries by delaying interruption in order to complete the current constituent.

However, Levelt proposed an alternative explanation for the result, namely that attention in monitoring is heightened towards the end of constituents. Thus, error detection is more likely towards constituent ends, which results in a higher number of suspensions at constituent boundaries. Levelt (1983) assumed that if speakers actually respect linguistic boundaries, the proportion of suspensions at constituent boundaries should be higher for repairs in which interruption was delayed for constituent completion than for repairs in which the speaker interrupted immediately upon error detection. In contrast, if attention is heightened towards constituent ends, the proportions should not differ. For the analysis Levelt determined the proportion of suspensions at constituent boundaries for the two types of interruption (delayed and immediate). He assumed that in repairs with suspensions right after the reparandum the speaker interrupted immediately upon error detection, while in repairs with suspensions at words following the reparandum, the speaker may have delayed interruption in order to complete the linguistic unit. The results revealed that the proportions did not differ for possibly delayed interruptions and immediate interruptions, supporting the idea that during monitoring attention is enhanced towards constituent ends, hence suspensions occur more often at constituent-final po itions. This interpretation holds under the assumption that suspensions right after the reparandum are indeed due to immediate interruption upon error detection. The data, however, do not rule out that interruption was delayed in these cases until a constituent boundary was reached.

In order to provide independent evidence for the claim of heightened attention towards the end of constituents, Levelt (1983) furthermore computed the error detection probability for color naming errors at different syllable positions within the constituent. This was calculated as the proportion of erroneous color terms that the speaker

corrected at each position. As a measure of error detection this assumes that uncorrected color term errors had not been detected. It was found that the proportion of corrected color name errors was highest at constituent final position, suggesting that error detection was most likely at this position. Levelt concludes that the tendency to suspend at constituent boundaries, is not due to delayed interruption of speech in order to complete the linguistic unit, but rather to heightened attention in monitoring towards the end of a constituent. Since errors are more often detected towards the end of a constituent, speech suspensions occur more often at constituent final positions.

For word boundaries, Levelt's (1983, p. 60) data showed that the majority of the speech suspensions happened after words (74%), not within words (26%) for repairs with suspensions within or right after the reparandum (N = 542). In order to account for this tendency for word completion, Levelt (1983) computed proportions of within-word suspensions of reparanda and of words following reparanda. The results showed that the later the suspension of speech with respect to the reparandum, the lower the number of within-word suspensions: within-word suspensions occurred in 26% of the reparanda as compared to only 13% of words following the reparandum. Levelt concludes that there is a preference to complete suspension words that follow the trouble word.

A further analysis of the types of repairs revealed that more erroneous words than inappropriate words[3] were suspended within-word. While 23% of the error repairs (N = 399) showed within-word suspensions of the erroneous reparandum, only 7% of the appropriateness repairs (N = 290) showed within-word suspensions of the inappropriate reparandum. The findings that inappropriate words and neutral words following an erroneous reparandum are seldom suspended within-word and that three times as many erroneous words than inappropriate reparanda were suspended within-word, led to the following qualification of the MIR: "only erroneous words may be interrupted upon detection of the occasion for repair" (Levelt, 1983, p. 64). The results also led to the pragmatic motivation of the MIR, namely, that within-word suspensions

[3] Inappropriate expressions are not wrong but ambiguous, vague or lack "coherence with previously used terms or expressions" (Levelt, 1983, p.52).

signal the erroneousness of the suspension word. Levelt states: "So, the more general rule seems to be that correct words should not be interrupted, and this holds equally well for correct trouble words (i.e., in appropriateness-repairs) as for neutral words (i.e., in delayed interruptions). Interrupting a word signals that that word is wrong." (Levelt, 1983, p. 64). This in turn means that by completing the word, speakers signal that that word itself is correct. Levelt states: "By interrupting a word, a speaker signals to the addressee that that word is an error. If a word is completed, the speaker intends the listener to interpret it as correctly delivered." (Levelt, 1989, p. 481). Since inappropriate words are not erroneous, speakers do not suspend them within-word. When an erroneous word is detected so late that the suspension would result in a within-word suspension of a following neutral word, the interruption is delayed so that the articulation of the neutral word is completed before suspension. When erroneous words are completed, it is because the error was detected too late to avoid its utterance and to suspend the erroneous expression within-word.

Brédart (1991) sought evidence for the MIR in a corpus of French self-repairs. He found that within-word suspensions were less frequent for words following the reparandum than for suspensions of the reparandum itself, replicating Levelt's (1983) findings. Brédart (1991) furthermore tested the prediction of the MIR that the longer the erroneous word, the more likely the speaker is to detect the error and suspend within-word. In contrast, Brédart predicted that word length should not have an influence on the suspension of inappropriate words, since these words are completed for pragmatic reasons. He found the predicted positive relationship between word-length and the frequency of within-word suspensions for erroneous words. Long erroneous words tended to be suspended within-word, while short erroneous words tended to be completed. However, in contrast to the second prediction this positive relationship also held for neutral words following erroneous words and inappropriate reparanda, although the relationship was weaker: long words were only slightly more likely to be interrupted within-word than short words.

Evidence against the MIR hypothesis comes in two basic forms. On the one hand, studies have examined various aspects of disfluencies like acoustic features of the suspension word or the distribution of fillers like 'uh' or 'um' and the following silence

until resumption. These studies provide evidence that speakers are more flexible in planning when and how to interrupt than suggested by the MIR hypothesis. Furthermore, studies examining the time course of speech suspension and resumption have yielded data for which the mechanism of speech interruption and repair processing suggested by the MIR hypothesis cannot account. By and large, these types of evidence are consistent with the main assumption of the Delayed-Interruption-For-Planning hypothesis, namely that interruption does not take place immediately upon error detection. In the following section, we will first lay out the basic assumption concerning the mechanisms of error detection and speech interruption of the DIP hypothesis. Thereafter the different types of evidence will be reviewed.

3.1.2 The Delayed-Interruption-For-Planning hypothesis

As mentioned in the General Introduction, the DIP hypothesis is based on a suggestion by Blackmer and Mitton (1991). They put forth the idea that repair readiness could be the crucial factor triggering speech interruption based on findings on the temporal characteristics of disfluencies, which will be discussed in detail below. The DIP hypothesis states that interruption of speech is not initiated immediately upon error detection but delayed for repair processing. The basic idea is that speakers simultaneously strive to maintain fluency by continuing to speak and thereby process the repair as covertly, namely while they are still speaking. The underlying cognitive mechanism proposed is as follows. First, the speaker does not initiate speech interruption immediately upon detection of trouble, and repair planning can start prior to the initiation of interruption. Second, when the repair has been processed up to a point at which it is accessible to the monitoring process, the speaker makes a decision whether to interrupt speech or not. Finally, speech is inevitably suspended when the speaker runs out of prepared material in the formulator and the articulatory buffer.

The delay in interruption is possible because of incremental speech production and the temporary storage of encoded material in the articulatory buffer (Levelt, 1989; see also Chapter 2). Buffering allows processing asynchronies between the formulator and the articulator to be adjusted and enables fluent speech. At the moment of error detection the formulator and the articulator are working in parallel and incrementally on different chunks of an utterance. Chunks of the phonetic plan are stored in the

articulatory buffer. Levelt (1989, p. 473) estimates that up to a few phonological phrases can be stored in the buffer. Thus, while speakers start processing the repair upon error detection, they can go on speaking until the formulator and the articulatory buffer run out of prepared material. If the repair processing is fast enough to reach final stages of encoding before the speaker runs out of prepared material, speech can be interrupted and the repair can be articulated immediately or very soon after suspension. When the repair processing is not fast enough to be completed or to reach final stages of production before the speaker runs out of prepared material, the speaker is compelled to finalize replanning after suspension. Thus, the effectiveness of delaying speech interruption for repair planning depends on the amount of material already being processed in the formulator and the articulator before error detection.

The mechanism suggested by the DIP hypothesis enables replanning to proceed while speech interruption is delayed. We will now review the available evidence for the basic assumption of the DIP hypothesis that interruption does not take place immediately upon error detection but is delayed.

3.1.3 Evidence for different types of planning prior to interruption

A number of studies have provided evidence for different types of planning that seem to require relatively much time prior to the interruption of speech and thus a delay of speech interruption. Speakers seem to have more options for suspension than those assumed by the MIR hypothesis namely, completing the word under articulation or not. Such different types of planning prior to the initiation of interruption indicate that errors or problems in formulation have been detected and that speakers have an estimate of the complexity of the repair process. It is assumed that speakers decide upon error detection not only if speech should be interrupted but also where and how to suspend speech.

One strategy speakers seem to apply is to lengthen the unit under articulation. Thereby the speaker continues to speak and buys time to resolve underlying speech production problems before they become overt. Bell, Fosler-Lussier, Girand, Gregory, and Gildea (2003) and Bell, Daniel, Fosler-Lussier, Girand, and Gildea (1999) analyzed the influence of planning problems as manifested in three types of disfluencies (silent

pauses, filled pauses, repetitions) on the pronunciation of preceding function words. Function words were roughly twice as long when preceding a disfluency as when preceding a fluently articulated word. Function words before filled pauses ('uh' and 'um') were lengthened most. This result indicates that trouble detection happened early and that speakers did not interrupt upon detection but lengthened the unit under articulation. However, in cases of prolongation it is not clear whether the speaker experienced plan completion problems, like a temporary delay in word form retrieval, or whether an error was detected that was edited out covertly. Nevertheless, the data support the idea that error detection may not necessarily lead to an immediate initiation of speech interruption but that speakers may prolong the unit under articulation to buy time to repair covertly.

Berg (1986) provided evidence that not only the moment of speech suspension is planned but also the features of the suspension itself. In a corpus of naturally occurring German repairs Berg found instances of within-word suspensions resulting in word fragments, which were pronounced according to the devoicing rules for word-final position in German. This phonological accommodation must have been planned in advance, which in turn suggests that interruption was not initiated immediately upon error detection. The point of the actual suspension of speech within-word and also the way the fragment was pronounced were planned.

Findings of two studies by Fox Tree and Clark (Clark & Fox Tree, 2002; Fox Tree & Clark, 1997) point in the same direction. Fox Tree and Clark (1997) provided evidence that speakers signal the initiation of a speech suspension by modifying the phonetic realization of the suspension word. In this study, Fox Tree & Clark (1997) investigated the realization of the two variants of the English article *the*, 'thuh' ([ðʌ]) with a reduced vowel and 'thiy' ([ðɪ:]) with a non-reduced vowel (see also Jefferson, 1974). The authors found that 'thiy' was followed by a suspension of speech in 81% of the instances, while only in 7% of the instances was speech suspended after 'thuh'. Fox Tree and Clark (1997) argue that speakers interactionally signal incipient trouble and the upcoming suspension of speech by realizing *the* as 'thiy' instead of 'thuh'. The results seem to indicate that speech suspension is delayed until after 'thiy', and that the realization of the suspension word as 'thiy' is planned.

Note that, as in the study of Bell et al. (1999, 2003), one open question is whether speakers in the analyzed instances were facing a temporary failure of word retrieval or whether an error was actually detected. Another open question is whether inappropriateness or an error was detected, since the MIR hypothesis applies only for the latter case. Because Fox Tree and Clark use a classification scheme for repairs that does not take erroneousness and inappropriateness into account, the actual distribution of error and appropriateness repairs is not provided. Fox Tree and Clark (1997) report that 'thiy' was followed by repairs like replacements of words and fresh starts (abandoning of the original utterance and fresh start of a new and different utterance), in which the cause for the repair can be an inappropriate expression as well as an erroneous expression.

In a second study Clark and Fox Tree (2002) reported that speakers clitizised 'uh' and 'um' to the last word of the original utterance, suggesting that speech is not suspended upon error detection, but that the suspension itself is planned together with the respective filler. Moreover, upon error detection speakers seemed to estimate the time needed for resolving the problem and signaled the expected delay with 'uh' or 'um' accordingly. Clark and Fox Tree, (2002) examined two corpora (the London-Lund corpus and the corpus of English conversations (Svartvik & Quirk, 1980)) with respect to the occurrence of 'uh' and 'um' (see also Smith & Clark, 1993). They found that 'uh' was followed by minor delays, while 'um' was followed by major delays. The authors suggest that speakers detect an error and estimate how long it will take to resolve the problem. Depending on the expected amount of time needed for problem solving speakers chose and realized 'uh' or 'um' to signal the duration of the following delay in the progress of speech. Interestingly, Fox Tree and Clark (2002) found that 'uh' and 'um' were often clitizised onto the last word of the original utterance and syllabified, leading to the creation of a single prosodic word. For instance, 'but uh' was syllabified to 'bu.tuh'. According to the authors, the syllabification indicates that upon error detection the speaker planned to signal the delay with 'uh' or 'um' and to suspend speech thereafter. Thus, the suspension word, for example, consisting of 'but', and 'uh', was selected and phonologically encoded as a single prosodic unit.

There is also evidence that the initiation of interruption is postponed while features of the repair are planned. According to the MIR hypothesis, the initiation of interruption should not be affected by the characteristics of the following repair, since interruption is determined by the moment of error detection and the erroneousness of the trouble word. Catchpole, Hartsuiker, and Pickering (2003) conducted an experiment in which they elicited lemma substitution errors by means of a picture naming task developed by Van Wijk and Kempen (1987). In a small number of trials the first picture changed after 300 ms into another picture while participants were still naming the first picture. Subjects were instructed to self-interrupt as quickly as possible and to name the second picture. The second picture varied in terms of degradedness, which prolonged naming latencies (Meyer, Sleiderdink, & Levelt, 1998). Catchpole et al. (2003) found that the suspension latency increased when the picture was degraded. Their results indicate that speakers did not interrupt immediately upon error detection (the moment the picture changed) but that the moment of initiation of interruption was influenced by the planning of the to-be-named item on the degraded picture.

The studies reviewed above have provided evidence for the necessity of planning prior to speech interruption requiring a delay of speech interruption. Another set of studies has provided timing evidence speaking to the differential predictions of the MIR hypothesis and the DIP hypothesis. We will review these studies in detail in the following.

3.1.4 Evidence from timing studies

Studies on the characteristics of the time course of error detection, speech suspension, and resumption provide a major source of evidence regarding the cognitive mechanism underlying self-monitoring. These studies have focused on the temporal intervals from onset of the reparandum (trouble morpheme or phoneme) to speech suspension (referred to as *error-to-cut-off interval*) and the so-called *cut-off-to-repair interval* (see Figure 3.1 below), which represents the duration from suspension to resumption (Blackmer & Mitton, 1991; Kormos, 2000; Oomen, 2001; Van Hest, 1996).

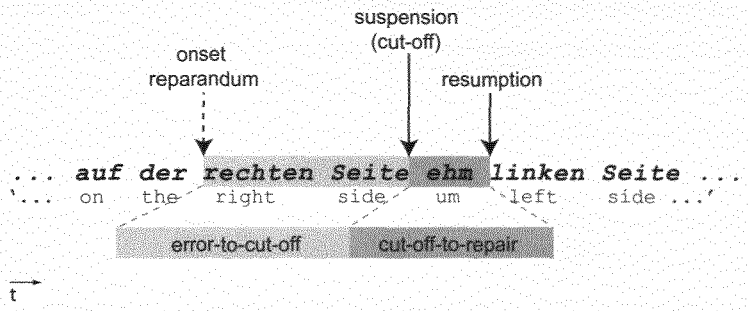

Figure 3.1. Schematic representation of error-to-cut-off and cut-off-to-repair interval. The error-to-cut-off interval represents the time from onset of the reparandum to speech suspension and the cut-off-to-repair interval the time from speech suspension to onset of the resumption.

A main reason why much attention has been devoted to timing issues and especially to the cut-off-to-repair interval is that the MIR hypothesis makes explicit predictions about timing that can be tested. Under the MIR hypothesis, upon error detection all subcomponents of the speech production system are interrupted with an assumed latency of 200 ms and thereafter replanning follows (Levelt, 1983, p. 56). This assumption implies that the error-to-cut-off interval reflects detection latency plus interruption latency. It also entails that there must be a lag of some length between speech suspension and resumption because some amount of time is needed to plan and process the repair (see Figure 3.2 below). This amount of time needed for replanning should be reflected in the cut-off-to-repair interval. Because no replanning occurs before the actual speech suspension, replanning is entirely *overt* under this assumption.

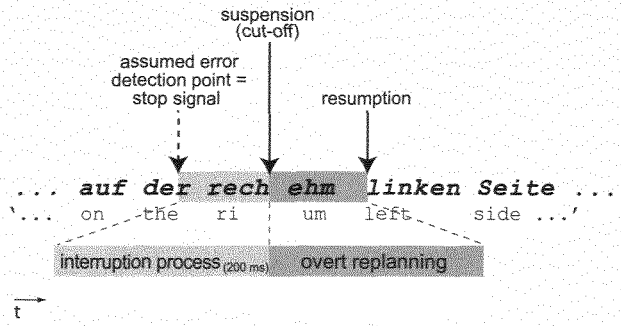

Figure 3.2. Schematic representation of the temporal sequence of error detection, interruption and replanning following Levelt (1983). In this example the assumed error detection point is 200 ms before speech suspension, the time needed to complete the interruption process (Levelt, 1983). After suspension, the overt replanning follows during the cut-off-to-repair interval.

However, timing studies have found that resumptions can follow speech suspensions immediately without any lag, which suggests that the planning of the repair must have been completed at the moment of speech suspension (Blackmer & Mitton, 1991; Van Hest, 1996; Oomen & Postma, 2001). In order to distinguish replanning taking place before speech suspension from replanning taking place after speech suspension we will call the former *covert replanning* and the latter *overt replanning*.

Blackmer and Mitton (1991) investigated the timing characteristics observed in a corpus of speech disfluencies from a Canadian radio call-in show. Critically, they observed disfluencies with no time lag between the point of suspension and the point of resumption, which they refer to as zero ms cut-off-to-repair intervals (Blackmer & Mitton, 1991, p. 189). Zero ms cut-off-to-repair intervals have also been observed by Van Hest (1996), who reported 50 instances in her corpus of L1 and L2 disfluencies. In the Blacker and Mitton (1991) study such zero ms cut-off-to-repair intervals were encountered for 12.4% of all (covert and overt) repairs (N = 1525) and for 19.2% of the overt repairs (N = 339). The authors found zero ms cut-off-to-repair intervals not only for appropriateness repairs, but also for error repairs. Note that, according to the MIR hypothesis, the initiation of interruption in appropriateness repairs is delayed so that the suspension occurs when the word under articulation is completed. In cases of

appropriateness repairs zero ms cut-off-to-repair intervals can be accounted for if it is assumed that replanning starts upon error detection while interruption is delayed for word completion. Thus, if replanning proceeds fast enough, the repair can be ready for articulation at the moment of speech suspension. However, zero ms cut-off-to-repair intervals should not exist for error correction, because the MIR hypothesis requires an interval after speech suspension during which the replanning of the repair takes place. In contrast, the DIP hypothesis is subject to no such restriction, because any kind of replanning can start upon error detection and can even come to completion before speech is suspended. Thus, the repair can be articulated immediately after speech is suspended.

Oomen and Postma (2001) also observed zero ms cut-off-to-repair intervals when investigating effects of increased speech rate on self-monitoring. The authors tested a hypothesis derived from Levelt's perceptual loop theory, namely that the effectiveness of prearticulatory monitoring and repair depends on the amount of buffered material in the articulatory buffer. Oomen and Postma (2001) assumed that with increased speech rate, buffering time in the articulatory buffer diminishes. Therefore, the time available for the system to monitor via the inner loop decreases. According to Oomen and Postma this predicts a smaller number of zero ms cut-off-to-repair intervals at a fast speech rate than at normal speech rate, because zero ms cut-off-to-repair intervals can only occur when buffering is present. Participants were instructed to describe a network of pictures. The order in which the pictures had to be named was determined by a dot that was moving through the network. Speech rate was manipulated by varying the speed with which the dot moved through the network, resulting in a speech rate of 3.6 syllables and 4.5 syllables per second respectively. The authors measured the cut-off-to-repair interval for 84 error and 75 appropriateness repairs in the normal condition, and for 99 error repairs and 64 appropriateness repairs in the fast condition. Zero ms cut-off-to-repair intervals were observed for appropriateness repairs as well as for error repairs in both speed conditions at similar relative frequencies: normal speech rate: 27% for appropriateness repairs, 7% for error repairs; fast speech rate: 27% for appropriateness repairs, 10% for error repairs. The results do not support the idea of shorter monitoring time due to diminished buffering at increased speech rate.

Currently speakers were able to detect errors and re-plan faster when speaking fast.

However, the results do suggest that error detection and replanning can take place before interruption is initiated, even at an increased speech rate.

However, the phenomenon of zero ms cut-off-to-repair intervals may not be so problematic for a modified version of the MIR hypothesis proposed by Hartsuiker and Kolk (2001). In the original version of the MIR hypothesis, the entire speech production system is halted upon error detection with a constant latency of 200 ms (Levelt, 1983) between initiation and completion of the interruption process. Thus, replanning can only begin after the interruption process has completed, which entails that speech suspension must be followed by a lag during which the replanning takes place. In contrast, Hartsuiker and Kolk (2001) proposed that interruption and replanning are initiated simultaneously and run in parallel. Zero ms cut-off-to-repair intervals can occur if the time needed for replanning is less than or equal to the time needed to complete the interruption process. In other words, the replanning is accomplished during the time needed for completing the interruption process and the articulation of the repair can start immediately after speech is suspended.

Hartsuiker and Kolk (2001) implemented this modified version of Levelt's perceptual loop theory as a computational model. They proposed that *not* all components of the speech production system are halted upon error detection, only the articulator. Thus, because the formulator is not suspended it can start working on the repair immediately after error detection. At the same time that the stop signal to the articulator is sent off, repair processing starts (see Figure 3.3 below). The time intervals they assume in their simulations for interrupting speech and for replanning allow for the completion of replanning during the latency needed for completion of the interruption process of speech. It is an important question whether the assumed time intervals are sufficiently long to account for repairs of varying complexity. We will discuss this question by taking into account the details of Hartsuiker and Kolk's (2001) simulations in section 3.1.5 (p. 45).

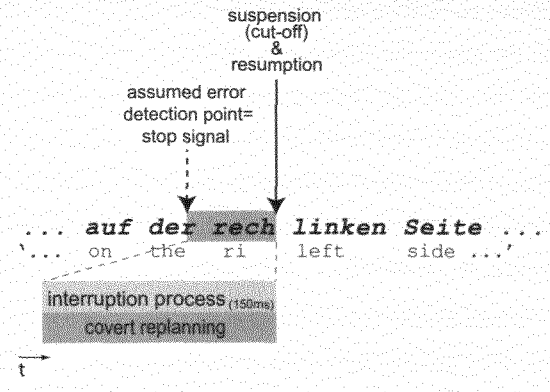

Figure 3.3. Schematic representation of the mechanism and the temporal requirements for zero ms cut-off-to-repair intervals following Hartsuiker and Kolk (2001). Covert replanning and the interruption process start at the same time and are completed at the same time.

Hartsuiker and Kolk (2001) further specified Levelt's model by assuming two stages of articulation: a selection stage for selection and activation of motor programs, and a command stage controlling execution. Using this model, Hartsuiker and Kolk (2001) were able to successfully simulate many findings, including the effects of speech rate on the distribution of cut-off-to-repair-intervals and the zero ms cut-off-to-repair intervals obtained by Oomen and Postma (2001). Because of the advantage of the modified MIR hypothesis over the original MIR hypothesis in explaining zero ms cut-off-to-repair intervals we will henceforth take the modified MIR hypothesis as the main competitor for the DIP hypothesis.

While it is possible for the modified MIR hypothesis to account for zero ms cut-off-to-repair intervals, this hypothesis seems less able to account for other findings from timing studies. The modified MIR hypothesis suggests that the length of the error-to-cut-off interval should be independent of the length of the cut-off-to-repair interval. This is because the duration of the error-to-cut-off interval purely reflects the latency of error detection plus the interruption latency. These durations are independent of the repair processing. Since interruption latency is constant, covert replanning time is constant in error repairs (150-200 ms). In appropriateness repairs the covert replanning time can be no longer than the word under articulation. In contrast, the DIP hypothesis

assumes a negative correlation between the error-to-cut-off interval and the cut-off-to-repair interval: the longer the error-to-cut-off interval, the shorter the cut-off-to-repair interval should be. This is because replanning starts upon error detection and interruption is delayed as long as the speaker can go on speaking and as long as repair processing takes. Hence, covert replanning time is not constant but depends on the amount by which interruption is delayed. Because the time spent before suspension is used for replanning, less time will be needed during the cut-off-to-repair interval itself.

Blackmer and Mitton (1991) found such a negative correlation, supporting the DIP hypothesis. They grouped the repairs in their corpus into fast and slow repairs (cut-off-to-repair interval < 250 ms vs. > 250 ms) based on Goldman-Eisler (1968) and Brotherton (1979), who used a value of 200-250 ms as a threshold to differentiate between short silences during articulation and longer hesitation pauses used for planning. 68% of the overt repairs (N = 339) had cut-off-to-repair intervals of 250 ms and shorter. For this data set of fast repairs Blackmer and Mitton (1991) found a significant negative correlation between error-to-cut-off times and cut-off-to-repair times. The longer the error-to-cut-off interval was, the shorter was the cut-off-to-repair interval. This suggests that the longer interruption is delayed (as reflected in the error-to-cut-off-interval), the further covert replanning can proceed. Hence, the cut-off-repair interval is short. However, in the 32% slower repairs, in which the cut-off-to-repair interval was equal to or longer than 250 ms, no such correlation was found. Blackmer and Mitton (1991) suggest several explanations for the lack of a correlation in the subset of slow repairs. One suggestion is that at the moment of error detection there was no buffered material that would have allowed the speaker to go on speaking, resulting in a long cut-off-to-repair interval. Another explanation they offer for slow repairs is that the suspension occurred soon after detection because the speaker preferred to interrupt immediately instead of going on speaking, for example to prevent saying something socially inappropriate.

However, a later study by Van Hest (1996) did not replicate the finding of a negative correlation between error-to-cut-off times and cut-off-to-repair times in fast repairs, although several features of her study make it incommensurable with that of Blackmer and Mitton (1991), including the use of a different operationalization of

cut-off-to-repair interval and the inclusion of errors produced in speakers' first and second language.

Finally, a finding that might be considered as problematic for the MIR hypothesis is that silent cut-off-to-repair intervals following within-word suspensions tend to be shorter than those following after-word suspensions (Nakatani & Hirschberg, 1994). According to the MIR hypothesis, the length of the cut-off-to-repair interval reflects the entire replanning time, because replanning only starts after speech suspension. The finding that this interval is shorter for within-word suspensions could be seen as problematic for the MIR hypothesis, but only under the assumption that replanning after within-word suspensions (cases of error repairs) takes the same amount of time as replanning following after-word suspensions (cases of error and appropriateness repairs). If this assumption holds, some replanning must have taken place before speech suspension in cases of within-word suspensions.

Taken together, there is evidence suggesting a relationship between the covert and the overt replanning time, such that the lengthening of the former seems to lead to a shortening of the latter. The modified MIR hypothesis has difficulties to account for this type of relationship, because it assumes some stable amount of covert replanning starting with the initiation of interruption and before speech suspension. We will now discuss under which conditions the amount of available replanning time is sufficient to explain zero ms cut-off-to-repair intervals.

3.1.5 *Limitations of the modified Main-Interruption-Rule hypothesis*

The most problematic finding for the MIR hypothesis—zero ms cut-off-to-repair intervals— can in principle be accounted for under the modified version of the MIR hypothesis proposed by Hartsuiker and Kolk (2001). As discussed above, to account for these intervals, Hartsuiker and Kolk (2001) proposed that interruption and repair processing start simultaneously upon error detection and are processed in parallel. Their simulations demonstrated that under this assumption, repair processing can come to completion in less than or in the same amount of time that is required to interrupt the articulator, such that articulation of the repair proper follows speech suspension immediately (see Figure 3.3 above). The simulations depend critically on the temporal

values assigned to the processes of interrupting and repair processing (see Table 3.1 below for an overview of all intervals).

Hartsuiker and Kolk (2001) assume 150 ms to be necessary for interrupting the articulator (*Interrupting*). At the same time as the interruption process starts, the repair processing begins. The critical latency between error detection and onset of repair articulation comprises the intervals of *Restart planning* (50 ms), *Phonological encoding* (110 ms) and articulatory encoding, which in turn consists of *Selection* (100 ms) and *Command* (100 ms). The restart planning interval represents the "duration of repeated execution of selection processes [i.e., conceptual and grammatical encoding] before phonological encoding minus the time benefit obtained from priming the to-be-selected units" (Hartsuiker & Kolk, 2001, p. 128). Since the authors limit their model to simple error repairs and do not distinguish different types of errors (syntactic, lexical, phonological), the assumed interval for restart planning (50 ms) can be very short. Hartsuiker and Kolk (2001) suggest that the conceptualization and grammatical encoding (i.e., restart planning) for such repairs will only take a very short amount of time because the correct representations are still available. They assume that, due to the preceding attempt to encode the same units, a facilitatory priming effect speeds up the production of the repair in the processes both preceding and following phonological encoding (Hartsuiker, personal communication).

Table 3.1. Temporal parameter set in the Hartsuiker and Kolk simulations (Table from Hartsuiker & Kolk 2001, p. 128).

Basic Duration of Each Time Interval in the Model			
Stage	Symbol	Duration (ms)	Per unit*
Phonological encoding	T_{phon}	110	σ
Selection	T_{sel}	100	ω
Command	T_{com}	100	σ
Audition	T_{aud}	50	ω
Parsing	T_{parse}	100	ω
Comparing	T_{comp}	50	ω
Interrupting	T_{int}	150	ω
Restart planning	$T_{restart}$	50	ω

*ω = word, σ = syllable

However, for more complex repairs, in which new conceptual and syntactic representations are generated, the assumed 50 ms for restart planning seem to be a very short duration. Indefrey and Levelt (2004) assume a considerably longer time window (250 ms) for the pre-phonological encoding stages of conceptual preparation and lemma selection in picture naming (see Table 3.2 below for an overview). If the repair involves the generation of new material, the production at the conceptual stage and the lemma retrieval stage cannot be sped up by a facilitatory priming effect for repair processing.

A similar problem arises for the later processing stages. Hartsuiker and Kolk (2001) assume a total interval of 310 ms consisting of the stages of phonological encoding and articulatory encoding (110 ms *Phonological encoding* + 100 ms *Selection* + 100 ms *Command*). Since already 50 ms of the 150 ms interruption latency are taken up by restart planning, this leaves only 100 ms for the interval of phonological encoding up to speech onset. Thus, for the repair processing to be completed at the moment of suspension, the 310 ms needed for phonological encoding, selection, and command would have to be sped up by a facilitatory priming effect of about 210 ms. However, if a more complex repair has to be processed and new material is generated, the speech production system cannot take advantage of a facilitatory priming effect. In such cases, the completion of phonological and articulatory encoding should require the full 310 ms. Thus, the same problem arises for phonological and articulatory encoding has for conceptual and grammatical encoding: namely the assumed time intervals do not appear to be sufficient for more complex repairs.

Table 3.2. Estimated time windows for successive operations in single spoken word encoding (Table from Indefrey & Levelt 2004, p. 108).

Estimated time windows for successive operations in spoken word encoding	
Operation	Duration in ms
Conceptual preparation (from picture onset to selecting target concept)	175
Lemma retrieval	75
Form encoding	
Phonological code retrieval	80
Syllabification	125
Phonetic encoding (till initiation of articulation)	145
Total	600

In sum, detailed temporal considerations show that the possibility of zero ms cut-off-to-repair intervals should depend on the complexity of repair processing; that is, on the extent to which new material will be incorporated in the repair. Specifically, the modified MIR hypothesis assumes that repairs following zero ms cut-off-to-repair intervals should consist primarily of repeated material. The more complex the repair, the less likely it is, that the repair processing will fit into the 150 ms required for interruption. Hence, any amount of repair processing that cannot be completed within the covert replanning time window must necessarily be completed during overt replanning time, the cut-off-to-repair interval.

Zero ms cut-of-to-repair intervals can be seen as a specific case of a more general principle: the less covert replanning time is available, the more overt replanning time—as reflected in the cut-off-to-repair interval—should be necessary. Thus, the complexity of the replanning could provide a test bed for the hypotheses. This is because the hypotheses differ with respect to the amount of replanning that can take place before speech suspension (covert replanning). While for the MIR hypothesis covert replanning time is constant, it is variable following the DIP hypothesis. The more complex the replanning, the more time should be required which should be reflected differently in the cut-off-to-repair intervals depending on either hypothesis. Based on these considerations we conducted the quantitative corpus analysis presented in the following section. We will first discuss in greater detail the rationale of the study and introduce the predictions of the two hypotheses. We will then turn to the empirical data.

3.2 Corpus Study 1: Suspension type, cut-off-to-repair interval, and repair complexity

Both the DIP hypothesis and the MIR hypothesis as modified by Hartsuiker and Kolk (2001) assume that replanning begins upon detection of an error and proceeds in parallel with ongoing speaking up to the moment of speech suspension. Processing of the repair can take place covertly, while the speaker continues speaking (covert replanning), or overtly, during the cut-off-to-repair interval (overt replanning). Thus, the covert replanning phase comprises the period from the detection of the error to the suspension of speech and the overt replanning phase comprises the subsequent period from the suspension of speech to onset of the resumption (see Figure 3.4 below).

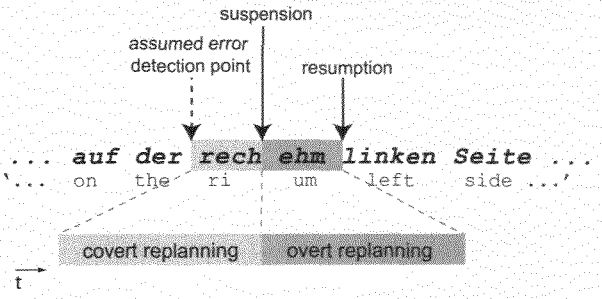

Figure 3.4. Schematic representation of covert replanning phase and the overt replanning phase.

A critical difference between the DIP hypothesis and MIR hypothesis regards to what extent the processing of the repair happens covertly versus overtly. The MIR hypothesis assumes that the covert replanning phase will be no longer than 150 ms necessary for completion of the interruption process in case of an error repair or no longer than the time needed to complete a word under articulation in case of an appropriateness repair. Further replanning necessary to complete the repair processing will be overt, during the cut-off-to-repair interval (see Figure 3.5, upper panel).

In contrast, the covert replanning phase of the DIP hypothesis is subject to no such constraint (see Figure 3.5, lower panel). The DIP hypothesis assumes that speakers

wish to minimize delays in speaking, and thus will strive to continue speaking while trying to covertly complete as much of the replanning as possible before they interrupt. In other words, speakers strive to minimize the overt and maximize the covert replanning time. Therefore, one way of testing these hypotheses is to investigate how much of repair planning is conducted covertly versus overtly.

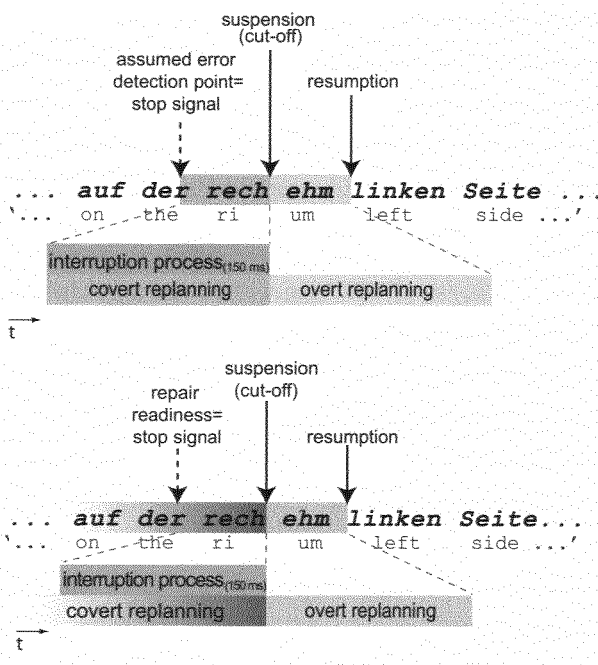

Figure 3.5. Schematic representation of covert replanning phase in relation to interruption latency according to the modified MIR hypothesis (upper panel) and the DIP hypothesis (lower panel).

To this end, it is necessary to obtain measures of the covert and overt replanning phases of the total replanning time. The overt phase can be observed because it is reflected in the cut-off-to-repair interval. However, it is not possible to directly measure the amount of covert replanning that has taken place, since the moment of error detection cannot be observed. Nonetheless, it is possible to draw inferences about covert processing time based on the way in which speech is suspended; specifically, whether speech suspension occurs within- or after-word.

For the MIR hypothesis, the suspension type is determined by the erroneousness of the word under articulation at the moment of detection. According to the MIR hypothesis' strategic signaling account of within-word suspension, suspending the word under articulation before it has been completed signals that this word was wrong. Therefore, the type of suspension can be used to infer the type of underlying interruption process, namely whether it was immediate or delayed. Within-word suspensions should be the result of immediate interruption upon detection. This means that at the moment of suspension, replanning has proceeded covertly for the 150 ms that are required to complete the interruption process (Hartsuiker & Kolk 2001). The remaining amount of time required to complete replanning must therefore proceed overtly, during the cut-off-to-repair interval.

Under the MIR hypothesis, after-word suspensions present a less straightforward case because they could be the result of immediate or delayed interruption. Interruption may be delayed in order to complete the word currently under articulation, for example, because the reparandum was inappropriate, not erroneous. Compared to within-word suspensions where the amount of covert replanning time was identical to the interruption latency, in after-word suspensions an additional amount of replanning may have taken place during the completion of the word. However, after-word suspensions can also be the result of immediate interruption. This would be possible if an error happens to be detected 150 ms before the end of the word under articulation, given that it takes the same amount of time for the interruption process to complete. The predictions of the MIR hypothesis for after-word suspensions depend upon the proportion of cases that fall into either category. For instance, if all after-word suspensions were the result of immediate interruption, the MIR hypothesis would predict that the cut-off-to-repair interval should be equal for within-word and after-word suspensions. If at least some after-word suspensions fall into the category of delayed interruption, the MIR hypothesis would predict that less of the replanning time on average should be overt in the after-word than in the within-word case. Thus, the average cut-off-to-repair interval should be longer or equal in the within-word case compared to the after-word case.

Under the DIP hypothesis, the suspension type depends on factors that are altogether different from the erroneousness of the word under articulation. The interruption process can be characterized as a race between the repair planning processes and the emptying of prepared material from the formulator and articulatory buffer. Within-word suspensions reflect cases in which the planning processes finish before the formulator and the articulatory buffer have been emptied. In other words, for the DIP hypothesis a within-word suspension indicates that the replanning process was completed covertly. Because of this covert processing, little or no overt processing during the cut-off-to-repair interval is required. In contrast, after-word suspensions reflect cases in which the buffer is emptied before the replanning process can finish. Since the formulator and the articulatory buffer run out of prepared material, speech is compelled to cease at the end of the last buffered word, and the replanning process will enter an overt phase. Thus, the DIP hypothesis makes the opposite prediction from the MIR hypothesis; namely, it predicts that the cut-off-to-repair interval should be longer when preceded by an after-word suspension than by a within-word suspension.

It is possible to make further predictions that differentiate the hypotheses based on the complexity of the processing that must be undertaken in order to produce the repair. Some repairs are more complex than others, and therefore will require more time to process. Consider cases of major repairs like fresh starts, in which the original utterance is abandoned and a new syntactic construction is generated, like in *wenn man links in ehm vorm Haus war eine Garage* ('when one left into um in front of the house was a garage'). In this case, a new preverbal message has to be generated, which then undergoes the complete encoding process from lemma retrieval and syntactic frame generation to morpho-phonological encoding, syllabification and phonetic encoding. Since everything has to be generated anew, no advantage can be taken of residual priming. These replanning processes should take more time than a minor repair, such as a phoneme repair in which only one element has to be exchanged and advantage can be taken of residual priming due to the previous production of most of the segments.

Both hypotheses predict an influence of repair complexity on cut-off-to-repair times since major repairs will require more processing overall, which would result in a corresponding increase in the total repair time. The two hypotheses differ, however,

with respect to whether or not this additional repair processing can partially or totally occur during covert replanning time. According to the MIR hypothesis, covert replanning cannot exceed the time needed to complete the interruption process (150 ms) or the time it takes to complete a word under articulation. Because neither time depends on repair complexity, the covert repair time stays constant, while the overt repair time increases with repair complexity (see Figure 3.6 and 3.7 below).

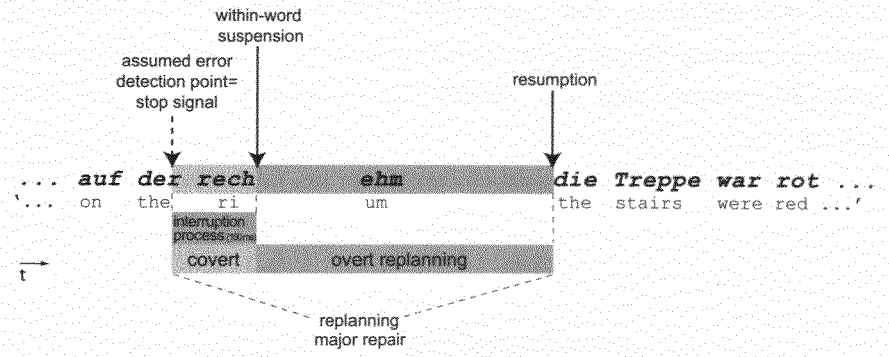

Figure 3.6. Covert and overt replanning time for a major repair following the MIR hypothesis.

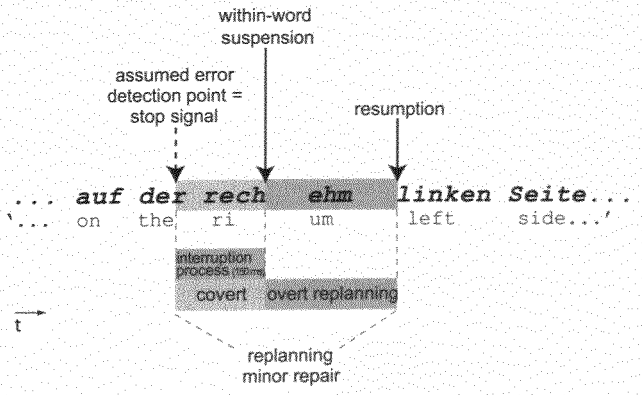

Figure 3.7. Covert and overt replanning time for a minor repair following the MIR hypothesis.

According to the DIP hypothesis, in contrast, covert replanning time may increase with more complex repairs. Note, however, that the effect of repair complexity should be conditional upon the type of suspension. For within-word suspensions, in which most replanning has been completed when speech interruption is initiated, the cut-off-repair interval should be the same for major versus minor repairs because no or only minimal overt replanning would be necessary (see Figure 3.8 and Figure 3.9 below). In contrast, in after-word suspensions, where repair planning has exceeded the amount of buffered material, replanning is partially overt. Thus, the length of overt planning should depend upon repair complexity, with longer cut-off-to-repair intervals for major than for minor repairs.

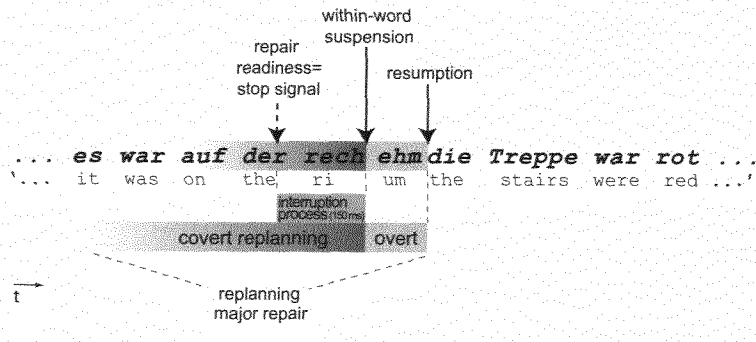

Figure 3.8. Covert and overt replanning time for a major repair following the DIP hypothesis.

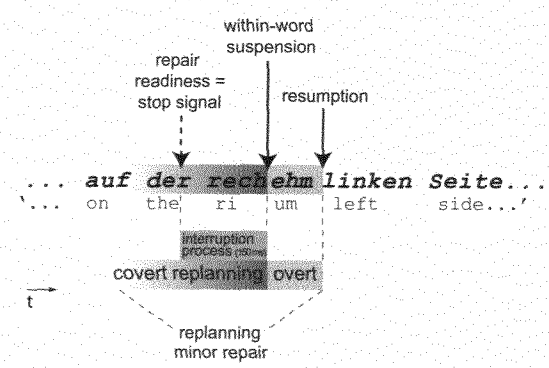

Figure 3.9. Covert and overt replanning time for a minor repair following the DIP hypothesis.

Finally, the hypotheses can be further tested by examining the distribution of major and minor repairs following zero ms cut-off-to-repair intervals. Given the temporal parameter settings in the simulations of Hartsuiker and Kolk (2001), the modified version of the MIR hypothesis can account for zero ms cut-off-to-repair intervals when the replanning phase is very short. This is true for minor repairs but not for major repairs. In major repairs, a new syntactic frame has to be generated and no advantage can be taken of a facilitatory priming effect due to the previous production of the same material. When the interruption is immediate as reflected in within-word suspension, replanning can only be completed if a minor repair was processed, but not when a major repair was processed. Hence, the MIR hypothesis predicts that when preceded by a within-word suspension, zero ms cut-off-to-repair intervals can only be followed by minor repairs.

According to the DIP hypothesis, zero ms cut-off-to-repair intervals indicate in all cases that repair processing was completed independent of its complexity. The processing of major repairs as well as minor repairs can be completed if enough prepared material is available in the formulator and the articulatory buffer. Thus, the DIP hypothesis predicts that when preceded by within-word suspensions, zero ms cut-off-to-repair intervals can be followed by major as well as by minor repairs. Table 3.3 summarizes the predictions.

Table 3.3. Overview of the predictions.

Dependent Variable	**Prediction by MIR hypothesis**	**Prediction by DIP hypothesis**
Cut-off-to-repair interval	after-word ≤ within-word	after-word > within-word
Cut-off-to-repair interval	minor repair < major repair	within-word: minor repair < major repair after-word: minor repair < major repair
Repairs that can follow a within-word suspension with a zero ms cut-off-to-repair interval	minor repairs	minor repairs & major repairs

3.2.1 Method

3.2.1.1 Data

For the purpose of this study, a corpus of German speech disfluencies was compiled. The data collection took place in Berlin and Mainz, Germany. The data were collected within a semi-experimental setting using a fixed task. Participants had to describe houses and apartments to an interlocutor. The performance was audio- and video-recorded and analyzed for the occurrence of speech disfluencies. In the following sections, the data collection will be explained in detail.

3.2.1.2 Task

A main objective of the study was to obtain a data set that was, on the one hand, ecologically valid and representative of everyday speech, and on the other hand, rich in varieties of speech disfluency. To this end, the data were collected by asking participants to provide living space descriptions (Linde & Labov, 1975; Ulmer-Ehrich, 1982). In living space descriptions, speakers describe houses or apartments in which they live or have lived, including explanations of such aspects as the layout of the space, the arrangement of the furniture, the location of windows, and so on. Describing a living space to someone is an activity of everyday life and therefore a fairly natural task (Linde & Labov, 1975; Ulmer-Ehrich, 1982). The setting (i.e., describing the space to an interlocutor) situates the participants within an interactional situation At the same time the task is fairly complicated and results in a considerable amount of speech disfluencies of various kinds. The information to be expressed has to be selected and to be linearized (Levelt, 1981; 1996), a process that is especially demanding in this particular case because the speaker has to transform three-dimensional space into the linear structure of speech. In addition, a perspective has to be chosen from which the space is described (e.g., bird's eye perspective, gaze tour or walking tour perspective (Ulmer-Ehrich, 1982). Furthermore, the speaker has to choose the appropriate words and constructions in order to convey the selected and linearized spatial information in a comprehensible way.

3.2.1.3 Data collection

The corpus consists of 12 semi-natural conversations. Seven of the recordings were made in private settings, six of them with the author as the interlocutor, one with a volunteer as interlocutor. The other five recordings were made at the Freie Universität Berlin, with a volunteer as interlocutor. Each session lasted between 20 and 40 minutes. For ten of the sessions, only the first 8-9 minutes were analyzed. In the other two sessions, the interlocutor pairs switched roles after the description of the first house or apartment, with the listener becoming the second speaker. In this case, only the description of the first speaker was selected, which resulted in two segments of around 6 minutes. Altogether, 96.3 minutes were analyzed.

3.2.1.3.1 Participants

All participants (age 25-32) were native speakers of German. They were undergraduate or graduate students of the Freie Universität Berlin and of the Universität Mainz. The 12 speaker/listener pairs of interlocutors consisted of four female/female pairs, two male/male pairs, four male/female pairs, and two female/male pairs.

3.2.1.3.2 Procedure and instructions

The experimenter provided spoken instructions to the participants. Participants were told to describe houses in such a way that the interlocutor would be able to recognize the place and, for example, locate the room of the speaker. Participants were free to describe the houses/apartments in as much detail as they wished. In all cases, the places being described were unknown to the interlocutor. The interlocutors were free to ask questions in order to understand the description. When the interlocutor was a volunteer (six out of twelve sessions, see section 3.2.1.3) the experimenter left the room after giving the instructions so that the participant and interlocutor were alone.

3.2.1.3.3 Equipment

The sessions were videotaped with a digital PAL DV camera. The camera was placed on a tripod in front of the participants at a distance of about 3-4 m. The camera was equipped with an external microphone to ensure high quality audio recording. In addition, speech was recorded with a minidisk recorder. The microphone of the minidisk recorder was attached to the clothing in front of the chest.

3.2.1.3.4 Digitization and analysis tool

The video data were digitized and transformed into CINEPAC format. Transcription and coding were performed with the annotation software tool MediaTagger developed at the Max Planck Institute for Psycholinguistics (Brugman & Kita, 1995). The annotation tool enables the tagging and labeling of video segments. The time data (starting time code, and ending time code of tagged intervals) are registered automatically. MediaTagger provides a noise-free stable still-picture of each frame (at 40 ms intervals), which allows a frame-by-frame analysis of movement sequences and concurrent speech. Moreover, MediaTagger allows the user to listen to a specific stretch of movie during the playback, for which in addition a waveform display of the specific segment is provided. For each level of coding (e.g., speech, disfluency, suspension point) a tier can be assigned in which the specific segments are tagged. For the present study, multiple tiers were assigned for the speech transcription and the coding of the features of the disfluency (i.e., suspension point, cut-off-to-repair interval, and resumption point). The transcription and coding of the respective speech segments was based on a frame-by-frame analysis of the movie. Thus, the time resolution was 40 ms.

3.2.1.4 Transcription and coding of speech

3.2.1.4.1 Transcription

The speech was transcribed verbatim from the digitized video in orthography that was adapted to capture the actual speech pattern as closely as possible. For example, word sequences like *so ein* ('such a') are mostly contracted in conversational German to *son* ('sucha'). This was transcribed as it sounded in order to preserve the characteristics of conversational speech. Filled pauses (*eh*, *ehm*, *mh*), silent pauses, indications of speech suspensions like glottal stops, laryngalization, and truncated words were also transcribed.

3.2.1.4.2 Speech coding

All disfluencies were categorized as covert or overt repairs (see section 2.1.1.3.2, p. 24). Overt repairs were defined as disfluencies that contained a clear indication of a speech suspension (e.g., a glottal stop, laryngalization, filled pause, or silent pause greater than 200 ms) and a resumption in which a modification of the original delivery had taken

place. In contrast, covert repairs were defined as disfluencies that consisted only of a filled pause (eh, ehm, mh) or a repetition of a phrase or a word.

Covert repairs were excluded from the analysis because in this type of repair it is unclear what the cause of the suspension was and whether a covert editing process had taken place. These disfluencies do not contain a reparandum in the original delivery. The resumption constitutes either a well-formed continuation of the original delivery (e.g., *das Schlafzimmer war ehm auf der rechten Seite*, 'the bedroom was um on the right side') or an immediate repetition of items of the original delivery followed by a well-formed continuation (e.g., *es war auf der der linken Seite*, 'it was on the the left side'). Disfluencies in which the speaker suspended speech, but did not resume in the same turn because the interlocutor took the next turn, were also excluded from analysis.

The remaining overt repairs were coded on the basis of the transcript, the audio and the video files. The coding proceeded according to the following steps, each of which is defined and described in more detail in following sections. First, the moment of speech suspension *(suspension point)*, the duration of the *cut-off-to-repair interval*, and the onset of the speech resumption *(resumption point)* were identified. In the next step, it was coded whether the last word of the original delivery was suspended within-word or after-word (*suspension type*). In addition, the resumption was classified with respect to the modification that had taken place in the resumed delivery (*resumption type*). The resumption types were then further classified as major or minor repairs (*repair type*).

3.2.1.4.3 Temporal assessment of speech disfluencies

The timing of the suspension point, the cut-off-to-repair interval, and the resumption point was assessed on the basis of the video and audio file with the annotation tool MediaTagger. In addition to the audio file, the waveform generated by MediaTagger was used as a visual aid. Begin and end points of the respective cut-off-to-repair intervals were tagged. This was done by listening to the cut-off-to-repair interval segment. Then, the end of the suspension word and the beginning of the resumed delivery were tagged. The tagged interval was shortened video-frame by video-frame

until no element of the suspension word or the resumed delivery could be heard anymore (see Figure 3.10 below for an illustration of the tagging points).

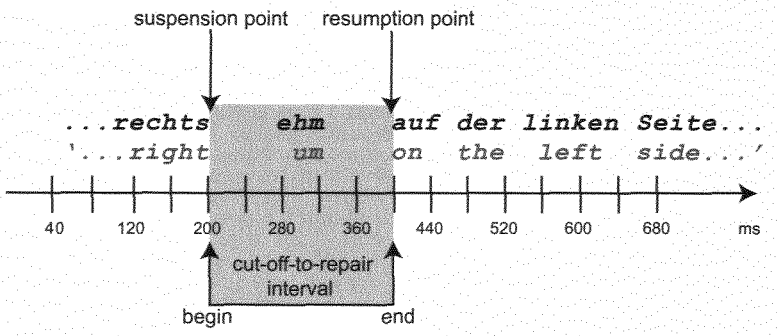

Figure 3.10. Schematic presentation of the tagging of the suspension point, the beginning and the end of the cut-off-to-repair interval and the resumption point. In the above example the cut-off-to-repair interval has a duration of 200 ms.

Note that due to the limitations of video, a 40 ms increment size was employed. In cases where there was no time interval or a time interval shorter than 40 ms between the suspension and resumption of speech, a one frame cut-off-to-repair interval (i.e., 40 ms) was assigned by tagging the estimated boundary between the last phoneme of the preceding word and the first phoneme of the following word.

3.2.1.4.4 Suspension type

Every disfluency was coded according to whether the suspension occurred within a word or whether the suspension word was completed. Following Levelt (1983), every case of premature truncation of the suspension word was coded as a within-word suspension. This included cases in which the suspension truncated the last phoneme of the suspension word. In these cases, articulation was prematurely terminated, usually by a glottal stop or laryngalization. Every suspension in which the suspension word was fully articulated was coded as an after-word suspension.

3.2.1.4.5 Resumption type

In order to determine the complexity of the repair, the resumed delivery was first classified with regard to the kind of modification that had taken place. Second, the

resumptions were classified with respect to the complexity of the replanning process and the level of the production process involved in the replanning.

For the resumption coding, Clark's classification scheme (1996) was used and further specified, since for the present study the complexity of the repair operation was the central measure. Clark's (1996) scheme classifies repairs based on the modification that has taken place in the resumption. For instance, in the resumed delivery an element can be added, deleted, or substituted. The basic procedure was to compare the original delivery to the resumed delivery and to determine which parts of the original utterance remained the same, which parts were modified, and in what fashion. The following five resumption types were distinguished:

Substitution: An item of the original utterance is substituted by an item of the same category in the resumption (e.g., a noun can only be substituted by another noun, a prepositional phrase can only be substituted by another prepositional phrase):

*auf der **rechten** em **linken** Seite*
'on the right um left side'

Addition: An element is added in the resumption as compared to the original delivery. A defining feature of additions is that there has to be back tracing (i.e., repetition of elements of the original utterance) with an item added compared to the original utterance:

ging nochmal son langer Flur son ganz schmaler langer Flur
'went again sucha long hallway...sucha very narrow long hallway'

Since compounding is a productive process in German, the same rule was applied to cases where a word is added to another word as in the example below:

*das Zimmer eh **Wohn**zimmer*
'the room uh living room'

Deletion: An element is deleted in the resumption compared to the original utterance. As in additions, a defining feature of deletions is that there has to be back tracing

(repetition of elements of the original utterance) with an item deleted compared to the original utterance.

auf der ganz linken auf der linken Seite
'on the very left on the left side'

Fresh start: In a fresh start the original delivery is abandoned and a completely new syntactic frame is generated (part of commit-and-repair strategies in Clark, 1996, p. 272).

wenn man links in eh vorm Haus war eine Garage
'when one left into uh in front of the house was a garage'

Additional category

In addition to Clark's (1996) categories, a fifth category was added to the resumption types, a mixed category.

Mixed: Two different types of processes take place in the resumption. These are mainly cases where an element of the original utterance is omitted in the resumption, while another element is added, like in the following example:

und hatte en grossen Kamin eh en Eckkamin
'and had a big fireplace uh a corner fireplace'

Here one element, *grossen* ('big') was deleted and one element *Eck* ('corner') was added in the resumed delivery.

3.2.1.4.6 Repair type

The resumptions were classified as major or minor repairs. The underlying rationale for this distinction was to obtain two classes of resumptions that could be distinguished with respect to the complexity of replanning and the time needed for replanning. The resumptions differed with respect to whether or not there was material of the original utterance that could be re-used in replanning. The basic assumption was the following: when previously produced linguistic units of the original utterance are taken over in the resumption, the replanning is faster than when everything has to be generated anew. For example, in phonological repairs, the conceptual and grammatical representations are

still available and do not have to be altered. The generation of the repair can be fast since the system can take advantage of residual activation of the previously processed material (Hartsuiker & Kolk, 2001). In contrast, a fresh start involves the generation of a new preverbal message, which has to undergo the complete encoding process through all encoding levels, from lexical concept activation, through lemma selection and phonological encoding, to phonetic encoding. Since new representations have to be generated in every encoding step, no advantage can be taken of residual activation. This should take considerably longer processing time than minor changes.

The various resumption types differ with regard to whether or not there is re-used material. In additions, an element is added to the original utterance but the syntactic frame of the original utterance is only altered in that one syntactic unit is added. A new word or even a new noun phrase has to be generated, which entails that a new lexical concept has to be selected and grammatically and phonologically encoded. However, the syntactic frame of the original utterance is maintained and can be re-used in replanning.

In deletions, the speaker reuses parts of the original utterance by backtracking to some prior element of the original utterance and repeating it, this time omitting some part.

In substitutions, the syntactic frame of the original utterance is maintained while one element is replaced by another element of the same class (e.g., noun by noun, adverb by adverb). In this case, the error can have originated at the conceptual level or at the lemma level (for a more detailed account see Levelt, 1989, p. 218). In the former case the wrong lexical concept was selected. The speaker might have chosen the wrong lexical concept because he originally thought, for example, that a room was located on the right of the hallway, but realized that it was actually on the left. In this case replanning involves selecting the correct lexical concept (i.e., left) and encoding it grammatically and phonologically. However, it is also possible that the error originated at the lemma level: the correct lexical concept (i.e., left) was chosen, but the lemma selection mechanism failed in selecting the target lemma among co-activated lemmas. Consequently, the wrong item (i.e., right) was encoded. In this case the replanning might be sped up because the lexical concept and the corresponding lemma (i.e., left)

might still have some residual activation. Because only the new word-form (i.e., left) remains to be encoded, this replanning process should be faster than a replanning process that involves lexical concept selection and grammatical and phonological encoding. It is mostly not possible to determine at which level the error originated and thus which of these two replanning processes has taken place.

In mixed resumptions, the syntactic frame of the original utterance is maintained while one element is omitted and one element is added or substituted. The replanning involves a combination of re-using previously encoded material, omitting previously encoded material and generating new material. Thus, some part of the replanning can be sped up, while some part of the replanning needs time to generate new material. It is not exactly clear how these two processes influence the overall replanning time.

Each resumption was classified with respect to whether or not it repeated old material, assuming that replanning is sped up when parts of the original utterance remain, while encoding new material slows replanning down (see Table 3.4 for an overview). We assumed that only fresh starts could be unambiguously classified as major repairs. The resumption types addition, deletion, substitution, and mixed were considered to be minor repairs. Although some part may be novel in these resumptions, nonetheless some part is taken up again, such that processing can be sped up.

Table 3.4. Overview of resumption type categories, example, altered item, and repair type.

Resumption type	Example	Repair type
Fresh start	*wenn man links in ehm vorm Haus war eine Garage* 'when one left into um in front of the house was a garage'	major
Addition	*ging nochmal son langer Flur son ganz schmaler langer Flur* 'went again sucha long hallway sucha entirely narrow long hallway'	minor
Deletion	*auf der ganz linken auf der linken Seite* 'on the very left on the left side'	minor
Substitution	*auf der rechten eh linken Seite* 'on the right uh left side'	minor
Mixed	*und hatte en grossen Kamin eh en Eckkamin* 'and had a big fireplace uh a corner fireplace'	minor

3.2.1.4.7 Reliability

A reliability check was performed on 15% of the coded disfluencies randomly selected from the corpus (N = 1202). A second trained rater independently transcribed and coded for suspension indication, suspension type and resumption type. The raters agreed on 89% of the suspension types. They also agreed on 74% of the resumption type coding. This percentage is comparable to the 76% agreement reported in Blackmer and Mitton (1991) and to the 73% agreement reported in Levelt (1983).

3.2.1.4.8 Statistical Analysis

A repeated measures 2 x 2 ANOVA was conducted on cut-off-to-repair intervals with suspension type (within-word, after-word) and repair complexity (major, minor) as factors. In addition, planned comparisons were carried out in order to test the differences between the cut-off-to-repair intervals for major versus minor repairs in within-word suspensions and for major versus minor repairs in after-word suspensions. This was done with t-tests with Bonferroni adjustment of the alpha level.

3.2.2 Results

3.2.2.1 Characteristics of the speech and the disfluencies in the corpus

Overall, the corpus consisted of 96.95 minutes of speech. On average, participants spoke for 8.08 minutes with a range from 5.80 to 9.04 (SD = 1.05) minutes. During the overall speaking time of 96.95 minutes, participants uttered 15,078 words (including word fragments) (M = 1256, SD = 285, range 780-1694). The mean speech rate was 156.16 words per minute, with a range of 89.85 to 188.46 (SD = 30.59). For an overview by participant see Appendix, Table 6.1.

Participants produced 1,202 disfluencies in total. On average, participants produced 100.17 disfluencies, with a range from 61 to 125 (SD = 18.44). The mean rate was one disfluency every 12.81 words, with a range of 6.96 to 17.11 (SD = 3.12). In temporal terms the mean rate was one disfluency every 4.94 seconds, with a range of 3.84 to 6.17 (SD = 0.80). For an overview by participant see Appendix, Table 6.2.

As mentioned above, a number of disfluencies were excluded from the main analysis. Six disfluencies were excluded because the interlocutor took the turn directly after speech suspension by the speaker. Disfluencies that had only a filled pause (N = 429) and disfluencies with an unaltered repetition of some item of the original utterance (N = 176) were also excluded, since it could not be determined whether and what kind of covert replanning process had taken place. Finally, 87 disfluencies were excluded because they were unclassifiable in terms of the resumption.

Data from two participants were excluded from the analysis. One participant did not provide any data points in the category of major repairs with within-word suspensions. The second participant was excluded because his means exceeded the group means by two standard deviations in three of the factor combinations (major repairs with within-word suspensions, major repairs with after-word suspensions, and minor repairs with after-word suspensions).

In the remaining set of 448 overt repairs, 31.9% (N = 143) of the suspensions were within-word suspensions, and 68.1% (N = 305) were after-word suspensions. Of the overt repairs, 44.6% (N = 200) were major repairs and 55.4% (N = 248) were minor repairs. The numbers of major and minor repairs following within-word and after-word suspensions are shown in Table 3.5.

Table 3.5. Number of major and minor repairs following within-word and after-word suspensions.

		Suspension type		
Repair type		Within-word	After-word	Total
	Minor repair	100	148	248
	Major repair	43	157	200
	Total	143	305	448

3.2.2.2 Effects of suspension type and repair type

The two hypotheses predicted opposite effects of suspension type on cut-off-to-repair intervals (see Table 3.3, p. 55). Our results showed that cut-off-to-repair intervals following after-word suspensions were on average 266 ms longer than cut-off-to-repair

intervals following within-word suspensions (430 ms vs. 164 ms). This difference was significant (main effect of suspension type, $F(1,9) = 14.86$, $MSE = 29.71$, $p < .01$). This result was predicted by the DIP hypothesis.

With respect to repair type, the MIR hypothesis predicted a main effect (longer cut-off-to-repair intervals for major repairs) and no interaction. The DIP hypothesis predicted an interaction with longer cut-off-to-repair intervals for major repairs only in the case of after-word suspensions. Cut-off-to-repair intervals preceding major repairs were on average 161 ms longer than cut-off-to-repair intervals preceding minor repairs (378 ms vs. 217 ms). This difference was significant (main effect of repair type, $F(1,9) = 8.35$, $MSE = 19.34$, $p < .05$).

However, there was also a significant interaction between suspension type and repair type, ($F(1,9) = 5.251$, $MSE = 23.35$, $p < .05$). As shown in Figure 3.11, the effect of repair type depended on the level of suspension type.

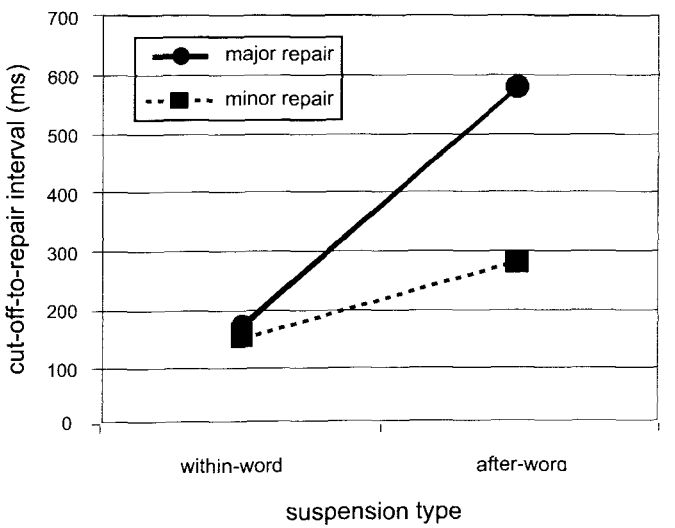

Figure 3.11. Effects of suspension type (within-word vs. after-word) and repair type (major vs. minor).

In planned comparisons, the differences in cut-off-to-repair duration between repair types were tested separately for within-word and after-word suspensions. Following after-word suspensions, the cut-off-to-repair interval was significantly longer for major repairs (580 ms, SD = 344) than for minor repairs (280 ms, SD = 106), ($t(9) = 2.705$, $p < .05$). In contrast, no significant difference between major repairs (175 ms, SD = 113) and minor repairs (154 ms, SD = 69) was observed for within-word suspensions ($t(9) = .578$, n.s.)

Finally, we tested whether minor as well as major repairs co-occurred with within-word suspension with 0-40 ms cut-off-to-repair intervals. Cut-off-to-repair intervals below or equal to 40 ms were observed in 39.2% (N = 176) of all cut-off-to-repair intervals. It was found that 9.5% (N = 19) of the major repairs (N = 200) and 19% of the minor repairs (N = 248), had within-word suspensions that were followed by 0-40 ms cut-off-to-repair intervals (see Table 3.6 below for the distribution of 0 to 40 ms cut-off-to-repair intervals by suspension type and repair type). This result was predicted by the DIP hypothesis but not by the MIR hypothesis.

Table 3.6. Repairs with 0 to 40 ms cut-off-to-repair interval following within-word and after-word suspensions.

		Suspension type		
Repair type		Within-word	After-word	Total
	Minor repair	48	59	107
	Major repair	19	50	69

3.3 Discussion

The present study investigated the mechanisms of error detection, replanning and speech suspension. The leading question was: do speakers interrupt their delivery immediately upon error detection as assumed by the Main-Interruption-Rule hypothesis (MIR hypothesis), or do they delay their speech suspension until the repair processing is in its final stages as assumed by the Delayed-Interruption-For-Planning hypothesis (DIP hypothesis)? In other words, what determines the timing of speech interruption: error detection or repair readiness? In order to test these competing hypotheses, we investigated the relationship of different speech suspension types (within-word vs. after-word), cut-off-to-repair interval duration, and different repair types (major vs. minor), for which the two hypotheses make different predictions.

The first analysis examined whether there was a difference between cut-off-to-repair intervals when preceded by a within-word suspension versus an after-word suspension. The corpus had a distribution of within-word suspensions (31.9%) and after-word suspensions (68.1%) similar to that obtained in Levelt (1983) with Dutch speakers (20.3% within-word, 79.7% after-word), suggesting that the method of data acquisition did not alter the data in this respect.

The analysis of cut-off-to-repair interval and suspension type revealed that the cut-off-to-repair interval was significantly shorter when speech was suspended within-word compared to after-word. This finding seems to support the DIP hypothesis, which assumes that in cases of within-word suspensions replanning has proceeded to the final stages of speech production when speech interruption is initiated. The major part of the replanning process must thus have taken place before the within-word suspension. In contrast, when speech is suspended after-word, some part of the replanning process must take place after speech suspension. Thus, the cut-off-to-repair interval is expected to be longer.

However, the existence of a possible confound leaves open an alternative explanation of these results that is consistent with the MIR hypothesis. Note that in the present study, following the MIR hypothesis, it was assumed that within-word suspensions are the result of immediate interruption upon error detection, while after-

word suspensions were assumed to be the result of delayed interruptions together with immediate interruptions that happened to coincide with the end of a word. In terms of error repairs and appropriateness repairs this means that repairs following within-word suspensions should all be error repairs, while repairs following after-word suspension could be a mixture of appropriateness repairs, and error repairs. It is possible that error and appropriateness repairs are correlated with major and minor repairs, in that error repairs are mainly minor repairs that can be processed fast, while appropriateness repairs are mainly major repairs that take more processing time. If in the analyzed data set a greater proportion of minor repairs followed within-word suspensions, the average cut-off-to-repair interval would be shorter than when speech was suspended after-word. This was indeed the case: in the present corpus, 51.5% of the after-word suspensions were followed by major repairs and 48.5% by minor repairs, while 30.1% of the within-word suspensions were followed by major repairs and 69.9% were followed by minor repairs. The fact that there were more major repairs following after-word suspensions lends plausibility to this explanation.

However, the MIR hypothesis cannot easily account for the results of the second analysis, which concerned the question of whether the cut-off-to-repair intervals differed for major and minor repairs. The results showed that the cut-off-to-repair intervals for major and minor repairs differed in the case of after-word suspensions: the cut-off-to-repair interval was shorter when a minor repair followed than when a major repair followed, which was predicted by both hypotheses. However, when the suspension was within-word the cut-off-to-repair intervals did not differ for major and minor repairs. The results suggest that in after-word suspensions the cut-off-to-repair interval reflects the duration of the replanning phase, while in within-word suspensions, it does not. This is not compatible with the predictions made by the MIR hypothesis, according to which the cut-off-to-repair interval should reflect the replanning time regardless of the suspension type. Rather, the present findings provide support for the DIP hypothesis, which holds that cut-off-to-repair duration does not reflect the replanning time when speech is suspended within-word, since replanning is about to be finished at the moment of cut-off.

The third and final analysis examined the types of repairs following very short cut-off-to-repair durations of less than or equal to 40 ms. The corpus included a considerable amount of such very short intervals, confirming the reports of very short cut-off-to-repair durations by Blackmer and Mitton (1991) for English, and Van Hest (1996) and Oomen and Postma (2001) for Dutch. More than a third (39.2%) of the overt repairs had cut-off-to-repair durations of less than or equal to 40 ms. This proportion in the present study is in agreement with the proportions reported by Blackmer and Mitton (1991), who observed zero ms cut-off-to-repair intervals in 19.2% of the overt repairs, and 0-100 ms cut-off-to-repair intervals in 48.6% of the overt repairs.

The analysis revealed that within-word suspensions with very brief cut-off-to-repair intervals in the present corpus were followed by minor as well as by major repairs, which is in line with the DIP hypothesis. Because the DIP hypothesis assumes that replanning starts upon error detection and speech is interrupted when the replanning is completed, 0 to 40 ms cut-off-to-repair durations are predicted by the DIP hypothesis, irrespective of the complexity of the following repair. In contrast, the number and the distribution of within-word suspensions and very short cut-off-to-repair times followed by major repairs are at variance with the MIR hypothesis. According to the original version of the MIR hypothesis (Levelt, 1983), the complete speech production system is interrupted upon error detection. The replanning starts only after the suspension, which means that there has to be a cut-off-to-repair interval of some length during which the replanning can take place. Thus, the original MIR hypothesis can in general not account for zero ms cut-off-to-repair intervals in cases of error repairs in which interruption should be immediate. For the modified version of the MIR hypothesis by Hartsuiker and Kolk (2001), zero ms cut-off-to-repair intervals associated with major repairs are problematic, since only 50 ms are allowed for replanning. It is unclear how short cut-off-to-repair intervals for major repairs with immediate interruption can be simulated successfully given the current set of assumptions in their model concerning the time required for error detection, speech suspension, and replanning.

In sum, the result that concerned cut-off-to-repair intervals when speech was suspended within-word as opposed to after-word is compatible with both hypotheses. While the result is predicted by the DIP hypothesis, it is not predicted by the MIR

hypothesis. Nevertheless, the MIR hypothesis can account for the result, because there is a preponderance of minor repairs associated with within-word suspensions. However, the findings on the length of the cut-off-to repair interval followed by major and minor repairs when speech is interrupted within-word are in line with the DIP hypothesis but not with the MIR hypothesis. The results indicate that speakers do not interrupt upon error detection but delay interruption in order to process the repair covertly. If the replanning is fast, speakers cut off their delivery immediately and articulate the repair. Thus, within-word suspensions do not reflect immediate interruption upon error detection followed by a cut-off-to-repair interval that reflects the replanning time, as suggested by the MIR hypothesis. Rather, these are cases where the suspension was delayed until the replanning had proceeded to final stages of production or had even come to completion.

A potential alternative explanation for the results might be attempted on the basis of the repair classification scheme used in the study. Note that in the classification of major and minor repairs a criterion was applied according to which only fresh starts, in which the original utterance is abandoned and an entirely new utterance is generated, were classified as major repairs. This is of importance with respect to the second prediction, which concerned the question whether the duration of cut-off-to-repair intervals differed for major and minor repairs. One might argue that all cases of repairs in which some element has to be generated anew constitute a major repair. This could be the case for additions, in which an element is added while the syntactic frame is maintained, and for substitution, in which one element is replaced by another element of the same category. Under this view, the classification procedure for major and minor repairs might have underestimated the proportion of repairs that are major and thus might have overestimated the difference in cut-off-to-repair times between major and minor repairs. Crucially, however, a reclassification of some minor repairs would not change the finding that cut-off-to-repair durations did not differ for within-word suspensions. This is because for this type of suspension even repairs involving complete replanning did not lead to longer cut-off-to-repair intervals than minor repairs in the present classification. Moreover, both the MIR hypothesis and the DIP hypothesis

predicted a difference in cases of after-word suspensions, which means that, with respect to this prediction, a possible bias would work for both hypotheses.

Further, a possible reclassification of some repairs as major would not change the findings with respect to the third prediction (see above, p. 71). This prediction concerned zero ms cut-off-to-repair intervals and their association with major and minor repairs when speech is suspended within-word. The MIR hypothesis predicted that when speech is suspended within-word, zero ms cut-off-to-repair intervals can only be followed by minor repairs. If anything, the present coding scheme potentially introduced a bias in favor of the MIR hypothesis in that it underestimated the number of major repairs. A re-classification of some minor repairs as major repairs would lead to an even higher number of major repairs following zero ms cut-off-to-repair intervals.

Given that the findings of the present study support the DIP hypothesis rather than the MIR hypothesis, it is now important to consider whether the DIP hypothesis can also account for findings from Levelt's (1983) original work that motivated the MIR hypothesis.

One of the main observations supporting the MIR hypothesis was that speakers interrupt speech at any given point regardless of linguistic boundaries. The same holds for the DIP hypothesis according to which the initiation of interruption depends on repair readiness. Speech is interrupted regardless of the stage of the word under articulation. In contrast, word completion is a result of running out of prepared material in the formulator and in the articulatory buffer before replanning has come to completion.

A second observation from the Levelt study (1983) was that suspensions occurred above chance at constituent boundaries. According to Levelt, this is because attention in monitoring is heightened towards ends of constituents. Therefore, errors are more often detected towards constituent-final positions and speech is suspended more often at these positions. The DIP hypothesis can account for this finding based on the assumption that constituents are planning units in speech production. This means that speakers are more likely to run out of prepared material at the end of constituents than within constituents. It is plausible that, if replanning is initiated in the middle of a constituent under

articulation, the speaker may already have buffered sufficient material to continue speaking until the constituent ends.

A third observation from Levelt (1983) that the DIP should be able to account for was that more erroneous words than merely inappropriate words and neutral words were interrupted within-word. Levelt (1983) interpreted repairs in which the speech suspension occurred after or within a neutral word following an erroneous word as being due to a delay in error detection. The result that more erroneous than inappropriate and neutral words were suspended within-word, was reconciled with the MIR by assuming that for pragmatic reasons inappropriate as well as neutral words are not interrupted. The DIP hypothesis can explain the data by assuming that the monitoring and replanning for appropriateness repairs is more complex and takes longer than for error repairs and therefore, speakers are more likely to run out of prepared speech material. What evidence is there to support this assumption? One possibility is that in Levelt's (1983) corpus the majority of appropriateness repairs were major repairs for which the replanning was not completed before the speaker ran out of prepared material, leading to the completion of the suspension word. This explanation seems unlikely, however, since as far as can be inferred from the distributions and the coding scheme in Levelt's study, most appropriateness repairs seemed to involve the substitution of a word. These substitutions would constitute minor repairs in the present study.

However, the assumption of greater complexity of processing underlying appropriateness repairs could still be viable even if it is not the appropriateness repairs themselves that differ in complexity, but the detection and decision processes that underlie such repairs. Others have assumed that appropriateness monitoring and repair is a relatively slow process compared to error monitoring and repair (Van Hest, 1996; Oomen & Postma, 2001). Speakers can probably detect errors faster than inappropriate expressions, since the incongruity between an intended and a formulated utterance is greater for errors than for inappropriate expressions. For an inappropriate expression to be detected, common ground and the previous discourse must be evaluated, including the potential ambiguity of an expression given the context, the use of the appropriate level of terminology, and the coherence with the previous discourse (Levelt, 1983).

These evaluations are centrally governed, high-level processes residing at the conceptualizer level, which is relatively slow.

After detection, speakers probably also evaluate how disruptive the inappropriate expression would be for the interlocutors' comprehension process (Berg, 1986, 1992) and decide whether or not to correct it. If the decision is made to correct, they must also decide how to correct, for example, with a more specific term or even a completely new message. Making such a correction involves evaluation of common ground and the previous discourse. In contrast, for an error the speaker knows the original intention and therefore does not need to engage in additional processing. Taken together, it is conceivable that the monitoring, detection, and the replanning of an inappropriate expression could take more time since more complex and high level decisions have to be made (Oomen & Postma, 2001; Postma, 2000; Van Hest, 1996). As a result, it is more likely that speakers run out of prepared material and suspend speech with a word completion.

Support for this argument comes from studies that measured the error-to-cut-off interval (from onset of the problematic element to speech suspension). In her study of self-monitoring of L1 and L2, Van Hest (1996) found that the mean error-to-cut-off times were significantly longer for appropriateness repairs than for error repairs (622 ms vs. 287 ms). Also, the overall mean repair time (from error to resumption) for appropriateness repairs was longer than for error repairs (1141 ms vs. 648 ms). Oomen and Postma (2001) obtained a similar result in a study that investigated the influence of increased speech rate on self-monitoring in a network description task. At a normal speech rate, the mean error-to-cut-off time was longer for appropriateness repairs than for error repairs (788 ms vs. 453 ms). When speech rate was increased, the overall mean repair times (error to resumption) decreased to the same duration for error and appropriateness repairs (555 ms), but the mean error-to-cut-off duration was around 100 ms longer for appropriateness repairs than for error repairs (408 ms vs. 311 ms).

These results suggest that compared to error monitoring and repair, appropriateness monitoring and repair takes more time. Speakers are then more likely to run out of prepared material and therefore the suspension words are completed more often. Conversely, if monitoring and repair is faster in the case of errors, it is more

likely that repairs are completed before the speaker runs out of buffered material. As a consequence, repairs would result more often in within-word suspensions.

In sum, the DIP hypothesis cannot only account for the data of the present study but also for evidence previously offered in support of the MIR hypothesis. This means that the DIP hypothesis' account of speech interruption in disfluent utterances is to be preferred over the MIR hypothesis suggested by Levelt (1983, 1989) and modified by Hartsuiker and Kolk (2001). Since the MIR hypothesis is tightly linked to Levelt's (1989; Levelt et al., 1999) production model, the question arises, whether the evidence in favor of the DIP hypothesis also constitutes evidence against the feedback mechanism of the speech production model of Levelt. In other words is the mechanism of triggering speech interruption upon repair readiness compatible with Levelt's model?

The cut-off-to-repair intervals following within-word suspension provide the clearest cases for assessing at what point during repair processing the interruption was initiated, because within-word suspensions should be the result of interruption upon repair readiness. The average cut-off-to-repair interval for within-word suspensions was 180 ms, with a range of 0-400 ms. When interruption is initiated upon completion of the preverbal message, it should result in a cut-off-to-repair interval of at least 175 ms, depending on the complexity of the message to be formulated. This is calculated as follows. Indefrey and Levelt (2004) estimate that it takes 425 ms from completion of the preverbal message to speech onset (see p. 47). Some of this processing can take place in parallel with the monitoring and interruption processes required to stop speech, which in turn, can be estimated to require about 250 ms: 100 ms for the self-monitoring process to verify the repair readiness,[4] and an additional 150 ms to complete the interruption process (Hartsuiker & Kolk, 2001). The remaining 175 ms (425 ms – 250 ms) of processing must be completed after speech suspension during the cut-off-to-repair

[4] This estimate is based on the following considerations: parsing of the phonological representation should take less time than Levelt's (1989) estimate of 150 ms for word recognition based on the phonetic plan (see also p. 25). Moreover, the verification of repair readiness does not necessarily require the recognition of the complete first phonological word of a repair utterance.

interval. This value corresponds well to the observed average cut-off-to-repair interval of 180 ms.

According to the perceptual loop theory of Levelt et al. (1999), the self-monitoring process has furthermore access to the output of phonological encoding via the internal loop (see section 2.1.1.3.2, p. 24 for details). It is assumed that the internal loop has access to a repair utterance after phonological encoding of its first syllable (Wheeldon & Levelt, 1995; Indefrey & Levelt, 2004). When interruption is initiated upon monitoring phonological encoding, zero ms cut-off-to-repair intervals should result. According to Indefrey and Levelt (2004), it takes 245 ms after phonological encoding of the first syllable for further syllabification and phonetic and articulatory encoding of the repair. This processing takes place in parallel with repair readiness verification (100 ms) and interruption (150 ms). Hence the repair is ready for articulation as soon as repair readiness verification and interruption are completed, resulting in 0 ms cut-off-to-repair intervals (245 ms – 250 ms). It can be concluded that the DIP hypothesis is in principle compatible with the speech production model and the suggested feedback loops of Levelt (1989; Levelt et al., 1999). While the average cut-off-to-repair interval can be explained by assuming that repair readiness is verified upon completion of the preverbal message of the repair, the model also provides the possibility to account for zero ms cut-off-to-repair intervals by assuming the repair readiness detection occurs via the inner loop.

So far we have discussed that the findings of the present study support the DIP hypothesis, which is based on the assumption that speakers strive to minimize the cut-off-to-repair interval by continuing speaking while planning the repair. These findings raise additional questions about the pragmatic factors motivating timely repair. Given the structural organization of turn taking in conversation (Sacks et al., 1974), speakers probably have an interest in minimizing the cut-off-to-repair interval since these intervals are potential points at which the interlocutor might initiate a repair or even execute it. Thus, the preference for self-initiation and self-repair in conversation (Schegloff, 1979) can be seen as one factor driving the strategy to maximize the covert replanning time and thereby to minimize silences and gaps in the speech flow. In addition, shorter silences and gaps in the speech flow avoid long deviations from the

main goal of the utterance, for example, telling a story, describing an apartment or making a request. By maximizing covert replanning, speakers repair within their turn. They thereby keep the space a repair takes and its prominence as small as possible. Moreover, they maintain the 'progressivity' of the sequence they are engaged in (Schegloff, 1979).

This kind of strategy, however, appears to risk confusing the listener by making erroneous information available for longer than necessary, and may even cause the listener to interrupt in order to, for example, correct or ask for clarification. However, it should be noted that as long as the speaker continues speaking, the likelihood of an interruption will remain low. Furthermore, speakers have resources at their disposal, which they can use to flag an erroneous fragment, and thereby minimize the risk of misunderstanding. One such resource are editing terms like 'no', 'I mean', or 'rather,' which signal the type of repair that a speaker is making (Levelt, 1983; Clark 1994).

What can be suggested here is that there is a fluency versus accuracy trade-off in conversation in which the factor of fluency is favored. Note, however, that there are probably circumstances in discourse in which speakers would wish to maximize accuracy. The decision about how disruptive, infelicitous, and socially consequential an error or an inappropriate expression is depends on the context in which it occurs. The decision of whether and when to interrupt is driven by a moment-to-moment evaluation of the resulting social impact and of the resulting impact on the flow of conversation. On the one hand, speakers will probably interrupt as soon as possible when they detect an inappropriate expression that can have socially drastic consequences, as the taboo word study by Motley et al. (1982) indicates. On the other hand, the relevance of a phonological error such as a 'cuf of coffee', for instance, might not be as consequential for the understanding within a given discourse.

It is an empirical question how flexibly the processing system can adapt to changing discourse circumstances. Some hints to its flexibility can be found in individual differences in the speakers in the corpus. One speaker (speaker NI) spoke very slowly and had a preponderance of very long cut-off-to-repair intervals. Notably, the vast majority of his disfluencies were repairs in which only an *eh* ('uh') or an *em* ('um') was uttered and the utterance was continued. This suggests either a strategic

decision to maximize accuracy within the interactional situation, or simply an acquired disposition to speak accurately. In contrast, another speaker (speaker SE) showed a pattern that was very much the opposite, making about twice as many errors, with many overt repairs in which elements are altered in the resumption and repetitions and very short cut-off-to-repair intervals. This speaker seemed to be maximizing fluency.

In sum, the DIP hypothesis provides a parsimonious account of the results from the present study and the evidence that motivated the MIR hypothesis. Of course, speakers do sometimes interrupt speech immediately, but it seems that immediate interruption is not the default in conversation. The results suggest that speakers interrupt when the repair processing is in its final stages, so that the cut-off-to-repair duration can be kept minimal. Speakers have various strategies at their disposal to handle malfunctions in speech production in conversations. Minimizing gaps and silences caused by replanning is just one of them.

4. Gestures and speech disfluency

Chapter 4

4.1 Introduction

The aim of this chapter is twofold: we will investigate whether speech-accompanying gestures are sensitive to speech disfluency and whether gesture can provide evidence for the self-monitoring process in speech production. The previous chapter investigated the processes underlying the self-monitoring and the interruption of speech. However, during conversation, speakers not only speak but also gesture. This raises the question of what happens to speakers' gestures when they encounter problems in speech. Currently little is known with regard to this question. For this reason, the current chapter investigates the effects that speech disfluency in the corpus had on concurrent gestures. In addition, we will investigate whether the gestural data provide further evidence for the processes underlying the interruption of speech.

The previous chapter as well as some of the studies discussed there investigated the temporal characteristics of self-monitoring, self-interruption and repair to infer underlying processing mechanisms and strategies. A shortcoming of these studies is that the moment of error detection cannot be directly observed. The disfluent utterance does not exhibit an unambiguous indication at which point in time the error leading to speech suspension was detected. Hence, studies on the time course of self-monitoring and repair have investigated two other measures, the error-to-cut-off and the cut-off-to-repair interval. They have taken the error-to-cut-off interval as reflecting to some extent detection latency and the cut-off-to-repair interval to some extent repair processing time. However, the results of the previous chapter have shown that at least in some

cases the error-to-cut-off interval includes some amount of replanning time. Similarly the results have indicated that the cut-off-to-repair interval may only correspond to a part of the replanning time. In order to provide further evidence as to whether the speech interruption process is initiated upon error detection, delayed for word completion when the word under articulation is not erroneous (MIR hypothesis), or delayed for repair planning (DIP hypothesis), a further measure of the time lag between error detection and speech suspension is required.

Here we explore the possibility of using speech-accompanying gesture as a further source of evidence for the processes underlying self-monitoring and speech interruption. We will start by providing an example of a speech-accompanying gesture illustrating the structural organization of gesture.

4.1.1 The structural organization of gestures

Gestural movement sequences can be broken down into discrete so-called *gesture phases* based on formal and functional features (Kendon, 1972, 1980; Kita, van Gijn, & van der Hulst, 1998; McNeill, 1992). Consider a gesture depicting a spiral staircase as shown in Figure 4.1 below. The speaker lifts one hand from the lap up to chest height and holds it there briefly. From there she rotates the hand with an extended index finger while moving it up in front of her forehead. The hand is held briefly and then drops back onto the lap.

a) Preparation b) Pre-stroke hold c) Stroke d) Post-stroke hold

Figure 4.1. Phases of gestural movement depicting a spiral staircase.

In this example, the hand departs from its resting position on the lap (see Figure 4.1, a) to the location up to the height of the chest, from where the expressive part, the stroke, is going to be deployed. This phase is referred to as the preparation phase. While rising, the hand shape and the orientation of the hand are also prepared in such a way that the index finger is extended pointing upwards, the other fingers are bent and the palm is facing to the side. The preparation phase is followed by a static phase, the so-called pre-stroke hold, where the hands are held still in the preparation-final and stroke-initial position. In the example, the hand stops moving as it arrives at chest height (see Figure 4.1, b). The pre-stroke hold is then released by the stroke, which constitutes the expressive phase (see Figure 4.1, c). The stroke displays the meaning of the gesture. In the given example the arm moves upwards while rotating the hand from the wrist, depicting the spiral shape of the staircase. The stroke is again followed by a static phase, the post-stroke hold. As the rotating hand reaches forehead height, it stops moving. The hand is held in the air with the index finger extended (see Figure 4.1, d). The gesture ends with the retraction of the hands back into resting position, for example, the hand drops back on the lap. Note that the hands may be only partially retracted. For example, a hand held in the air with the index finger extended, drops down to chest height while the hand shape is released. This phase is called a partial retraction, since the gesture is only partially retracted by releasing the hand shape and dropping down to chest height but not back onto the lap, the full resting position. Furthermore, pre-stroke holds as well as post-stroke holds as described in this example do not always occur. In natural conversation one can observe a succession of strokes without the hands going into a hold in between or being retracted after each stroke.

The succession of the gesture phases results in the following sequential organization (Figure 4.2).

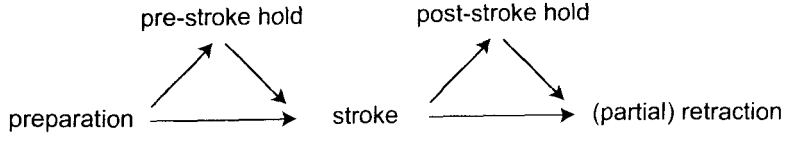

Figure 4.2. Sequential organization of gesture phases

4.1.2 The temporal and semantic co-ordination of gesture and speech

In order to use gesture as an additional source of evidence it has to be established first whether gesture is sensitive to speech disfluencies. There is reason to believe that this might be so. The modalities are temporally and semantically coordinated, such that the meaningful part of the gesture is coordinated with the co-expressive parts of speech. Moreover, gesture execution adapts to features of speech like the location of stress. These kinds of evidence for a close coordination between gesture and speech suggest that the modalities are cognitively linked, such that information exchange about the time course of their execution is possible.

Let us first review studies investigating how gesture and speech are temporally and semantically interrelated. Studies on the temporal organization of gesture and speech have demonstrated that gesture is synchronized with the semantically and pragmatically co-expressive portion of concurrent speech (Kendon, 1993; McNeill, 1992; Morrel-Samuels & Krauss, 1992). The synchronization of gesture and speech indicates that there is direct or mediated information exchange between the modalities about the time course of their execution. Kendon (1993) illustrates the temporal and semantic organization with an example in which the speaker retells the fairy tale Little Red Riding Hood. The example shows how the stroke, the meaningful part of the gesture is coordinated with the co-expressive part of speech itself (Kendon, 1972, 1980; Kita, 1993; McNeill, 1992; Schegloff, 1984). The speaker says "and he took his hatchet and with a mighty sweep sliced the wolf's stomach open." (Kendon, 1993, p. 45). During this utterance the speaker produces two different gesture strokes. In the first gesture stroke, the speaker raises her hands as if holding the hatchet into an upraised position slightly before she starts saying *and he took his hatchet and with*. She holds the hands still at that location while she says *a mighty sweep*. As she says *sliced*, she performs the second gesture stroke, a sweeping motion with her hands to the left. The movement is meaningful in that it presents an enactment of the hunter's action of slicing the wolf's stomach open. The example illustrates the characteristic organization of gesture and speech. In order to coordinate the meaningful part of the second gesture, the stroke, with the co-expressive part of speech (*sliced*), the speaker holds her hands in

upraised position until she starts articulating the co-expressive part of speech *sliced*. Only then does she execute the slicing motion, the gesture stroke.

The phenomenon of the *pre-stroke hold* motivated the explanation that gesture "waits for speech" in order to achieve temporal coordination with the co-expressive elements in speech (Kita, 1993). Since gesture, unlike speech, does not have to undergo complex grammatical encoding, gesture encoding is faster and gesture execution can start earlier than speech (McNeill, 1992). Therefore, gesture preparations are pre-positioned with respect to the co-expressive part of speech. Depending on where in speech the co-expressive portion will come, a preparation can be followed by a pre-stroke hold in order to synchronize the stroke with the respective co-expressive part of speech. Hence gesture execution is temporally adjusted to the linear structure of speech, in which co-expressive elements occur in the middle or towards the end of the utterance.

Secondly, there is evidence that gesture adapts to changes in features of speech like speech onset (Levelt, Richardson, & La Heij, 1985) and the location of contrastive stress (De Ruiter, 1998). Gesture-speech synchronization is maintained when temporal parameters of speech onset in speech production are experimentally manipulated (Levelt et al., 1985). This indicates that the gesture production process is informed about the timing of speech production. Levelt et al. (1985) investigated the nature of the synchronization process of pointing gestures and deictic expressions. Participants had to point to lights at different distances and refer to them with deictic expressions (*this light*). Speech adapted to features of the gesture execution in that speech started later when the pointing gesture started later because it was directed at a light further away. Gesture also adapted to features of speech planning and execution. In one experimental condition speakers had to choose between two deictic expressions (*this/that light*) which prolonged speech planning. When speech planning was prolonged, gesture initiation was also delayed, indicating that gesture also adapted to speech.

The synchronization of speech and gesture is also maintained when the locus of contrastive stress in a complex noun phrase is manipulated (De Ruiter, 1998, 2000). In a study of pointing gestures in which speakers had to point to and name a picture, De Ruiter (1998) varied the location of contrastive stress in utterances participants had to produce (the GREEN crocodile vs. the green CROcodile). The gesture was initiated

earlier when contrastive stress was located earlier in the utterance, namely at the adjective, than when it was located later in the utterance, at the noun. Also, the launching movement of the gestures was prolonged when the stress came later in the utterance, providing evidence that gesture adapted to the characteristics of speech.

4.1.3 Models of the integration of speech and gesture

The observed temporal and semantic coordination has led to various accounts concerning the cognitive linkage of the two modalities. Investigators have proposed that the modalities are linked either both in working memory and at the formulator level, or only at the conceptualizer level, or throughout all levels.

According to the first view, which assumes a link in working memory and at the formulator level, gestures are an epiphenomenon of the lexical retrieval process, in that they facilitate lexical retrieval via cross-modal priming at the formulator level of speech production (Hadar & Butterworth, 1997; Krauss, Chen & Chawla, 1996; Krauss, Chen, & Gottesman, 2000; Rauscher, Krauss, & Chen, 1996). Proponents of this account restrict their claims to spontaneous non-conventionalized gestures that resemble some part of their referent iconically, so-called lexical gestures. According to Krauss et al. (2000), these gestures originate from representations in working memory, which will be also expressed in speech (see Figure 4.3 below). A memory representation, the so-called source concept, combines a set of features in propositional and non-propositional representational formats. Propositional representations of features are transformed into linguistic structures, while spatial/dynamic representations of features are transformed into gestures. The Spatial/Dynamic Feature Selector translates spatial/dynamic representations into a set of abstract movement properties. These specifications are the input to the Motor Planner module. The Motor Planner transforms the specifications into a motor program that consists of a set of instructions for movement execution. These instructions are the input of the Motor System, which translates them into overt gesture, the lexical gesture. The lexical gesture is kinesthetically monitored by the Kinesic Monitor. Mediated by the Kinesic Monitor, the lexical gesture is fed into the phonological encoder. Specific features of the source concept, which are represented motorically, then facilitate word-form retrieval by cross-modal priming. Krauss et al. (2000) do not further specify how a word-form is primed by a gesture.

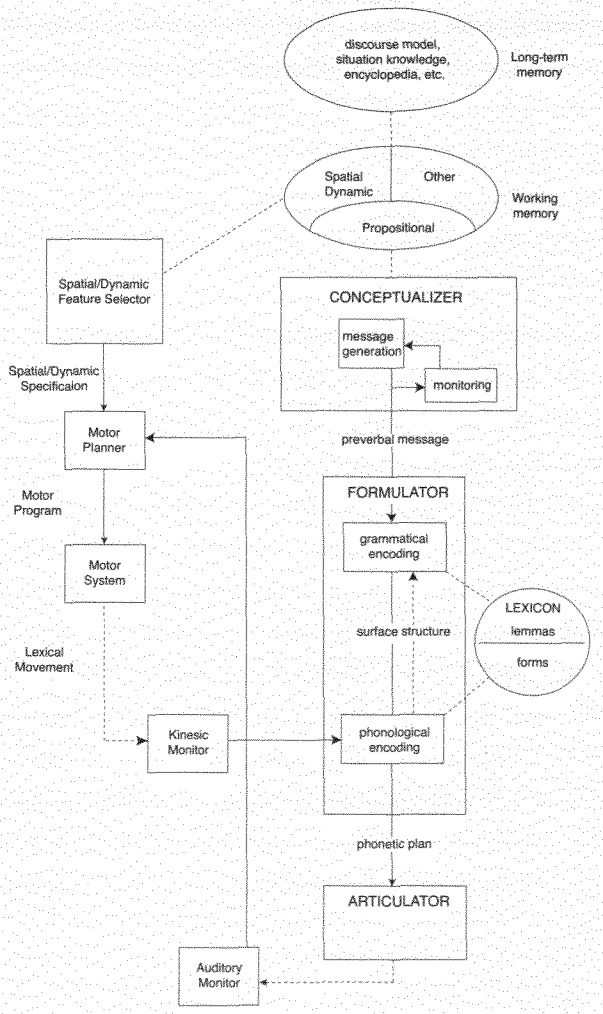

Figure 4.3. Krauss et al.'s (2000) model of gesture and speech production.

Gesture is terminated in reaction to uttering the co-expressive part of speech. Krauss et al. (2000) assume an auditory monitor for overt speech that is linked to the Motor Planner. As soon as the speaker hears via the auditory monitor that the co-expressive part of speech is being uttered, the gesture is terminated. This termination mechanism entails that gesture is reactive to events in speech.

A shortcoming of this account is that the claims are restricted to only one type of gesture. No specific account is provided for synchronization phenomena, such as the pre-stroke or the post-stroke hold. It remains unclear why gesture is sometimes terminated with a retraction and sometimes with a hold, and how these different types of suspensions are planned and executed. Moreover, no account is given about how gesture and speech might interact in the case of problems in speaking that are not related to lexical retrieval problems.

A different theory is proposed by McNeill (1992; McNeill & Duncan, 2000), who assumes that gesture and speech form a fully integrated system and that the modalities interact throughout their production. This approach does not provide an information-processing model with distinct processing steps in which mechanisms are explicitly specified. Instead, McNeill and Duncan (2000, p. 155) see gesture and speech as "embodied cognition" and not as outputs of separate cognitive production processes. McNeill (1992) puts forth the so-called *growth point theory*. In contrast to Krauss et al. (2000), who only account for lexical gestures, the growth point theory accounts for a broader range of types of speech-accompanying gestures. On the one hand, these comprise *iconic* (equivalent to Krauss et al.'s (2000) lexical gesture) and *metaphoric gestures*. These are spontaneous and idiosyncratic creations of the speakers representing iconically aspects of their concrete (*iconics*) or abstract referents (*metaphorics*). On the other hand, they comprise *deictic gestures*, which indicate objects and events, and *beat gestures*, which can be characterized as short bi-phasic up- and downward or back- and forth-movements.

According to the growth point theory, a growth point is the minimal psychological unit of thinking. The information expressed by a growth point is the newsworthy element against the background of its immediate context, and it combines imagistic and linguistic categorical contents that are semantically and pragmatically related. A growth point is a 'seed' of an utterance consisting of co-expressive speech and gesture. The 'growth' of the seed is a dynamic dialectic process between the imagistic and linguistic thinking. The imagistic part of the growth point grows into the gesture stroke. The linguistic part becomes the words that are synchronized with the gesture, and these words are the starting point for the linguistic development of the

utterance. The utterance linguistically 'grows' around these words, such that the utterance is syntactically correct and fits into the discourse context.

The growth point theory provides the following explanations for different phases of gesture production (McNeill, 2000, p. 322-323). Onsets of gestural movements (preparations) represent the moment where the growth point begins to form. Strokes automatically synchronize with the linguistic unit of the respective growth point. Pre-stroke holds occur when the stroke of a gesture is delayed, since other elements in speech have to be articulated first. Holds after strokes occur when the co-expressive parts in speech are longer than the actual stroke.

The growth point theory entails that synchronous gesture and speech are semantically and pragmatically co-expressive. Moreover, since according to McNeill (1992; McNeill & Duncan, 2000) gesture and speech constitute a single inseparable system and evolve together into a multi-modal utterance, it can be assumed that a disfluency in one modality should affect the other modality.

Other investigators have proposed that gesture production is linked to speech production at the level of the conceptualizer (Kita & Özyürek, 2003; De Ruiter, 1998; Levelt et al., 1985; Melinger & Kita, 2001). De Ruiter (1998; 2000) proposes the so-called *sketch model*, a model of gesture speech linkage at the conceptualization level, where the speech monitor also resides (Levelt, 1983). This model is an extension of Levelt's (1989) speech production model (see Figure 4.4 below). It specifies the mechanisms underlying the temporal and semantic gesture-speech integration for the broadest range of gesture types, compared to the accounts of McNeill (1992) and Krauss et al. (2000). This is because in addition to all types of speech-accompanying gestures, it includes pantomimic gestures as well as conventionalized gestures such as the thumbs-up gesture.

In the sketch model it is assumed that gesture is part of the communicative intention of the speaker. In the conceptualizer information intended to be expressed is assigned to gesture and speech. Propositional content is converted into a preverbal message, while simultaneously imagistic content is converted into a so-called sketch.

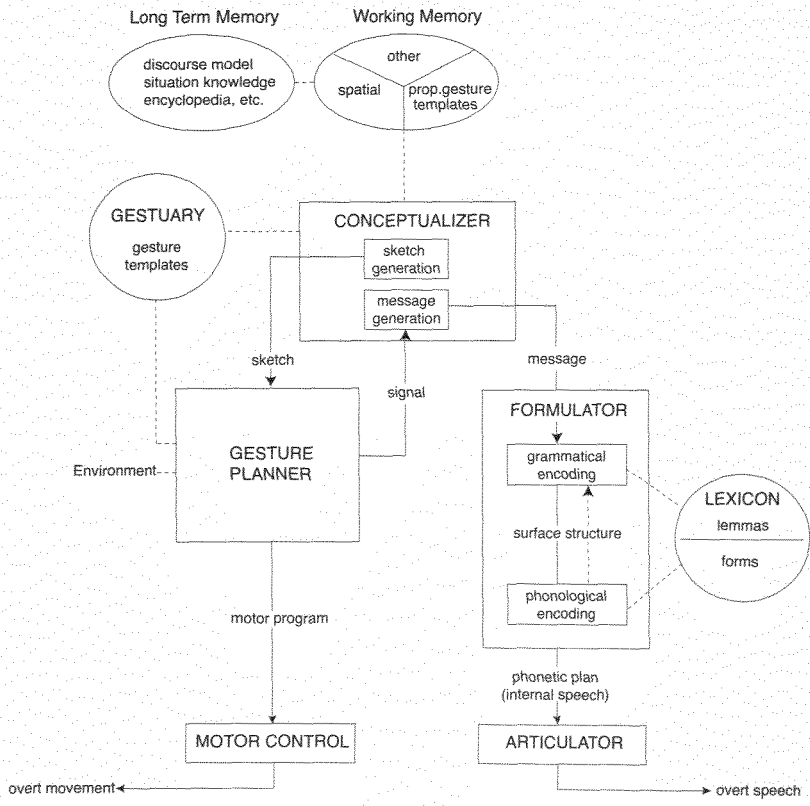

Figure 4.4. De Ruiter's Sketch Model of integrated gesture and speech production.

In case of a conventionalized gesture like the thumbs up gesture, the conceptualizer accesses the Gestuary, which is a repository of conventionalized gestural shapes. In case of an iconic gesture the shape of the gesture is determined by an imagistic representation. The Sketch is then sent to the Gesture Planner where the respective motor program is generated. The motor program is sent to the Motor Control module, which transforms the motor program into overt gestural movement.

The synchronization of gesture and speech is achieved by a feedback loop providing information about the processing stage of gesture from the Gesture Planner to the conceptualizer. Synchronization is coordinated at the conceptualizer, where the

speech monitor is located in Levelt's model (1989; Levelt et al., 1999). Thus, information about the timing of processing steps of both speech and gesture available to the conceptualizer through feedback loops can be utilized to coordinate the modalities temporally.

The Sketch Model of De Ruiter (2000) also provides an account for cases of error detection. As an error is detected in speech via the internal or external monitoring loop, a stop signal is sent to the formulator and to the gesture planner module. These modules pass the signal on to the lower modules, which then interrupt gesture and speech. The fact that the conceptualizer assumes the role of coordinator of the modalities means that the gesture-speech relationship is flexible. In other words, synchronization is not fixed but can be adapted to a speaker's communicative goals in a given situation.

In sum, De Ruiter's (1998; 2000) and McNeill's (1992; McNeill & Duncan, 2000) accounts of gesture-speech coordination entail that speech and gesture production can be adjusted to each other online to a certain degree. In contrast, this possibility is not given following Krauss et al. (2000), since his model only accounts for gesture initiation and termination upon hearing what is being said, which means gesture can only be reactive to events in overt speech.

4.1.4 Gesture in disfluent utterances

Although there are only a few studies that have investigated gestural behavior in cases of disfluent speech, different types of disfluencies and repairs in speech have been found to have a direct impact on concurrent gestures. In this section we consider previous research on this topic.

The investigation of gestural behavior in stutterers provides evidence that stuttered disfluencies affect concurrent gestures. Mayberry and Jaques (2000; see also Mayberry, Jaques, and Dede, 1998) found that during stuttered disfluencies, gestures were rarely co-produced. In the few cases where gesture was co-produced, the gesturing hand either fell to rest or stopped moving (went into a hold by freezing the ongoing gestural movement) during the moment of stuttering. The gesture was resumed by rising from rest position again or by releasing the hand from a hold "within milliseconds"

(Mayberry & Jacques, 2000, p. 206) of resumption of speech fluency. In some cases the gesture that fell to rest during the moment of stuttering, was abandoned by remaining in rest position when speech fluency was resumed. Temporal details about when the gesture was retracted/frozen in relation to the stuttered disfluency are not reported.

Gestures also appear to be affected by non-pathological disfluencies. Ragsdale and Silvia (1982) investigated "kinesic hesitation phenomena" (head, hand, arm, leg, and foot movement, posture change and body shift that occur in conjunction with vocal hesitation phenomena). The majority of the kinesic hesitations occurred just before or simultaneously with the speech disfluency (sentence change, repetition, stutters, omissions, sentence incompletion, tongue slip, intruding incoherent sounds, excluding filled pauses) but rarely after. Ragsdale and Silvia do not describe in detail what they define as a kinesic hesitation phenomenon, nor do they give a temporal measure.

When there is co-occurrence of gesture and disfluent speech, the temporal synchrony between a gesture stroke and its co-expressive part in speech can be maintained, even when speech is severely disfluent. This has been shown by McNeill (1992), who impeded speech performance by delayed auditory feedback. While participants were recounting a cartoon, they heard their own voice played back to them with a delay of 200 ms. The manipulation led to slowing down speech, stammering and stutters. Gesture strokes were still synchronized with the co-expressive part of speech. While speech was hampered by multiple disfluencies, gesture was withheld until the co-expressive part of speech could be uttered so that the stroke could be executed synchronously.

Similarly, for pointing gestures, the temporal synchronization with speech seems to be maintained in the face of disfluency through temporal adjustments of gesture. In a temporally fine-grained study, De Ruiter (1998) analyzed the synchronization of speech and pointing gestures in order to test the predictions of his Sketch Model (see Figure 4.4). Participants were instructed to point to near or further away pictures and to name the color and object of the picture. In some of the trials, participants produced disfluencies (11 self-interruptions followed by a repair and 17 hesitations between words). For these 28 disfluent trials, De Ruiter found that speech onset started on average 166 ms later than in the fluent trials (1021 ms vs. 1187 ms), even though the

disfluency occurred after speech onset. This suggests that speakers delayed speech onset for a certain amount of time but then started speaking although the problematic lexical element had not yet been retrieved.

For the gesture analysis of the 28 disfluent trials, De Ruiter measured onset and duration of the launching movement and onset and duration of the apex (the hold of the pointing hand) in relation to the co-occurring disfluent speech. These values were then compared to the corresponding values for fluent trials. This analysis revealed that the onset and the duration of the launching movement as well as the onset and the duration of the apex were adapted to the timing characteristics of the co-occurring disfluent speech. The onset of the pointing gesture adjusted to the delay of speech onset in that gesture was initiated in the disfluent trials on average 67 ms later than in the fluent trials (628 ms vs. 695 ms). Furthermore, the duration of the launching movement was on average 117 ms longer in the disfluent trials as compared to the fluent trials (691 ms vs. 808 ms). This total delay of 184 ms (67 ms + 117 ms) was sufficient to synchronize the gestural apex with the onset of speech. As a result, the temporal interval between speech onset and apex in the disfluent trials was nearly identical to the corresponding interval in fluent trials (316 ms and 298 ms respectively). In addition, the duration of the apex was prolonged by 211 ms, which closely corresponds to the 204 ms duration of the hesitation. The results indicate that planning and execution of the gesture took into account the amount of delay in speech.

Gesture suspensions seem to co-occur with speech suspensions, suggesting that speech disfluency might influence gesture execution. Kita (1993) analyzed self-interruptions in speech and gesture for repairs and repetitions. Repetitions were defined as cases where a portion of a sentence is repeated without any alteration (e.g., *there is a a shock*, example from Kita, 1993, p. 62). Repairs were defined as involving the alteration of speech in the resumption as compared to the original delivery. Eight participants contributed 54 repairs and 34 repetitions, which were accompanied by gestures. Kita investigated if gesture was suspended when speech was suspended followed by a repair vs. a repetition. To provide a baseline for comparison, Kita assessed the frequency of gesture suspensions in fluent utterances (for each participant, half as many fluent baseline utterances as disfluent utterances). In the fluent baseline

utterance, he randomly picked one word and checked if the concurrent gesture was terminated before the following word began. Kita found that in speech repairs an accompanying gesture phrase was more likely to be terminated before the resumption word began than in the fluent baseline utterances. By contrast, in repetitions the accompanying gesture phrases tended not to be terminated before the resumption word began. Kita deployed the same analysis for a subset of 10 repetitions and 22 repairs, investigating whether a gesture stroke was suspended when speech was suspended. Compared to the baseline, gesture strokes in speech repairs were more likely to be terminated before the resumption started. In contrast, for repetitions, there was no evidence that the stroke was terminated before the onset of the resumption in comparison to the baseline. However, these results should be interpreted with caution in the light of the small number of data points. Moreover, the fact that gesture suspensions were more likely before speech resumptions in repairs than in baseline utterances might be due to the fact that speech suspensions are often followed by a pause. This pause before the resumption provides more time for gesture to be terminated in disfluent utterances than in fluent utterances without pauses.

Taken together, the majority of researchers assume some type of gesture reaction to speech disfluencies. We will now provide some examples from the corpus of the present study illustrating that gesture can indeed react to speech disfluencies but that the types of gestural reactions may differ. Furthermore, we will exemplify that gestural reactions cannot be observed in all cases of speech disfluencies.

4.1.5 Examples of gesture suspensions accompanying speech disfluencies

In the first example, the speaker is describing the location of a hallway to the left of a door. In her verbal utterance she confuses the words *links* ('left') and *rechts* ('right'). She disrupts the word *rechts* ('right') with a glottal stop resulting in the fragment *re* and repairs it immediately with the correct word *links* ('left'). Speech disfluency and repair have a direct impact on the concurrent gesture (see Figure 4.5). The speaker executes a deictic gesture but interrupts it midway by pulling back to the starting position. She then deploys the same gesture again.

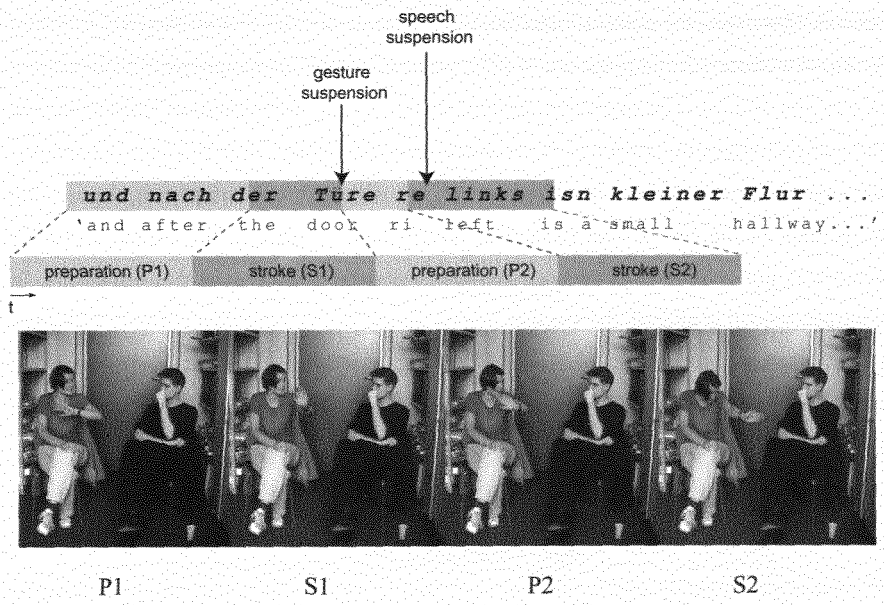

Figure 4.5. Temporal patterning of gesture suspension and speech suspension. The arrows indicate the moment of gesture suspension and of speech suspension.

The speaker prepares for a direction-indicating gesture by moving her left hand from the lap up in front of her chest, with the back of her hand facing towards her (P1). The preparation takes place while she says *und nach der* ('and after the'). She then extends her arm to deploy the stroke, by which she indicates the direction to her left, as she says *der Tuere* ('the door') (S1). Midway, before the hand has reached the end position indicating the direction left, she pulls the hand back towards the starting position in front of her chest (P2). As she utters and interrupts mid-word *re* ('ri') and corrects immediately with *links* ('left'), the gesture is repeated and this time completed, by extending the arm and turning the hand to the left (S2).

The above example illustrates how gesture can be affected by speech disfluency. The error *rechts* ('right') must have been detected in a prearticulatory phase, since *re* is interrupted 150 ms after its onset, which would be too short for detection via the external auditory monitoring and subsequent interruption. The speaker interrupts the correct gesture stroke, re-prepares and restarts the stroke again, before she resumes

speech. Moreover, the gestural reaction, the suspension of the first stroke (S1) followed by an immediate second preparation (P2), precedes the moment speech is suspended at *re* ('ri') by 160 ms. The suspended gesture stroke (S1) up to the point of suspension and the completed gesture stroke (S2) are performed identically. In both cases the speaker extends her arm and turns her hand to her left, although not completing the turning movement in the first stroke (S1). The fact that the shape and the execution of the gesture is not changed but repeated indicates that the gesture was not erroneous. In contrast, speech is altered, the speaker begins saying *re* ('ri') but then corrects with *links* ('left'). Gesture is not altered but repeated while speech is altered, which suggests that the detection of the error in speech triggered the gesture suspension and the restart of the gesture even before the erroneous fragment *re* was uttered.

In the next example, the speaker starts out with a construction, which she then abandons. She says: *genau man ging dann* (120 ms) *es war ja eigentlich auch ne ganze Wohnanlage* ('exactly one went then (120 ms) it was actually also an entire housing complex'). She pauses for 120 ms between her speech suspension after *dann* ('then') and the resumption (indicated in the transcript by the number in brackets). She starts out preparing a gesture with both hands, which she then abandons by retracting the hands back into rest position (see Figure 4.6 below, note that the gestural movement of each hand is exemplified separately).

In the example the right hand is suspended twice, first after the preparation and then after the hold, while the left hand is only suspended once, after the preparation. As the speaker starts out saying *genau man* ('exactly' followed by the German impersonal pronoun *man* ('one')), she raises her right hand from the lap up to chest height, with the back of the hand facing upwards (P2). She halts the hand at that location (H1) while she says *ging* ('went'). This is the first suspension of the right hand. The onset of this first suspension begins 560 ms before speech suspension. At the same time she moves her left hand from her face down on the raised right hand (P1). As she says *dann* ('then'), both hands retract by folding and dropping back to rest position onto her lap (R1 & R2). She suspends speech after the word *dann* ('then'). Thereafter the speaker starts out with a new construction talking about the entire housing complex. The prepared gesture stroke is not deployed but the hands drop back into resting position. In this example, the

onset of the gestural suspension—both hands dropping back into rest position—is temporally prepositioned by 240 ms with respect to the moment of speech suspension occurring after the word *dann* ('then').

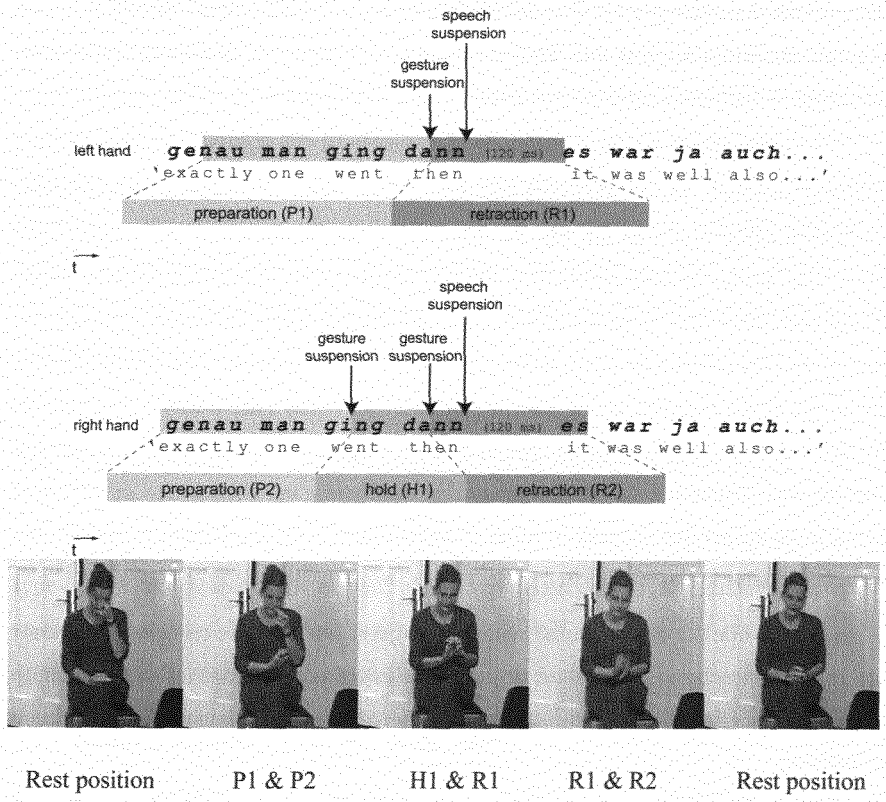

Figure 4.6. Temporal patterning of speech suspension and gesture suspensions of left (upper panel) and right hand (lower panel). The arrows indicate the moment of gesture suspension and of speech suspension.

This type of qualitative evidence suggests that speech disfluency affects gesture execution such that gestures are suspended when speech is suspended. It furthermore indicates that gesture suspension can be earlier than speech suspension. This suggests that error detection in speech can trigger gesture suspension at an earlier point in time than speech suspension. Hence, gesture suspensions might be closer in time to the moment of error detection than speech suspension.

However, only a minority of gesture suspensions is of the type described above, in which there is an unambiguous relationship between the gesture suspension and the speech suspension. The majority of gesture suspensions in disfluent utterances are suspensions that regularly occur in fluent utterances, for instance a pre-stroke hold (see example in section 4.1.2, p. 84). Thus it cannot be assumed that all gesture suspensions occurring in disfluent utterances are caused by the disfluency.

Consider the following example in which we can observe a pre-stroke hold in a disfluent utterance. The speaker says *und dann kam erst mal en To ehm kam ne Einfahrt* ('and then came first a ga um came a driveway'). As she says *und dann kam* ('then came') she brings up both hands from the rest position next to her face with the palms facing each other. She stops the movement and holds the hands still while she says *erst mal en To ehm* ('first a ga um'), where she interrupts within-word. She finally deploys the stroke as she resumes saying *kam ne Einfahrt* ('followed a drive way'). She extends her arms in front of her. For this type of gesture suspension it is not clear whether the gesture was suspended due to the upcoming disfluency or whether the gesture was going to be suspended anyway in order to synchronize with an element in speech that was uttered later in the utterance.

Thus, while the observational evidence suggests that there are cases in which error detection in speech triggers gesture suspension, this kind of evidence is not sufficient to unambiguously identify a relationship between a gesture suspension and a speech suspension.

4.1.6 Summary

To summarize, gesture and speech are temporally synchronized and semantically coordinated. The preparation phase of a gesture is prepositioned with respect to the co-expressive part of speech and the gesture stroke starts slightly before or synchronous with the co-expressive speech part. Moreover, gestures adapt to features of speech, such as speech onset and the location of contrastive stress. This suggests that there is information exchange between the modalities about the timing of respective processing steps. These synchronization phenomena can be accounted for by assuming a link at the conceptualizer level, where the speech monitor resides also (Levelt, 1983; 1989).

De Ruiter (1998) proposes that as the monitor detects an error, a stop signal is sent to the Gesture Planner module, which then passes it on to the motor level. The studies on gestures and disfluent speech reviewed above indicate that such a kind of coordination might in fact be employed, and that speech disfluencies affect gesture production and execution. However, the scope of the studies on gesture and speech disfluencies is limited in that either the kinds of hesitation phenomena are not described in detail, the number of observations is low or not given at all, or the temporal assessment of gestural responses to speech disfluencies is not very fine-grained. In order to overcome these limitations, the present study aims to establish quantitatively whether gesture is sensitive to speech disfluencies. If it can be shown that gesture is sensitive, the results of the gesture analysis can also be discussed with respect to the MIR hypothesis and the DIP hypothesis.

4.2 Corpus study 2: Gesture suspension in disfluent utterances

This section presents a quantitative corpus analysis examining the frequency and the types of gesture suspensions as well as the timing of gesture suspensions relative to speech suspensions. In the first two sections we will introduce the rationale for the quantitative analyses pursued in this study.

4.2.1 Gestural sensitivity to speech disfluency

If gesture is sensitive to speech disfluency, this should in some way be reflected in gesture suspensions. There are three ways in which such an effect of speech disfluency on gesture suspension might become manifest in the data. First, speech disfluencies might lead to additional suspensions of gesture. Hence the frequency of gesture suspensions in disfluent utterances might be higher than in fluent utterances.

Second, speech disfluencies might lead to a different timing of gesture suspension in disfluent and fluent utterances. If we assume that gesture and speech production are linked at the conceptualizer level (De Ruiter, 1998; Kita & Özyürek, 2003; Levelt et al., 1985) and if we assume with Levelt (1983, 1989) that speech monitoring takes place at the conceptualizer level, it is plausible that the gesture

production system can receive input about error detection and speech suspension from the monitoring and repair process. Hence, error detection or speech suspension might trigger gesture suspension and we might be able to observe a systematic relationship between the timing of gesture and speech suspension. For example, the detection of an error during speech monitoring might function as a trigger for gesture suspension. If error detection functions as a trigger, gesture and speech might stop at the same time. It is also conceivable that upon error detection a stop signal is sent first to the gesture production system and subsequently to the speech production system. In this case gesture suspension might systematically take place before speech suspension. Alternatively the trigger for gesture suspension could be the actual speech suspension itself. In this case gesture would systematically stop after speech.

Third, the position of suspended gesture phases might differ between disfluent and fluent utterances. Gestural suspensions can be categorized as *early* or *late* depending on their position with respect to the stroke phase, which expresses the gestural meaning. A gesture suspension can be considered to be *early*, if it occurs before the completion of a stroke; for example, in the middle of a preparation, after the preparation, or in the middle of a stroke. Such suspension preempts or stops the stroke. A gestural suspension can be categorized as *late*, if it occurs immediately after the stroke phase has been completed; for example, when a hold after a stroke is suspended by being retracted into rest-position. Assuming that speech disfluencies result in gesture suspensions, these suspensions may not respect the temporal order of gesture phases and hence interrupt gestures in relatively early phases of gesture execution. This might lead to a relative increase of *early* gesture suspensions compared to fluent utterances.

4.2.2 Gesture suspension in light of the MIR hypothesis and the DIP hypothesis

In the following we will lay out how gesture can provide evidence for either one of the hypotheses. Not only can the relationship between gesture and speech suspension provide insight into the coordination between gesture and speech, but it can also potentially be used to further investigate the timing of cognitive events underlying speech suspension more generally, such as the issue of whether speech suspension occurs immediately upon error detection or is delayed for the planning of a repair. The two hypotheses tested in the previous chapter, the MIR hypothesis and DIP hypothesis

make assumptions about speech suspension but not about the behavior of gesture during speech disfluency. Thus, the gesture behavior itself cannot be used to directly test these hypotheses. However, the timing of gesture suspension can be taken as an index of underlying processing, and therefore could potentially provide indirect evidence that distinguishes the hypotheses.

To this end, analyses were conducted that examined the synchrony of speech and gesture suspension for different types of speech disfluency and repair, focusing on overt repairs with within-word suspensions, and overt repairs with after-word suspensions. Overt repairs provide the critical data that could differentiate the two hypotheses, because only in these cases it is clear that the speech disruption was due to the necessity of a repair (see p. 26).

In a limited set of cases, the timing of gesture suspension may provide information about the moment of error detection. In the case where gesture stops in synchrony with, or later than, the stopping of speech, the timing of the gesture would provide no information regarding the moment of error detection. However, if gesture suspension takes place prior to speech suspension, it seems reasonable to assume that the gesture suspension was closer in time to the moment of error detection than was speech suspension. Thus, the logic is to see whether gesture stops earlier than speech and to see whether the gesture suspension latency is consistent with the hypotheses' assumptions regarding the moment of error detection and speech suspension.

The two hypotheses differ with regard to the question how much time can pass between error detection and speech suspension (see Chapter 3). For the MIR hypothesis, this amount of time is limited to either the suspension latency (within-word suspension) or to the completion of the word under articulation (after-word suspension). In contrast, the DIP is not subject to such restrictions, since interruption is initiated upon repair readiness. Hence, the interval between error detection and speech suspension is determined by the time it takes to complete the planning of the repair, which in turn depends on factors such as repair complexity. These differences between the hypotheses limit what they can account for in terms of possible speech-gesture asynchronies.

For cases of within-word suspensions, the MIR hypothesis assumes that speech was interrupted as soon as possible after the moment of error detection. The

MIR hypothesis predicts that gesture suspension should occur no sooner than the speech suspension, because an earlier suspension of gesture would suggest that speech interruption was not initiated immediately. For after-word suspensions, the lag between error detection and speech suspension could be bigger because at least some of the after-word suspensions would be the result of delayed interruption for word completion. Hence gesture suspension can occur prior to speech in after-word suspensions. However, the asynchrony between gesture and speech suspension should be no larger than the average amount of time it takes to articulate a word; otherwise, this would indicate that speech interruption was delayed for reasons other than word completion. Considering that average word length is about 400 ms (Levelt, 1989, p. 199), the delay due to word completion should not exceed 250 ms assuming 150 ms error detection latency (Levelt, 1989; see also Chapter 2 and 3).

In contrast, the DIP hypothesis assumes that there can be a bigger lag between error detection and speech suspension. When speakers have detected an error, and start to plan how to resume, they can suspend their gesture. In the meantime they go on speaking until the repair processing is in the final stages or until they have run out of words that can be uttered without further conceptual processing. Therefore, the DIP hypothesis is consistent with gesture stopping earlier than speech for both within-word suspensions as well as after-word suspensions.

4.2.3 Method

4.2.3.1 Data

The gesture analysis was based on the corpus of German living space descriptions by 12-native German speakers described in Chapter 2.

4.2.3.2 The task

Speakers described living spaces to an interlocutor (as discussed above in Chapter 3, section 3.2.1.2). Living space descriptions in an interactional setting were chosen because prior research has shown that speakers gesture more frequently in face-to-face settings than in settings where the speaker has no visually accessible addressee (Aboudan & Beattie, 1996; Bavelas et al., 2002; Bavelas, Chovil, Lawrie, & Wade, 1992). A second reason was that a high gesture rate could be expected, since the task is spatial in nature, and previous research has shown that gestures are especially prevalent when speakers talk about spatial content (Rauscher et al., 1996).

4.2.3.3 Recording

In nine of the recording sessions participants were seated on chairs without armrests at an angle of 90 degrees (see Figure 4.7, a). In three of the recording sessions the participants sat on a sofa next to each other (see Figures 4.7, b and c).

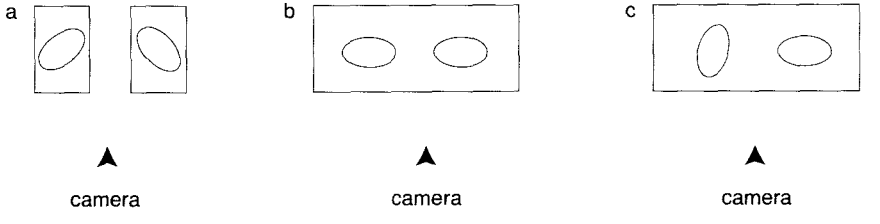

Figure 4.7. Sitting arrangement and camera position during the recording. a) sitting arrangement on chairs, b) and c) sitting arrangement on a sofa.

4.2.3.4 Transcription and coding of speech

4.2.3.4.1 Transcription

Speech was transcribed verbatim from the digitized video. Indications of speech suspensions such as glottal stops, laryngalization, truncated words as well as filled pauses and silent pauses were transcribed in detail.

4.2.3.4.2 Speech coding

For all disfluencies the moment of speech suspension and the moment of speech resumption were measured on the basis of the video and the audio file (as described in detail in the Chapter 3, see section 3.2.1.4.3, p. 59)

4.2.3.5 Segmentation and coding of gesture

For the analysis the gestural movements of each hand were first segmented into discrete gesture phases. These were then classified as preparation, stroke, hold, retraction, and partial retraction (as introduced in section 4.1.1, p. 82). Thereafter, each transition from one gesture phase to the next was coded whether it constituted a gesture suspension or a gesture continuation. Gesture phases were segmented with the annotation tool MediaTagger (see p. 58). The digitized video had a temporal resolution of 40 ms. The segmentation scheme and the gesture coding scheme were based on Kita, van Gijn, and van der Hulst (1998) but adapted for the specific purpose of the study. Criteria were developed for: a) the segmentation of the gestural movement; b) for the identification of the gesture phases; and c) for the identification of gesture suspensions. The criteria are laid out in detail in the following sections.

4.2.3.5.1 Segmentation of gesture

The gestural movements of both hands were segmented into discrete movement phases. Following Kita et al. (1998) a gestural movement was considered to have started with the initiation of a hand movement leaving rest position, and to have ended when the hands returned to rest position. The rest position could be the lap, a table or the armrests of a chair. Self-adapting movements (Kita et al., 1998; McNeill, 1992) like adjusting clothes or rubbing the nose as well as the object manipulations like grasping a coffee cup or a lighter were also considered as rest positions and not as gesture. The movement portion was then segmented into gesture phases. For each gesture phase beginning and

ending time codes were tagged by a frame-by-frame examination of the movement. For the segmentation procedure (i.e., identification of onsets and offsets of gesture phases) an actual frame-by-frame marking procedure was developed. The goal was to establish unambiguous coding criteria for obtaining consistent and frame-accurate timing of gesture phases.

Transition from a dynamic to a static phase: A dynamic phase ended when the hand came to a hold. The image quality of the video gave some indication as to when this was the case. Often the image of the hand was blurred when the hand moved, and it became clear again when the hand came to a halt. The first frame in which the hand was not blurred anymore but clear was considered to be the last frame of a dynamic phase. The next frame was considered the first frame of the static phase (see Figure 4.8 below).

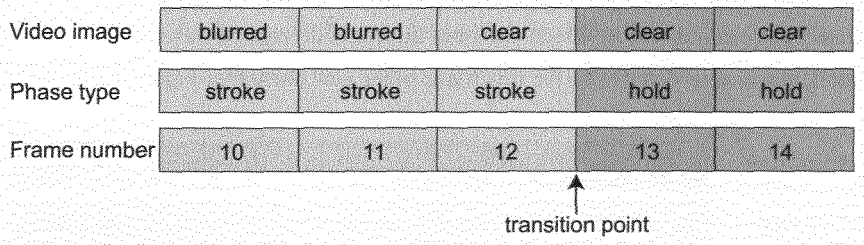

Figure 4.8. Coding of frame transition from a dynamic gesture phase to a static gesture phase. Each box represents a video-frame (40 ms each frame). In this instance frame number 12 is the transition point from stroke to hold.

Transition from a static to a dynamic phase: The first frame where a movement could be detected other than slight drifting (which often occurred in holds) was considered the first frame of the dynamic movement phase. Often a very slight movement could be detected early on, but the actual acceleration started later. In such a case the video image provided the cue for coding: as soon as the image became blurred, the new dynamic phase was considered to have started. The blurred frame was coded as the first frame of the new phase (see Figure 4.9 below).

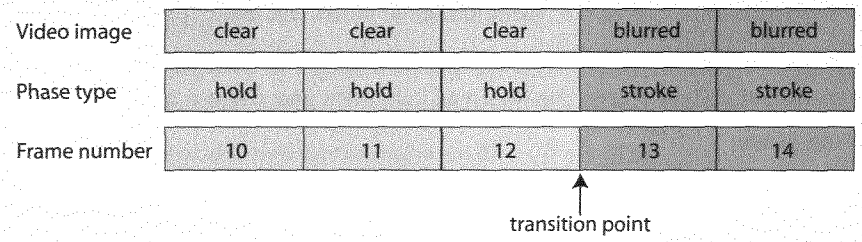

Figure 4.9. Coding of frame transition from a static gesture phase to a dynamic gesture phase. Each box represents a video-frame (40 ms each frame). In this instance frame number 12 is the transition point from hold to stroke.

Transition from a dynamic to a dynamic phase: When the direction or velocity of a movement changed abruptly, the hand slowed down considerably or possibly halted its movement for a short time. This was considered the last frame of the ongoing dynamic phase. The next frame in which the new direction or change in velocity could be observed was considered to be the first frame of the new phase (see Figure 4.10).

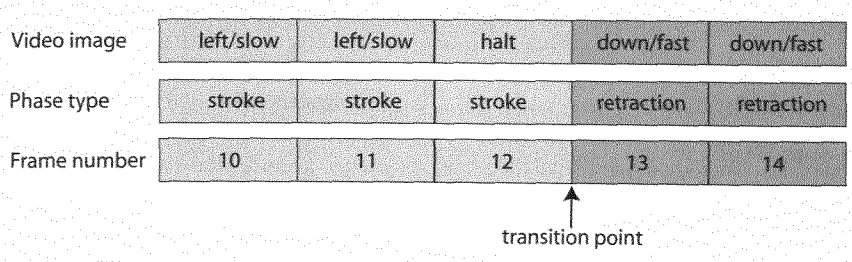

Figure 4.10. Coding of frame transition from a dynamic gesture phase to a dynamic gesture phase. Each box represents a video-frame (40 ms each frame). In this instance frame number 12 is the transition point from stroke to retraction.

4.2.3.5.2 Coding of gesture phases

In the gesture phase coding, the following general rule was applied: The stroke is the only phase type in which multiple direction changes (e.g., tracing a zigzagged line), or continuous direction changes (e.g., tracing a circle) can occur. All other dynamic gesture phases (preparation, partial retraction, retraction) cannot involve a direction

change. In other words, if a direction change occurred in the movement, it was either segmented into two phases or considered to be a stroke. Having segmented the gestural movement stream into gesture phases, the phases were labeled according to the following criteria.

Preparation: a *preparation* was defined as a movement of the hands to a location from where a stroke was deployed, for example, the hand moved from rest position (lap) up to chest height. During the rising movement the hand-internal preparation took place. The hand shape was configured, for example, all fingers were bent, the hand formed a fist, and the palm pointed downwards. The hands moved in the most direct way to the location from where the stroke was deployed. Preparations also took place between two strokes. It was possible to have two preparation phases in succession, for example, the hand rose and came to a hold at chest height. It started rising again up to forehead height. Then the stroke was deployed –the hand moved fast to the right indicating a direction.

Hold: a phase was labeled as a *hold* when the hand/s were held in a static position other than the rest position. Often the hands were not completely still, but drifted slightly. The drifting, however, was not a result of a directed movement, detectable by taking into account the neighboring dynamic phases that involved velocity changes.

When the velocity of a movement changed abruptly and a discrete direction change occurred, the hand possibly halted its movement for a single video-frame. A possible halting of movement for such a brief moment was not coded as a *hold* because it was considered a by-product of the change in velocity or direction and not an actively configured gestural state.

Stroke: a phase was labeled as a *stroke* when it appeared to display the meaning of the gesture (Kendon 1972, 2000; McNeill 1985, 1992). In most cases the stroke showed well-defined hand configuration and well articulated movement.

Gestural movements that were repeated several times (e.g., gestures depicting hammering, sawing, an object rolling down a hill or a gesture which repeatedly traced the outline of a room) posed the problem of whether to segment the movement as a single stroke or as a series of multiple strokes. For example, the movements involved in a depiction of hammering are part of a single expressive unit; the repetition of the movement is a feature of the iconic depiction of the action of hammering. As long as the movements were symmetrical and uniform in trajectory, velocity, and hand configuration they were coded as a single stroke. Similarly, the movements involved in repeatedly tracing the outline of a room were coded as a single stroke, although these seem to be rather a series of repetitions of a single expressive unit. The moment the movement became non-uniform, a new segment was considered to have started. When repeated movements were not uniform in trajectory and velocity, they were coded as sequences of preparations and strokes. Consider a sequence in which the downward pointing index finger traces a straight line, moving from left to right back and forth multiple times, depicting the location and extension of a hallway. In contrast to single strokes, which are uniform in execution, a preparation in preparation-stroke sequences could be distinguished from the stroke phase on the basis of the hand configuration, which was well articulated in strokes and more relaxed in preparations. Another indicator of preparations versus strokes was the velocity profile of the movements. The velocity of preparation and stroke differed in speed: one was deployed faster than the other. If any of the above mentioned features were observed, the movement sequence was segmented and coded as preparations and strokes.

The segmentation of gestures consisting of short up-and-down or back-and-forth movements was also somewhat problematic. This type of gesture is called a beat gesture (McNeill, 1992) or baton gesture (Efron, 1972). For coding, the sequences were broken down into preparation and stroke phases if the accented movement (stroke) and the preparatory movement were clearly distinguishable by one frame (40 ms); that is, if each movement was at least one frame or more. However, sometimes the distance traveled was so small and the speed so fast that a clear distinction was not possible with the available temporal resolution. In this case the whole beat sequence was coded as a single stroke. The same problem arose for superimposed beats–beats that were superimposed on a representational gesture depicting semantic content like the size or

shape of a room. For example, in a gesture where the hands were held in front of the body as if holding a piece of paper, the hands moved up and down repeatedly. These were coded as single strokes if a segmentation of the up-and-down or back-and-forth movement was not possible because the movement was too rapid.

Partial retraction: a phase was defined as a *partial retraction* when the hands moved towards a potential rest position (e.g., the lap), but came to hold before the rest position was reached, thus resting in an intermediate position. In these cases the movement was determined by a relaxation of the muscles; thus, it was not a directed movement. A phase was also coded as a partial retraction when a well-defined hand shape of a preceding stroke was relaxed. In these cases a change in the quality of tension occurred, for example, the fist in a hold was relaxed, and the hand configuration was released so that the hand lacked tension and was in a neutral shape (relaxed hand shape with fingers and palm naturally curving). The releasing movement was considered the partial retraction phase.

Retraction: a movement phase was defined as a *retraction* when the hands moved back into rest position (e.g., on the lap, arm rests, arms are folded in front of chest). Self-adaptors (rubbing the neck, adjusting clothes), practical actions (grasping a coffee cup) or object manipulation (playing with a lighter) were not considered to be gestures. Therefore, movements towards performing such actions were also considered retractions.

Interrupted preparation/stroke: (this category is not present in Kita et al., 1998). A movement phase was considered interrupted when a dynamic phase was abruptly ended and the abruptness was not part of the depiction. For example, in a gesture depicting something crashing into a wall the gesture stroke did not count as an interrupted stroke. Often it was not possible to determine whether the movement was a preparation for a gesture stroke or whether the actual stroke part was being executed, because the movement was suspended prematurely. In these cases the phase was coded as an

interrupted preparation/stroke. If the interrupted phase could be unambiguously defined as a preparation or a stroke, the phase was coded as *interrupted preparation* or *interrupted stroke* respectively.

4.2.3.5.3 Coding of gesture suspensions

Gesture suspensions were defined as a subset of all possible transitions between two adjacent gesture phases. Each transition of two adjacent gesture phases was coded as a gesture suspension if it met any of the following four criteria: 1) a phase transitioned into a hold, 2) a phase transitioned into a retraction, or partial retraction; 3) a preparation was followed by another preparation; or 4) a preparation or stroke was interrupted before it had been completed. These criteria are described in more detail below.

According to the first criteria a gesture was suspended when a dynamic gesture phase was halted. In terms of phase transitions this meant: all dynamic gestural movement phases were suspended by going into a hold (see Figure 4.11 below).

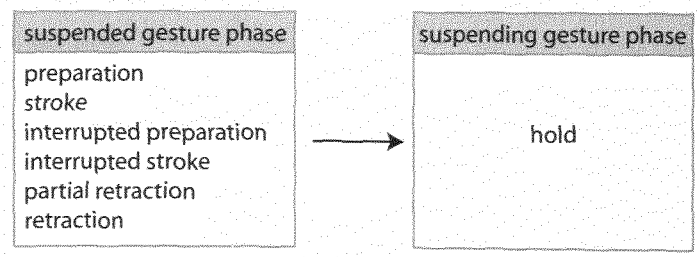

Figure 4.11. Gesture suspension by phase transitions from a dynamic gesture phase to a hold.

According to the second criteria a gesture was suspended when the hands were retracted partially (for instance by releasing the hand configuration) or when the hands were retracted back into rest position. For example, the hand rose and while it rose, the hand shape was configured (extending index finger); as the hand reached chest height but before the stroke phase was executed, the hand dropped back onto the lap. This would be a transition from a preparation to a retraction. Any gesture phase that was

followed by such a retraction or partial retraction was coded as a suspension (see Figure 4.12 below).

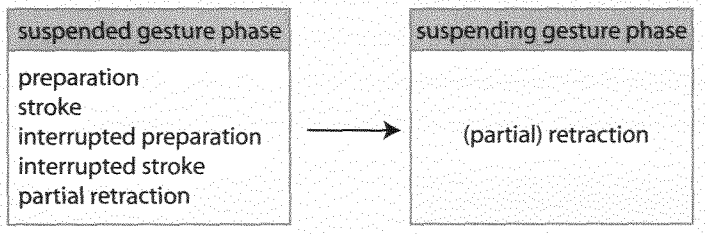

Figure 4.12. Gesture suspension by phase transition from a dynamic or static gesture phase to a partial retraction or retraction.

According to the third criterion for identifying suspensions, a gesture was suspended when the stroke for which the hands had prepared was abandoned and the hands prepared for a new gesture. For example, a pointing gesture was prepared first but was then suspended by a different gesture depicting the slope of a roof. The hand rose up from the lap, preparing for a pointing gesture (e.g., with the index finger pointing upwards). The hand stopped suddenly at chest height, and then immediately all fingers were extended, with the palm facing down, and the hand moved in a diagonal line downwards, for example, depicting the slope of a roof.

Sometimes the hands also prepared for the same stroke twice in succession by breaking up the preparatory phase into two preparations. For instance, the rising hand prepared for a pointing gesture but briefly halted midway, and then moved on up to shoulder height from where the stroke was executed. All phase transitions from preparation to preparation were coded as gesture suspensions (see Figure 4.13 below).

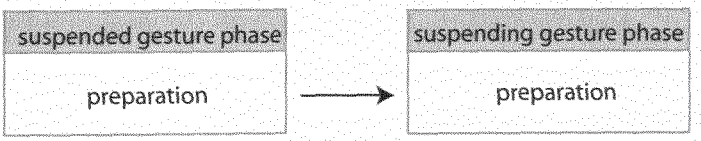

Figure 4.13. Gesture suspension by phase transition from preparation to preparation.

According to the fourth criterion, a gesture was suspended when a dynamic phase was interrupted. These were cases where a preparation or a stroke phase was prematurely truncated by a sudden abrupt halt or a sudden change in movement direction, which was not a feature of the referent being depicted. In these cases the phase transition from the interrupted phase to the next phase was always considered to be a gesture suspension, no matter what phase followed (see Figure 4.14 below).

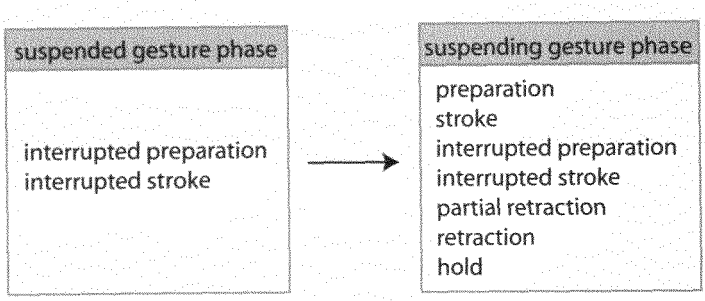

Figure 4.14. Gesture suspension by phase transition from interrupted preparation/stroke to any other phase type.

4.2.3.6 Analysis

In order to assess whether gesture is sensitive to speech disfluency, the frequency, the timing and the types of gesture suspensions were analyzed. Since gesture suspensions are a common phenomenon in fluent gesture-speech utterances (see section 4.1.5, p. 94) it was necessary to assess how many gestural suspensions during disfluent periods are related to the presence of the disfluency and how many occur independently. Therefore, all analyses were also conducted with fluent baseline utterances. These comparisons used *virtual speech suspension points* that were matched as closely as possibly to the timing of the real suspension points in disfluent utterances (see section 4.2.3.6.2, p. 116).

For the analysis of frequency differences in gesture suspension (*gesture rate analysis*) we assessed the rate of gesture suspensions per speech suspension in disfluent utterances and per virtual speech suspension in fluent utterances. However, speakers might simply be gesturing more in disfluent utterances as compared to fluent utterances. In this case the above measure might be misleading because the absolute increase in

gesture suspensions could not be directly related to the presence of speech disfluencies but could be due to an increased opportunity for gesture suspensions. To control for the possibility of frequency differences of overall gestural activity, the number of gesture phases per actual or virtual speech suspension (*phase rate*) was calculated as well as the rate of gesture suspensions per gesture phase (*suspension rate*) for both disfluent and fluent baseline utterances.

Similarly, the gesture phases might qualitatively differ in the disfluent and the fluent utterances such that speakers produced more stroke phases in one or the other type of utterance. Again this would have an impact on the number of suspensions, in that more opportunities for pre-stroke and post-stroke holds would occur. Note that the phase rate only assessed number of phases and was blind to qualitative differences in gesturing. We therefore calculated the rate of meaningful phases; specifically, the number of strokes per actual or virtual speech suspension (*stroke rate*). The stroke rate assessed whether speakers produced more stroke phases in disfluent or fluent utterances.

Secondly, we conducted a *gesture timing analysis*, which assessed the relative time differences between gesture suspensions and per actual or virtual speech suspensions. In this analysis it was assessed whether gesture was suspended on average before, at the same time or after speech was actually or virtually suspended in disfluent and fluent utterances.

Finally, we conducted a *suspension position analysis*. As described above (see section 4.2.1, p. 99), disfluent and fluent utterances might differ in the distribution of specific gesture suspension positions. Gestures might be suspended earlier with respect to the stroke in disfluent utterances than in fluent utterances. Suspensions within preparation or stroke phases or suspensions of gesture preparations and suspensions of pre-stroke holds preempt or stop the stroke. Such suspensions might be more frequent than suspensions of strokes or post-stroke holds. Hence, the frequency of different gesture suspension positions was assessed for disfluent and fluent utterances.

4.2.3.6.1 Data selection and observation window

From the corpus we selected speech disfluencies that were accompanied by gesture. Speakers often produced multiple disfluencies in close succession. In those cases it was not possible to unambiguously determine the link between the speech disfluency and the respective gestural behavior. In order to exclude that the observed gesture suspensions might be associated with a preceding or following speech disfluency, we took into account previous findings indicating that related events in speech and gesture happen within a 1-2 second interval (De Ruiter, 1998; Levelt et al., 1985; Morrel-Samuels & Krauss, 1992). A given disfluency was included in the analysis only if its speech suspension S_i was at least 2 seconds apart from the speech resumption of the previous disfluency R_{i-1} and the speech resumption of the respective disfluency R_i was at least 2 seconds apart from the speech suspension of the following disfluency S_{i+1} (see Figure 4.15 below). For each of the selected speech suspensions, a time window of one second to each side of the suspension point was the observation window for the rate, the timing, and the type analysis of the accompanying gestures.

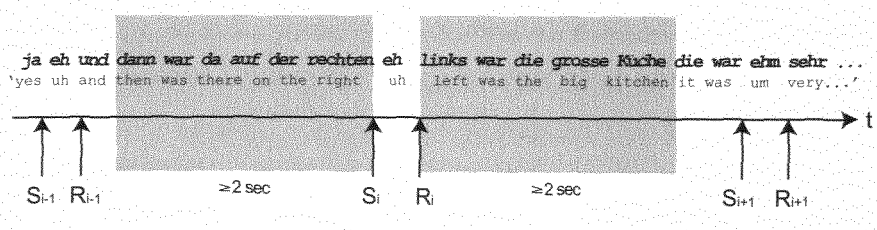

Figure 4.15. Procedure for choosing independent observation points: the speech suspension point S_i is selected if the previous resumption point R_{i-1} is at least two seconds apart and if the resumption point R_i is at least two seconds apart from the next suspension point S_{i+1}.

Gesture rate analyses

For the calculation of frequency differences in gestural activity, the instances of gesture phases, stroke phases and gesture suspensions that took place within the observation window were determined for both hands. Then the rates of gesture suspensions, phases, and strokes per speech suspension and gesture suspensions per gesture phase were calculated.

Gesture timing analysis

The second analysis concerned the timing of gestural suspensions relative to speech suspensions. Within the observation window only the gesture suspension closest to the speech suspension point was selected, regardless of whether it occurred before or after the speech suspension (see Figure 4.16 below). In case speakers gestured with both hands, only the gesture suspension of the hand that was closest to the speech suspension was included in the analysis.

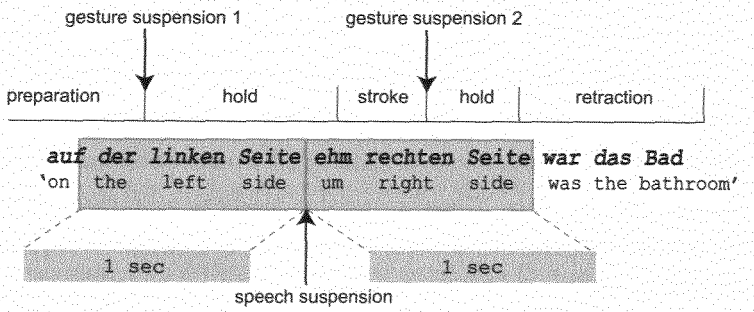

Figure 4.16. Example of a disfluent utterance and the disfluent period (as defined by 1 sec to each side of the speech suspension point) with two gesture suspensions (phase transition from preparation to hold (1); phase transition from stroke to hold (2)). The closest gesture suspension relative to the speech suspension, in this case gesture suspension 2, was included in the analysis.

The asynchrony between gesture and speech suspension was defined as the time of gesture suspension minus the time of the co-occurring speech suspension.

Gesture suspension position analysis

For the analysis of gesture suspension positions the frequencies of the following suspended gesture phases were determined (for an overview of gesture suspension coding, see section 4.2.3.5.3, p. 110): suspensions of strokes, suspensions of post-stroke holds, suspensions of partial retractions following strokes, suspensions of preparations, suspensions of pre-stroke holds, suspensions of partial retractions following preparations, and phases that were suspended within (interrupted preparations/strokes).

For the statistical analysis suspensions of strokes, post-stroke holds, and partial retractions following strokes were grouped and categorized as *late* gesture suspensions. Suspensions of preparations, interrupted preparations/strokes, pre-stroke holds, and suspensions of partial retractions following preparations were grouped and categorized as *early* gesture suspensions.

4.2.3.6.2 Baseline

A baseline for the assessment of distributional, temporal and qualitative differences of gestural activity was computed based on fluent speech (see also Levelt 1983; Kita 1993). For each speaker all fluent utterances that were accompanied by gesture were selected. Then *virtual speech suspension points* were determined. For each speech suspension selected for the analysis, the location of the speech suspension within the disfluent utterance was assessed by calculating the distance of the speech suspension from the onset of the clause. These suspension locations were then inserted into randomly selected fluent baseline utterances. For example, for a speech suspension that occurred 1600 ms after onset of a disfluent utterance, a virtual speech suspension point was inserted 1600 ms after onset of a randomly selected fluent baseline utterance by the same speaker (see Figure 4.17 below). Hence, the number and the respective location of virtual suspension points was matched to the number and the location of real speech suspension points that each speaker contributed to the analysis of the disfluent utterances. For each virtual suspension point it was checked that the observation window of one second to each side did not reach into a preceding or following utterance. This ensured that the observation window of a virtual speech suspension point did not overlap with the observation window of a speech suspension in a preceding or subsequent disfluent utterance. The resulting selection of virtual speech suspension points was subjected to the same analyses as the disfluent utterances (rates, timing of gesture suspension relative to speech suspension, and gesture suspension positions).

The temporal matching procedure for the baseline utterances controlled for possible systematic differences in the distribution of speech and gesture suspensions over the course of an utterance. For instance, gesture suspensions could have a greater rate of occurrence in the beginning of an utterance than speech suspensions, or vice

versa. Without taking into account the temporal location of these events, one might find an asynchrony that was only driven by the difference in the distributional characteristics of the location of speech and gesture suspensions.

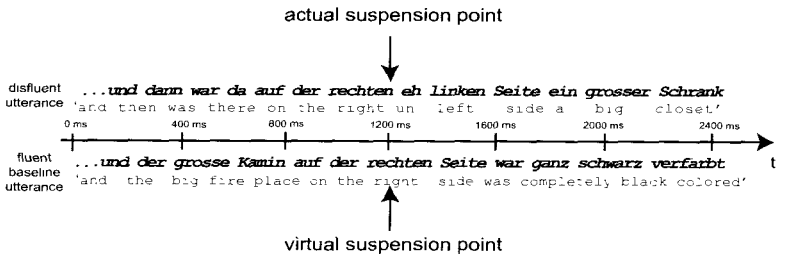

Figure 4.17. Baseline procedure. For each disfluent utterance the location of the actual speech suspension was determined (in this example at 1200 ms). A fluent utterance was selected from the same speaker who produced the disfluent utterance and a virtual speech suspension point was placed at the same location (after 1200 ms) as in the disfluent utterance. This virtual suspension point was then used as corresponding reference point for the various analyses.

4.2.3.6.3 Reliability

A reliability check was performed on a randomly selected 15% subset of the speech disfluencies that were 2 seconds apart and were accompanied by gestures within the one-second-window to each side of the respective speech suspension. A second trained rater independently segmented and coded gesture phases. The raters agreed on 74% in the coding of gesture phases. They moreover agreed on 71% of the gesture phase segmentation within 2 frames. This percentage is comparable to the 72% for gesture phases and 69% of gesture phase boundaries reported by Kita et al. (1998).

4.2.4 Results

4.2.4.1 General characteristics of gesturing time and gesture phases in the corpus

The total speaking time in the corpus was 96.3 minutes. During this time participants' hands departed from and returned to rest position 1072 times. Participants produced during that time 8785 gesture phases. Of all gesture phases 3037 were gesture strokes. Table 4.1. below provides an overview of the gesturing that occurred during the overall

speaking time for each participant (for an overview of gesture phases and strokes see Appendix, Tables 6.3 - 6.9). Note that results are presented for right and left hand independently as well as together, because some gestures included both hands, while others included only a single hand.

The proportion of speaking time spent gesturing indicates that the living space description task was successful in inducing a substantial amount of gesturing over all subjects. Of the 96.3 minutes of speaking time, 51.7 minutes, or 55%, was spent gesturing. Over participants, the time spent gesturing ranged from 2.29 to 6.40 minutes (*SD* 1.25). Expressed as a proportion of speaking time, this range was from 27% to 84% (*SD* = 20%).

Table 4.1. Time in minutes spent gesturing and speaking for each participant for both hands (time during which left and right hand overlapped in gesturing, total gesturing time and total speaking time and proportion of speaking time during which the speaker gestured).

Participant	Left hand	Right hand	Overlap	Gesturing time	Speaking time	Proportion gesture/ speech time
AN	4.82	4.40	4.14	5.08	6.08	0.84
AR	3.83	4.11	3.36	4.58	5.80	0.79
BI	3.46	5.98	3.03	6.40	7.86	0.81
FA	1.60	3.66	1.33	3.93	9.04	0.43
KA	2.78	5.65	2.62	5.81	8.21	0.71
MA	3.76	3.58	2.93	4.42	8.41	0.53
NI	0.41	2.97	0.33	3.05	8.68	0.35
NA	2.10	3.95	1.83	4.22	8.59	0.49
SE	2.54	3.38	1.07	4.84	8.41	0.58
SI	4.01	3.74	3.05	4.70	8.34	0.56
SM	1.69	2.12	1.44	2.38	8.36	0.28
TO	1.12	1.63	0.46	2.29	8.53	0.27
TOTAL	32.12	45.17	25.58	51.70	96.31	0.54

The corpus contained 1202 speech disfluencies. Ninety-three disfluencies were either due to interruption by an interlocutor or did not have an indication that the error was actually detected. These disfluencies were excluded from the analysis. Hence, 1109 speech suspensions remained. For the analysis of gesture suspensions in relation to speech suspensions only those speech suspensions were selected that were two seconds apart from the resumption of the previous disfluency; and their respective resumption was two seconds apart from the following speech suspension (see Figure 4.15 above). The procedure resulted in 432 speech suspensions. The considerable reduction of data points was due to the fact that the speech suspensions often happened in close succession.

Out of the 432 speech suspensions, 306 were accompanied by gestures. Of these 306 speech suspensions, 206 had gestures taking place within one second to each side of the speech suspension. For 178 speech suspensions a matching virtual suspension point could be located in a fluent baseline utterance, which was accompanied by gesture within the one-second window to each side of the virtual suspension point. Out of the 178 actual speech suspensions, 82 were followed by a covert repair (repetition of elements or filled pause with no evidence of change of content in the resumption) and 96 were followed by an overt repair (altered element/s in the resumption in comparison to the original delivery).

4.2.4.2 Gestural sensitivity to speech disfluencies

Gesture rate analysis

The different rate analyses were performed to see whether there are general frequency differences in the gestural behavior accompanying fluent as compared to disfluent speech. The variables for fluent and disfluent utterances were compared using a t-test for paired samples, treating participants as a random factor ($N = 12$).

As described in the Method section above, the analysis window consisted of a one second window to each side of the speech suspension point (see p. 115). In this disfluent period, speakers suspended their gestures at a mean rate of 2.10 gesture suspensions per speech suspension ($SD = .52$). This gesture suspension rate includes gesture suspensions

of either hand. Note that for every hand more than one gesture suspension can occur in the disfluent period. In the fluent baseline utterances, speakers suspended their gestures on average 1.94 times per virtual speech suspension (SD = .38). The difference in the suspension rates was not significant ($t(11) = 1.29$, n.s.)

Likewise, the gesture phase rate in the disfluent utterances was not different from the gesture phase rate in the fluent baseline utterances. The rate was on average 4.74 phases (SD = 1.07) per speech suspension. For the baseline the rate was 4.45 phases (SD = .78) per virtual speech suspension. This difference was not significant ($t(11) = 1.8$, n.s.)

The stroke rate in the disfluent utterances was also not different from the stroke rate in the baseline condition. In the disfluent utterances speakers produced 2.09 strokes per speech suspension (SD = .54), while in the fluent baseline utterances 1.94 strokes per virtual speech suspension (SD = .34) were produced. The difference between the stroke rate in fluent baseline and disfluent utterances was not significant ($t(11) = 1.25$, n.s.)

To control for the possibility that the number of gesture suspensions might differ in relation to the number of gesture phases, an analysis was conducted which considered gesture suspensions as a proportion of phase transitions in the disfluent versus fluent baseline utterances. In the disfluent utterances, gesture suspensions comprised on average .45 (SD = .07) of the overall phase transitions as compared to the fluent baseline utterances with .44 (SD = .05). The difference between disfluent and fluent baseline utterances was not significant ($t(11) = 0.43$, n.s.).

In sum, there were no significant differences in the frequency of gesture suspensions or gesture phases between disfluent and fluent baseline utterances.

Gesture timing analysis

In the following step, we examined those speech suspensions that were accompanied by gestures and that had a corresponding fluent baseline utterance with matching virtual speech suspension accompanied by gesture (N = 178). Figure 4.18 shows the distribution of gesture suspensions over time slots relative to speech suspension in fluent and disfluent utterances.

Because we applied stringent criteria for the data selection to ensure independence of neighboring disfluencies (see section 4.2.3.6.1, p. 114) as well as stringent criteria for the baseline selection (see section 4.2.3.6.2, p. 116) the data points that each subject provided for the analyses were considerably reduced. Unlike the rate analysis, in the timing and type analyses only disfluent utterances containing gesture suspensions were considered, resulting in further data reduction. This made it problematic to perform statistical tests using participants as the unit of analysis because this sometimes led to one or more participants with very few observations or even empty cells. Thus, in the gesture timing and in the gesture suspension position analysis we treated each disfluency as an independent observation.

The mean asynchrony between gesture suspensions and speech suspensions was -92 ms (2.29 frames) (SD = 404 or 10.11 frames), which was significantly different from zero, ($t(177) = 3.04$, $p < .001$). In other words, gesture suspension preceded speech suspension more often than would be expected if suspensions in the two modalities were unrelated. Preceding gesture suspensions were mainly observed in the time slot between 0 and –160 ms (see Figure 4.18 below).

Such a relationship was not observed in the fluent baseline utterances. Here the mean asynchrony between gesture and virtual speech suspensions was 16 ms (0.4 frames) (SD = 466 or 11.66 frames). This mean asynchrony was not significantly different from zero ($t(177) = 0.46$, n.s.) In other words, as suggested by the distribution in Figure 4.18, gesture suspensions in baseline utterances were equally likely to occur before and after virtual suspension points. The difference in asynchrony between the disfluent utterances and the fluent baseline utterances was statistically significant ($t(177) = 2.17$, $p < .05$).

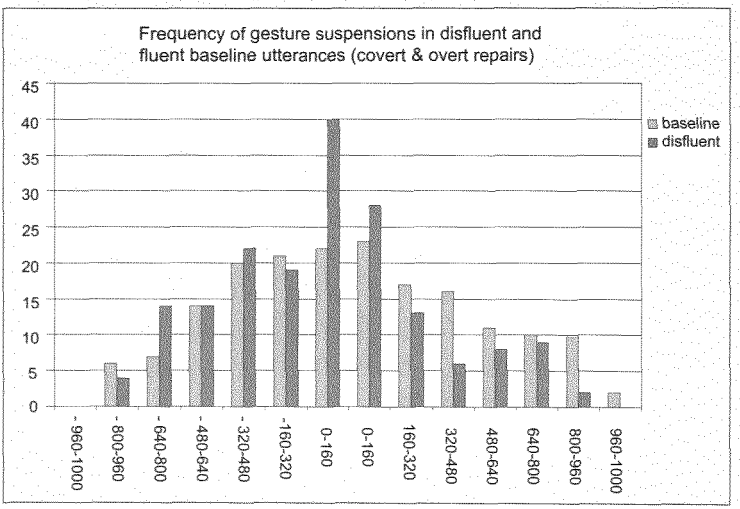

Figure 4.18. Frequency of gesture suspension around speech suspension points (0) in disfluent utterances (N = 178) and matched fluent baseline utterances (N = 178).

Gesture suspension position analysis

The frequencies of gesture suspension positions in disfluent and fluent baseline utterances are shown in Table 4.2 below.

Table 4.2. Numbers and proportions of gesture suspension positions for disfluent utterances with covert and overt repairs and corresponding fluent baseline utterances.

	Utterance type			
Gesture suspension position	Disfluent		Fluent	
stroke	105	(.59)	135	(.76)
post-stroke hold	16	(.09)	11	(.06)
partial retraction (post-stroke)	4	(.02)	5	(.03)
preparation	42	(.24)	20	(.11)
interrupted preparation/stroke	10	(.06)	6	(.03)
pre-stroke hold	1	(.01)	1	(.01)
partial retraction (pre-stroke)	0	(.00)	0	(.00)
Totals	178		178	

Grouping into early and late suspension positions (see p. 115) resulted in the frequencies listed in Table 4.3 below.

Table 4.3. Numbers and proportions of early and late gesture suspension positions for disfluent utterances with covert and overt repairs and corresponding fluent baseline utterances.

	Utterance type			
Gesture suspension position	Disfluent		Fluent	
early	53	(.30)	27	(.15)
late	125	(.70)	151	(.85)
Total	178		178	

A chi-square analysis was conducted to test for a difference in the distribution of gesture suspension positions (early vs. late) across disfluent and fluent baseline utterances. A significant association was found between gesture suspension position and utterance type ($\chi^2(1) = 10.90, p < .001$).

4.2.4.3 Evidence for MIR hypothesis or the DIP hypothesis

In the previous section, the question of gestural sensitivity to speech disfluency was addressed. A significant gestural response to speech disfluency was observed in the timing and the gesture suspension position analyses. Given that gesture was on average suspended before speech it may be possible to use gesture as a further source of evidence with respect to the MIR or DIP hypotheses. To this end, in this section we present the gesture timing and gesture suspension position analyses for overt repairs following within-word and after-word suspensions. The motivation for focusing on overt repairs, along with the specific predictions for the two hypotheses, have been outlined in section 4.2.2, p. 100.

For the following analyses, the baseline assessment was slightly altered from that used in the previous results section, in which the temporal location of the virtual speech suspension in the fluent utterance was exactly matched to the location of the actual speech suspension in the disfluent utterance (distance of speech suspension from the beginning of the utterance in ms; see p. 116). With the former procedure, the vast

majority of virtual suspension points ended up within-word. However, it is possible that not only the temporal location of the speech suspension point but also the location with respect to word boundaries—whether it is within- or after-word—has an influence on the asynchrony. In order to control for such a possibility, we used a different baseline procedure for after-word suspensions. For each after-word suspension in a disfluent utterance the location of the virtual speech suspension in the matching fluent baseline utterance was adjusted such that it was located at the closest word end. This procedure resulted in approximately equal numbers of relocations in each direction (forward 24, backward 21).

Gesture timing analysis

When only overt repairs were considered (68 after-word suspensions and 28 within-word suspensions) it was found that gesture suspensions still on average preceded speech suspensions. The mean asynchrony between gesture and speech suspensions was -82 ms (2.05 frames, SD = 404 or 10.11 frames). The asynchrony differed significantly from zero ($t(95) = 1.99$, $p < .05$). In the recalculated baseline, the mean asynchrony between gesture suspensions and virtual speech suspensions was 42 ms (1.04 frames, SD = 408 or 10.21 frames). The asynchrony between gesture and speech suspensions in fluent baseline utterances was not significantly different from zero ($t(95) = 0.32$, n.s.) The difference in asynchrony between the disfluent utterances and the recalculated fluent baseline utterances was statistically significant ($t(95) = 2.04$, $p < .04$).

For after-word suspensions alone, the mean gesture-speech asynchrony was -98 ms; that is, gesture stopped on average 98 ms (2.46 frames, SD = 400 or 10.04 frames) before speech stopped. This was significantly different from zero ($t(67) = 2.04$, $p < .05$). In the fluent baseline utterances gesture stopped on average 51 ms after speech stopped (mean asynchrony 51 ms or 1.28 frames, SD = 426 or 10.65 frames); the mean asynchrony between gesture and speech suspensions did not differ significantly from zero ($t(67) = 0.99$, n.s.) The difference in asynchrony between the disfluent and the fluent baseline utterances was statistically significant ($t(67) = 2.05$, $p < .04$).

For within-word suspensions alone, the gesture-speech asynchrony was -43 ms (1.07 frames, *SD* = 415 or 10.38 frames); that is, gesture was suspended on average 43 ms before speech. The asynchrony did not differ significantly from zero ($t(27) = 0.26$, n.s.) In the fluent baseline utterances the mean asynchrony between gesture and speech suspensions was 19 ms (0.46 frames, *SD* = 369 or 9.22 frames). This mean asynchrony was not significantly different from zero ($t(27) = 0.54$, n.s.) The difference in asynchrony between the disfluent and the fluent baseline utterances was statistically not significant ($t(27) = 0.55$, n.s.)

In other words, the separate analyses of within- and after-word suspensions suggest that the pre-positioning of gesture suspensions with respect to speech suspension is mainly driven by after-word suspensions.

Gesture suspension position analysis

The frequencies of gesture suspension positions in overt repairs following within-word and after-word suspensions are shown in Table 4.4 below.

Table 4.4. Numbers and proportions of gesture suspension positions in disfluent utterances with overt repairs and corresponding fluent baseline utterances.

Gesture suspension position	Speech suspension type											
	Within-word (*N* = 28)				After-word (*N* = 68)				All overt repairs (*N* = 96)			
	Disfluent		Fluent		Disfluent		Fluent		Disfluent		Fluent	
stroke	16	(.57)	24	(.86)	46	(.68)	56	(.82)	62	(.65)	80	(.83)
post-stroke hold	1	(.04)	2	(.07	2	(.03)	4	(.06)	3	(.04)	6	(.06)
partial retraction (post-stroke)	0	(.00)	1	(.04)	3	(.04)	0	(.00)	3	(.04)	1	(.01)
preparation	7	(.25)	1	(.04)	13	(.19)	6	(.09)	20	(.21)	7	(.07)
interrupted preparation/stroke	4	(.14)	0	(.00)	3	(.04)	2	(.03)	7	(.07)	2	(.01)
pre-stroke hold	0	(.00)	0	(.00)	1	(.01)	0	(.00)	1	(.01)	0	(.00)
partial retraction (pre-stroke)	0	(.00)	0	(.00)	0	(.00)	0	(.00)	0	(.00)	0	(.00)

Grouping into early and late suspensions (see p. 115) resulted in the frequencies listed in Table 4.5 below.

Table 4.5. Numbers and proportions of early and late gesture suspension positions in disfluent utterances with overt repairs and corresponding fluent baseline utterances.

Gesture suspension position	Speech suspension type											
	Within-word (N = 28)				After-word (N = 68)				All overt repairs (N = 96)			
	Disfluent		Fluent		Disfluent		Fluent		Disfluent		Fluent	
early	11	(.39)	1	(.04)	17	(.25)	8	(.12)	28	(.29)	9	(.09)
late	17	(.61)	27	(.96)	51	(.75)	60	(.88)	68	(.71)	87	(.91)

We compared the distribution of gesture suspension positions in the disfluent versus fluent baseline utterances. Separate chi-square tests were performed for all overt repairs and for the subgroups of within-word suspensions and after-word suspensions followed by overt repairs. For all overt repairs, a significant association between gesture suspension position (early vs. late) and utterance type (disfluent vs. fluent) was found ($\chi^2(1) = 12.08$, $p < .01$); that is, the distribution of early versus late gesture suspensions differed significantly between disfluent utterances and fluent baseline utterances. Specifically, an early gesture suspension was 3.1 times more likely in disfluent utterances than in fluent baseline utterances. Significant associations between gesture suspension position and utterance type (fluent vs. disfluent) were also found in separate analyses for within-word suspensions ($\chi^2(1) = 8.40$, $p < .01$) as well as for after-word suspensions ($\chi^2(1) = 3.96$, $p < .05$).

4.2.5 Discussion

Corpus Study 2 addressed the questions of whether gesture is sensitive to speech disfluency and if so whether the timing of gesture suspensions provides evidence for or against the MIR or the DIP hypotheses. Given the naturalistic character of the data, stringent criteria were adopted to maximize the likelihood that a specific gestural

response was associated with a particular speech disfluency. This procedure resulted in the exclusion of a considerable number of observations. However, due to the collection of a very large set of speech disfluencies the remaining number of observations was sufficient to conduct statistical tests in the various analysis steps. Out of the total of 1109 speech disfluencies that were observed, 178 cases (16%) were included in the analyses.

The first question was whether gesture is sensitive to speech disfluencies. We conducted a gesture rate analysis, a gesture timing analysis, and a gesture suspension position analysis.

The amount and type of gesturing did not differ between disfluent and fluent utterances. The rate analysis did not reveal differences in the frequencies of gesture suspensions, gesture phases, gesture strokes, or gesture suspensions as a proportion of overall gesture phases. In contrast, the timing analysis revealed differences between disfluent and fluent baseline utterances in the timing of gesture suspensions. In fluent baseline utterances no average asynchrony could be detected, which was expected under the assumption of no relationship between gesture suspensions and virtual suspension points. By contrast, in some types of disfluent utterances an asynchrony between gesture suspensions and speech suspensions was detected providing evidence that there is a systematic relationship between gesture suspensions and speech suspensions. Gesture was suspended on average before speech in an analysis of all repairs in disfluent utterances. Similarly, the gesture suspension position analysis revealed a difference between disfluent utterances and fluent baseline utterances. In disfluent utterances we found a relative increase in early gesture suspensions (i.e., suspensions that occurred before the end of the stroke , which was not observed in fluent baseline utterances.

In sum, neither the frequency of gesture suspensions, nor the relative amount of gestural activity turned out to be reliable indicators of gestural sensitivity to speech disfluency. Differences between fluent baseline and disfluent utterances were observed in the timing of gesture suspension in relation to speech suspension and in the location of suspensions relative to the stroke indicating that gesture is in fact sensitive to speech disfluencies.

How is it possible that there are changes in the timing and location of gesture suspension while the frequencies of gesture suspensions, phases and strokes remain constant? The results can be explained by assuming that gesture reacts to speech through some kind of structural reorganization. At least in disfluent utterances with overt repairs our results indicate that gesture was structurally adjusted such that suspending phases like holds occurred on average earlier in the overall sequence of gesture phases than in fluent utterances. Another way that gesture possibly adjusted to disfluent speech might be a temporal reorganisation as reported by De Ruiter (1998, see section 4.1.4, p. 91). Individual gesture phases might have become prolonged while others were shortened. For example, a gesture stroke might have been prolonged by slowing down execution or by repeating the movement multiple times. Since the overall frequency of gesture phases in the time interval used in the present study (2 sec) remained constant, phases following a prolonged stroke would have to be shortened. Further study is needed in order to investigate and better understand the nature of such temporal and structural re-organization. The present study can serve as a starting point for such studies.

The second question was whether the timing of gesture suspensions provides evidence for the MIR or the DIP hypothesis. As mentioned above (see predictions, p. 100), the relevant data for testing the hypotheses are the gesture-speech asynchronies that occur in overt repairs following within-word and after-word suspensions. Overall, it was found that in overt repairs there was a significant asynchrony of gesture suspension and speech suspension, with gesture being suspended prior to speech. Separate analyses of after-word suspensions and within-word suspensions suggested that this asynchrony was mainly driven by gesture suspensions accompanying after-word speech suspensions. No significant asynchrony was found for cases of within-word speech suspensions. The latter finding may mean that there is no systematic relationship between gesture and speech suspensions for this subgroup of disfluent utterances. Alternatively, the finding might mean that gesture and speech were suspended synchronously and therefore no timing difference was found. Our way of testing does not distinguish between the two possibilities. However, the gesture suspension position analysis supports the latter of these possibilities. The distribution of gesture suspension positions showed a significant difference between disfluent and baseline utterances both

for within- and after-word suspensions. Specifically, gesture was suspended more often in early phases of execution in disfluent utterances than in fluent utterances. Thus, despite the absence of a temporal asynchrony, there seems to be a relationship between gesture and speech suspensions in within-word suspensions.

What do these findings mean for the MIR hypothesis and the DIP hypothesis? The finding of no significant gesture-speech asynchrony in cases of overt repairs following within-word suspensions is in line with the MIR hypothesis as well as with the DIP hypothesis. This is because both may assume that gesture and speech are interrupted simultaneously.

According to the MIR hypothesis the interruption would be triggered by a stop signal upon error detection. However, not all gestural suspensions as operationalized in the study could be explained as the result of a simple stop signal. Gesture suspensions took different forms and the planning time required for some of these suspensions, for example, gestural fresh starts (preparation followed by a preparation for a different gesture), might exceed an estimated interruption latency of 150 ms (Hartsuiker & Kolk, 2001) during which parallel planning of a new gesture might take place. This is analogous to the situation of fresh starts in speech following zero ms cut-off-to-repair intervals. The replanning process of the fresh start is too complex and requires too much time to fit within the time required for the interruption process to be completed.

The DIP hypothesis can account for the absence of a significant gesture-speech asynchrony in cases of overt repairs following within-word suspensions by assuming that both gesture and speech are interrupted upon repair readiness. Moreover, the DIP does not run into similar problems as the MIR hypothesis in explaining the different ways in which gesture was suspended. This is because it is assumed that interruption is delayed for repair planning, and thus, the speaker also has time to plan how to suspend the gesture (e.g., by retracting or freezing the gesture). The delay could give the speaker even the time to completely replan the gesture.

The finding that gesture stopped before speech in cases of after-word suspension, can also be explained by both hypotheses. According to the MIR hypothesis, after-word suspensions are cases in which an error is detected but the word under articulation is completed in order to signal that that word is correct. In such a situation, the gesture

might be interrupted immediately, while the interruption of speech would be delayed for word completion. The critical question with respect to the MIR hypothesis is whether the observed asynchrony is too long to be accounted for by completion of the word under articulation. More specifically, the question is whether the prepositioning of gesture suspension exceeds some 250 ms that can be assumed for completion of a word under articulation. Note that the average asynchrony of 82 ms might underestimate the actual temporal prepositioning of gesture suspensions. This is because the average included gesture suspensions that were not related to the disfluency. A considerable number of gesture suspensions occurred in both fluent and disfluent utterances and therefore cannot be attributed to the disfluency. This can be seen in the overlapping portions of the fluent and disfluent distributions of gesture suspensions in relation to the speech suspension (see Figure 4.18, p. 122). Nevertheless, the number of gesture suspensions in disfluent utterances exceeded the number of gesture suspensions in fluent utterances mainly in the time slot of gesture suspensions occurring 0-160 ms before speech suspension. Hence, it can be concluded that the temporal prepositioning of the majority of gesture suspensions related to the disfluency does not exceed the time necessary to complete a word under articulation and is compatible with the MIR hypothesis.

Note, however, that the MIR hypothesis can only explain the gesture results for after-word suspensions and within-word suspensions assuming different gesture-speech relationships for the two types of speech suspensions. In the case of within-word suspensions, gesture would behave just like speech in that both are triggered by the same event—error detection. By contrast in the case of after-word suspensions the MIR hypothesis would have to assume that gesture is treated differently from speech, in that gesture interruption but not speech interruption would be triggered by error detection. The MIR hypothesis would have to explain for the after-word case why gesture would be interrupted immediately while the word under articulation would be completed.

The DIP hypothesis assumes that after-word suspensions are the result of running out of prepared material in the formulator and the articulatory buffer. Thus, similar to the MIR hypothesis, the DIP hypothesis might account for the gesture result by assuming that gesture is interrupted upon error detection, while speech is continued.

However, it seems inconsistent for the DIP hypothesis to deny any role of error detection in the triggering of speech interruption but to assume that error detection might be the triggering event of gesture interruptions accompanying after-word suspensions. The assumption of two different mechanisms is furthermore problematic because it would entail that upon error detection, speakers would need an estimate whether they will eventually suspend speech within-word or after-word in order to co-ordinate gesture suspension accordingly. To have such an estimate, the speakers would have to know for how long they will be able to go on speaking, how long it will take to process the repair, and how long the interruption itself will take. They would then have to adjust the timing of gesture suspensions to these estimates by chosing either error detection or repair readiness as trigger event. The processing entailed by such an account seems overly complex. Obviously this is not a viable mechanism.

In order to remain consistent the DIP has to assume the *same* underlying mechanism for speech and gesture suspension. In cases of within-word suspensions, the DIP hypothesis may assume that gesture and speech are interrupted simultaneously upon repair readiness. In cases of after-word suspensions, it may assume that the speaker has to cease speaking and gesturing because he runs out of prepared material in the speech buffer as well as in the gesture buffer. The pre-positioning of gesture suspension relative to speech suspension in after-word cases would then not be the result of interruption upon error detection. Instead, it would be the result of the gesture buffer running out of material before the speech buffer, because it is smaller.

This proposed explanation assumes that the suspension of gesture is not due to a trigger signal but is a consequence of running out of prepared material. However, the gesture suspension position analysis appears to contradict this assumption. This analysis found that in after-word just as in within-word suspensions, gesture was more likely to be suspended early—prior to stroke completion—in disfluent utterances compared to fluent baseline utterances. If after-word suspensions were the result of running out of buffered material, then there should be no preponderance of early gesture suspensions, since in the absence of a triggering event for gesture suspension gesture phases should run to completion. Note, however, that the class of after-word suspensions consists not only of cases in which the speaker ran out of buffered material and had to cease

speaking, but also of cases in which speech was interrupted upon repair readiness and the interruption happened to result in an after-word suspension. The DIP hypothesis might assume that the cases in which gesture was suspended early (before the completion of the stroke) should be instances in which it was suspended upon repair readiness. These cases would then not contribute to the observed temporal gesture-speech asynchrony. The cases in which gesture was suspended late (completion of the stroke or thereafter) should be cases in which the speaker went on talking and gesturing but then ran out of buffered material with the gestural buffer running out before the speech buffer. It would be these cases that drive the observed gesture-speech asynchrony. It is evident that these considerations are highly speculative and need further research in order to be confirmed.

In sum, the results of the gesture study with respect to the MIR and the DIP hypothesis do not provide conclusive evidence for either hypothesis. Both hypotheses encounter problems in providing consistent explanations for the observed pattern of findings.

4.3 Control experiment: Stopping latencies of speech and gesture

4.3.1 Introduction

Much of the above discussion is based on the assumption that when gesture stops earlier than speech, it is closer to the moment of error detection than speech suspension. In other words, it was assumed that a stop signal is sent first to gesture and subsequently a stop signal was sent to speech. However, this possibility hinges on the assumption of similar suspension latencies (time from the internal stop signal to suspension) of the two modalities. If the assumption of similar suspension latencies of the modalities did not hold, other scenarios would be possible. For example, if the suspension latency for gesture was generally shorter than the suspension latency for speech, in within-word suspensions the stop signal for speech interruption might have been released first. Hence, speech suspension would be closer in time to the trigger event than gesture suspension. For after-word suspension, the stop signal for gesture and speech interruption might have been released at the same time.

Conversely, if gesture suspension latencies were generally longer than speech suspension latencies, in both within-word and after-word suspensions, the stop signal for gesture interruption might have been released before the stop signal for speech interruption. Hence, the assessment of the suspension latencies is of crucial importance, since the underlying process of gesture-speech coordination and interruption in case of speech disfluencies has to be described differently depending on the suspension latencies. Therefore, a control experiment was conducted that assessed whether the stopping latencies of gesture and speech in sustained discourse differ.

Some evidence for stopping latencies of speech and hand movements has been obtained using the so-called stop signal paradigm. The literature on the stop-signal paradigm suggests that stopping latencies do not differ much across tasks. Similar latencies in the order of about 200-400 ms have been found across tasks and effector systems (Logan & Cowan, 1984). This includes hand movements like type-writing (Logan, 1982), key-presses (Logan, 1981), arm movements (McGarry & Franks, 1997)

as well as the articulatory movements in speaking (Ladefoged, Silverstein, & Papcun, 1973). However, the applicability of these findings to the present study is limited. The studies involving hand/arm movement suspension usually investigated simple movements like key presses or the squeezing of an object. In more complex movements like type writing, motor programs for key sequences (typing high frequent syllables/words) have to be inhibited. Such practiced movements differ from gesture in that the gestural movement is mostly created on the spot. Furthermore, gesture differs from button pushes and object squeezes in that the gestural movements are more complex. For example, they can involve the movement of both arms in front of the body outlining the layout of a space.

Furthermore, in contrast to the conditions in the corpus study, usually only one modality is investigated independently, for example, participants have to inhibit a button push or interrupt a single sentence. The processes underlying gesture and speech suspension in sustained discourse might differ, since the modalities are coordinated in specific ways and are produced simultaneously. Since the stopping latencies for speech and co-occurring gesture in discourse have never been evaluated, the question regarding the relative stopping latencies for the two modalities cannot be answered reliably on the basis of existing data.

One way of assessing whether the suspension latencies differ for the modalities is to provide an external stop signal for both modalities, which triggers two simultaneous internal stop signals for gesture and speech.

In Corpus Study 2, gesture and speech were suspended at varying points during their execution. Gesture suspensions occurred within a phase, after a preparation or a hold. Moreover, different kinds of gestures, such as pointing gestures or gestures depicting the shape of an object, were suspended. Also, speech was suspended at various points, within a word as well as after a word. In order to obtain data and results that are comparable to the corpus study, the conditions in the following experiment were kept as similar as possible to those in the previous study. Participants were involved in the same conversational task; namely in providing living space descriptions to an interlocutor. While participants were performing the task, external auditory stop

signals were presented at random intervals. Participants were instructed to interrupt gesture and speech as soon as possible after hearing this stop signal. The random presentation of the stop signal during sustained discourse allowed us to tap into the gesture/speech performance at different points of their execution.

4.3.2 Method

4.3.2.1 Participants

20 Dutch undergraduate students from the Radboud University, Nijmegen took part in the experiment, and were paid for their participation.

4.3.2.2 Procedure

The participants were instructed to describe houses or apartments they grew up in or were very familiar with to an interlocutor. They were asked to describe the spatial layout such that their interlocutor would be able to recognize the place and find their room. After describing a first place, they were to move on to the next house or apartment. The interlocutor was a confederate, who was instructed to try to understand the description and ask questions in case of problems. During the description, the participants were presented via a small earplug with a high or a low tone lasting 200 ms. The tones were presented in a randomized order and at randomized time intervals between 20-40 sec. The participants were instructed to stop speaking and moving as quickly as possible on hearing the tone and to say aloud if the tone they had heard was high or low. After responding, they were to continue with their description.

After the instructions, the high and low tones were played to the participants and the loudness was adjusted to a level the participant was comfortable with. Participants were then allowed to practice until they were ready and had no further questions.

4.3.2.3 Equipment

The sessions were videotaped with a PAL DV camcorder. The DV-data were digitized and compressed into MPEG 1 format and annotated with the annotation tool MediaTagger (Brugman & Kita, 1995).

4.3.2.4 Transcription and coding

All instances where the stop signal occurred while participants were gesturing and speaking were transcribed and coded.

4.3.2.4.1 Transcription of speech

For all instances in which the stimulus (high/low tone) occurred while participants were speaking and gesturing, the speech (≈15 sec before and after the speech suspension) was transcribed verbatim from the digitized video in exactly the same manner as described for Corpus Study 1 (see section 3.2.1.4, p. 58 for details). Filled pauses (eh, em), silent pauses (200 ms and longer), indications of speech suspensions like glottal stops, truncated words were included in the transcripts.

4.3.2.4.2 Coding speech

Each speech suspension that was accompanied by a gesture at the moment of stimulus presentation (high / low tone) was coded for suspension point and resumption point. This was performed by tagging the begin point and the end point of the cut-off-to-repair interval in the same manner as in Corpus Study 1 and 2 (see section 3.2.1.4.3, p. 59 for details).

4.3.2.4.3 Coding gesture and gesture suspensions

The gestural phases were first segmented and coded based on the scheme developed for Corpus Study 2 (see section 4.2.3.5, p. 104): preparation, stroke, hold, retraction, partial retraction, and interrupted preparation/stroke. Phase transitions were also coded as suspensions according to the same criteria as in Corpus Study 2 (see section 4.2.3.5.3, p. 110 for details).

4.3.2.5 Analysis

To assess the time it took speakers to suspend gesturing and speaking, the time interval from stimulus onset to movement/voice suspension was measured. In the experiment the stimuli (high/low tones) occurred at a randomized time interval of 20-40 sec. It was likely that some of the stimuli occurred at a point in time at which gesture or speech was about to end independently of the stimulus, simply because the speaker had reached the end of a sentence and/or the end of a gesture. In these instances, the suspension of

speech or gesture could not be considered as a response to the stimuli. As a first step to ensure that the suspension of gesture and speech was induced by the stimuli, responses with very short reaction times (below 120 ms) were excluded.

4.3.3 Results

Of the 20 participants, three had to be excluded due to technical recording problems. In one of the participants no stimuli occurred while the participant was speaking and gesturing at the same time. The remaining 16 participants heard 40 stimuli during 20 min of descriptions (total of 640).

Overall 7.8% of the gesture suspensions (N = 282) of both hands and 4.6% of the speech suspensions (N = 459) that occurred had reaction times of 120 ms or below, making it unlikely that they were due to the stimuli. Those gesture and speech suspensions were excluded from the analysis. In addition, instances in which the participants were either speaking but not gesturing or gesturing but not speaking were excluded from the analysis. Only the 148 instances in which the stimuli occurred while participants were speaking and gesturing were entered into the analysis. Each of the 16 participants contributed on average 9.25 data points, with a range of 5 to 25 observations per subject.

If the participant was gesturing with both hands when the stimulus occurred, either both hands were suspended at the same time or one hand stopped earlier than the other. In the former case only one response was included, in the latter case only the data point from the hand that showed the earliest reaction to the stimuli was included. All asynchronies that differed more than two standard deviations from the mean asynchrony were considered to be outliers and were excluded from the analysis.

The mean reaction time from stimulus onset to speech suspension was 392 ms (9.78 frames, SD = 141 or 3.53 frames). The mean reaction time from stimulus onset to the first gesture suspension differed only minimally: 391 ms (9.80 frames, SD = 133 or 3.34 frames). The mean asynchrony between gesture and speech suspensions for a given tone was 1 ms (0.02 frames, SD = 190 or 4.74 frames). The mean gesture-speech asynchrony was not significantly different from zero ($t(140) = .07$, n.s.), implying that gesture and speech stopped at the same time.

4.3.4 Discussion

This experiment addressed the question whether the suspension latencies for gesture and speech differ or not when presented with an external stop signal. The underlying assumption was that modality-specific differences might be due to different suspension latencies, which would show up in stop signal reaction time in a paradigm that provides a common stop signal for both modalities. No evidence was found that the latencies differ for the two modalities. This suggests that the observed suspension latencies can be interpreted as reflecting the latency of an internal stop signal plus a constant relay time, which is the same for both modalities. This in turn means that the asynchronies observed in Corpus Study 2 must have been due to differences in the timing of an internal stop signal rather than differences in relay times.

It cannot be excluded that the latency of an externally triggered internal stop signal is different from the latency of an internal stop signal triggered by an internal event such as error detection. However, both cases ultimately involve the internal generation of a stop signal. In the case of the current experiment the participant was explicitly instructed to stop speaking and gesturing, hence, to internally generate stop signals upon the external stimulus. There seems to be no a priori reason why an internal signal generated in this manner should differ from one generated on the basis of a perceived error, with only the latter resulting in modality specific differences. More importantly there is no reason to assume that the latency of an internal stop signal induced by our experimental procedure should be modality specific.

It may be noted that the stopping time for speech in the current experiment is longer than the 200 ms for stopping speech that has been estimated by, for example, Logan and Cowan (1984) from Ladefoged, Silverstein, and Papcun (1973). Close consideration of Ladefoged et al. (1973) reveals that this estimate is only valid for cases in which a signal to stop speaking arrives when the speaker is not engaged in planning. Ladefoged et al. (1973) found that when the signal to stop speaking arrived during a planning phase, latencies to stop speaking increased, falling in a range between 200-500 ms (see Ladefoged et al., Figure 4, p. 1107). Because in the current experiment, the signal to stop speaking arrived at random moments, the estimate of stopping time

would include many instances where speakers were engaged in planning. Hence, the observed mean stopping time of 390 ms falls within the 200-500 ms range obtained by Ladefoged et al. (1973).

An alternative explanation for the almost identical reaction times for gesture and speech that exceed the times reported in the literature (Logan & Cowan 1984) is that speakers might have aligned the stopping of the modalities, for example interpreting the instructions as implying that they should stop the two modalities simultaneously. In this case, one of the two modalities might have actually a faster stopping time, but the difference would be masked by the subjects efforts to synchronize the stopping times. To do so speakers would have to delay, for example, the interruption of gesture to align the stopping of speech. Coordinating the two modalities would seem to call for some amount of executive processing, which should inflate the stopping times for both. While this possibility cannot be completely ruled out, the fact that the observed speech stopping times were in the range of speech stopping times in the speech modality only in Ladefoged et al. (1973) seems to make this explanation unlikely. To provide definitive evidence would require further studies in which the instructions would require stopping of only one modality (speech or gesture) at a time.

5. General Discussion

Chapter 5

Traditionally, psycholinguistic approaches to self-monitoring have focused on how speakers avoid producing erroneous speech. The studies in this dissertation have attempted to address self-monitoring from a broader conversational perspective, in which the need to produce accurate speech is but one of various demands placed on the speaker in face-to-face conversation. This led to a consideration of how speakers trade off accuracy against fluency, as well as how they manage the multimodality of conversational performance during speech disfluency. To investigate these issues, a corpus of living space descriptions was recorded, and speech and gesture were analyzed. The dissertation reports the results from two studies based on this corpus. In this chapter, I summarize the findings and discuss issues related to their generalizability as well as identify some avenues for future research.

Summary of the findings

The first study addressed the question whether speakers interrupt their speech stream upon error detection (Main-Interruption-Rule hypothesis) or upon repair readiness (Delayed-Interruption-For-Planning hypothesis) by considering repair complexities, cut-off-to-repair durations and speech suspension types in a corpus of living space descriptions. The MIR hypothesis predicts that for within-word suspensions, the duration of the cut-off-to-repair interval should reflect replanning time (minus suspension latency). Thus, given a within-word suspension, this cut-off-to-repair interval should be longer for major repairs (such as a fresh start) than for minor repairs (such as a phoneme substitution). This prediction was not supported. Instead, it was found that the cut-off-to-repair interval for major and minor repairs did not differ when speech was suspended within-word. The result indicates that a part of the replanning

process must have taken place before suspension. This is consistent with the prediction of the DIP hypothesis, that speech is interrupted as soon as repair processing has come to completion no matter whether a major or a minor repair was processed.

Another problematic finding for the MIR hypothesis was that major repairs followed within-word suspension after very short or even nil (0-40ms) cut-off-to-repair intervals. This should not occur if replanning must wait until all subcomponents of the speech production system have been stopped upon error detection (Levelt, 1983). The result is still problematic even if, following Hartsuiker and Kolk (2001), it is assumed that replanning and interruption are processed in parallel with some 50 ms available for replanning. This time interval is too short for planning a major repair, such as a fresh start. The DIP hypothesis, in contrast, predicts such results since it assumes that speech interruption is initiated upon repair readiness.

The current study distinguished between within-word and after-word suspensions. However, not only are there variations in the location of the suspension but also in the articulatory patterns of the suspension word, such as lengthened syllables, word final devoicing, glottalization, laryngalization, and prosodic markings (Bell et al., 2003; Berg, 1986). These variations suggest that at least on some occasions, the way speech is suspended is planned. Hence, the interruption cannot be just the result of a simple stop signal released in a reflex-like manner upon error detection as suggested by the MIR hypothesis. Conversely, if the suspension is planned, it is also not just the result of running out of buffered material as suggested by the DIP hypothesis. A more fine-grained investigation of suspension words might have confirmed such variations in articulatory patterns also in the present corpus and provided further evidence for more complex suspension planning that theories of self-monitoring and speech interruption would need to accommodate.

In Corpus Study 2, two questions were investigated: whether gesture is sensitive to speech disfluencies and whether gesture provides a further source of evidence for immediate or delayed speech interruption. To this end, characteristics of gestural behavior during disfluent utterances and fluent utterances were compared. Separate analyses were conducted for overt repairs with within- and after-word suspensions in

order to discuss the implications of the results with respect to the DIP hypothesis and the MIR hypothesis.

No differences were found in the rate of gestural activity during disfluent utterances; however, differences were found in the timing of the gesture suspension as well as in its location within the gesture phrase. Gesture was on average suspended before speech in disfluent utterances but not in fluent utterances. This was true for covert repairs and overt repairs following after-word suspensions, but not for overt repairs following within-word suspensions.

The results of the gesture suspension position analysis showed that suspensions in early phases of gesture execution were three times as likely in disfluent utterances compared to fluent utterances. Specifically, in disfluent utterances, gesture suspension tended to occur within a preparation or a stroke or right after a preparation. Taken together, these results suggest that speech disfluency does not lead to changes in the overall amount of gesturing, but rather results in a temporal reorganization of the gesture phrase.

The fact that the gestural response was prepositioned in the timing analysis suggested that there were two stop signals released in succession, first to gesture then to speech. In order to control for the possibility that this result was due to intrinsic differences in the mechanics of stopping between the two modalities, a control experiment was conducted. Participants performed the same living space description task as in the corpus study, but were instructed to stop speaking and gesturing immediately upon hearing a tone. Because the stop signals were triggered simultaneously by the tone, any observed asynchrony in stopping time would be the result of a difference in suspension latencies between the two modalities. However, no such asynchrony was found, indicating that gesture and speech do not differ in this respect.

In sum, the gesture study provided evidence that gesture is sensitive to speech disfluency. Moreover, the experiment provided evidence that the prepositioning of gesture suspensions in relation to speech suspensions is not due to differences in relay

times of the two modalities. This result suggests that gesture suspension is closer in time to error detection than speech suspension.

We discussed this finding with respect to two specific scenarios that could differentiate the MIR hypothesis and the DIP hypothesis: (1) an earlier stopping latency for gesture in the case of within-word suspensions, which would provide evidence against the MIR and in favor of the DIP; and (2) a large asynchrony with gesture stopping earlier in the case of after-word suspensions, which would support the DIP and contradict the MIR hypothesis. Given that neither of these two critical scenarios was observed, the study was inconclusive with respect to these hypotheses. Instead, it was found that: (1) in the case of within-word suspensions, no difference was observed in the relative stopping times for gesture and speech, and (2) in the case of after-word suspensions there was a small but reliable stopping asynchrony with gesture stopping earlier than speech. These findings could be explained by either hypothesis.

A tentative reconciliation of the findings of the speech and the gesture study

How can the results of the speech study (Corpus Study 1) and the gesture study (Corpus Study 2) be reconciled with respect to the MIR and the DIP hypotheses? The results of the speech study have provided evidence in favor of the DIP hypothesis. It is unclear how the MIR could accommodate the speech result, in particular the observed similar cut-off-to-repair latencies for major and minor repairs following within-word suspensions, without giving up the central assumption of immediate speech interruption upon error detection.

This consideration leaves the DIP hypothesis as the only viable account for both the speech and the gesture data. However, the suggested explanations of the DIP hypothesis for the gesture data are not unproblematic as discussed in Chapter 4. A first possible account was to assume two different mechanisms for gesture suspension, namely gesture interruption upon repair readiness in cases of within-word suspensions and gesture interruption upon error detection in cases of after-word suspensions. We rejected this possibility because the speaker cannot foresee whether or not a repair

readiness signal will come in in time or not and therefore has no basis for the decision whether or not to interrupt gesture.

However, we saw that the DIP might also explain the gesture findings by assuming that gesture and speech are continued upon error detection, and that gesture stops before speech because the gesture buffer runs out of prepared material before the speech buffer. These considerations are highly speculative and require further testing given their post-hoc nature. Furthermore, they make assumptions about gesture, such as the existence of a gesture buffer, which currently have not been fully explored.

Nonetheless, it seems that only the DIP hypothesis provides the possibility to explain the observed data pattern in the speech study as well as in the gesture study in a consistent way. Furthermore, as has become clear in the discussion section of the speech chapter, the DIP hypothesis is compatible with the most explicit model of speech production (Levelt, 1989; Levelt et al., 1999).

Generalizability

The findings of the two studies were based on an analysis of the speech and gesture produced by speakers in a living space description task. To what extent might the results depend on the nature of the particular task used? One might argue, for instance, that the living space description task used in this study promoted fluency more than accuracy, because there were no consequences for inaccurate communication. Thus, it could be argued that these were not the optimal circumstances under which to find evidence for the MIR hypothesis. Although this cannot be ruled out, it must be noted that this argument tacitly accepts the position that the interruption process may be flexible with respect to the speaker's broader communicative goals, a position that in itself contradicts the MIR hypothesis.

A second consideration regarding the generalizability of the results concerns the types of gestures that might have been elicited by this task. Different types of discourse genres elicit different types of gestures. In the living space description task, pointing gestures and iconic gestures were prevalent, while other types of gestures such as beat gestures, emblems and conventionalized pragmatic gestures were rare. It could be the

case that the patterns of gesture stopping behavior observed in the current study are specific to the most common kinds of gestures elicited by living space descriptions. A more interactive type of conversational situation, such as a discussion or a negotiation, is certain to elicit other types of gestures, which may have different characteristics in timing and co-expressivity. This could potentially lead to a different overall pattern of gestural behavior during disfluent speech from that which was found in the current study. Thus, further study of different discourse genres is needed, with closer attention to the stopping behavior for different kinds of gestures. The methodology and findings from the current study can provide a foundation for these efforts.

The perspective adopted in the present study was centered on speech, in that the analysis selected cases of speech disfluencies and then examined gestural behavior in response to the speech disfluency. However, the direction of influence may not be only from speech to gesture, but also from gesture to speech. Kendon (2004) has illustrated how gesture and speech are organized with respect to each other, showing that either gesture or speech is held up in order to accommodate the performance of the other modality. The fact that gesture is suspended *at all* raises the possibility that it is not only speech that speakers monitor for problems, but also gesture, and even the semiotic coordination between the two modalities. Seyfeddinipur (2002) provides evidence that gestures are also monitored for correctness and recipient design features, such as visibility. In such cases, speech is suspended when gesture is suspended and speech resumes when gesture resumes. While the gesture is altered in the resumption, speech is not altered but either repeated or continued, indicating that the problematic information was provided by the gesture. For example, a speaker pointing to the left while saying *right* suspended her speech and gesture. In the resumption, she repaired her gesture and pointed to the right while repeating the word *right*. This suggests that speakers also suspend speech when there are problems in gesture in order to preserve the temporal and semantic coordination of the modalities.

Broader Implications

In this section we discuss the broader implications of the observed pattern of gestural sensitivity to speech disfluency. To begin, this finding has important implications for cognitive models of speech and gesture production. Different kinds of models have been proposed to account for the temporal and semantic coordination of gesture and speech by assuming linkage of the production system at different levels. McNeill (1992, 2000) proposes that gesture and speech form a single system, from which a continuous interaction between the two modalities follows. De Ruiter (1998, 2000) and others (e.g., Kita & Özyürek, 2003) assume a link between the systems at the conceptualizer level. Krauss et al. (2000) have put forward a gesture speech production model that assumes the link between the systems to be at working memory and at word form level. Which of these models could best account for the above results?

The results are most problematic for models that assume that gestures are an epiphenomenon of the lexical retrieval process (e.g., Hadar, Wenkert-Olenik, & Soroker, 1996; Krauss et al., 1996; Krauss et al., 2000; Krauss & Hadar, 1999). One prominent example is Krauss et al. (2000), who assume that gestures aid in the retrieval of items from the mental lexicon via cross modal priming at the word form level. Assuming that at least some of the disfluencies were due to lexical retrieval problems one might expect a higher rate of gestural strokes in disfluent contexts, since gesture aids in lexical retrieval. Our results suggest that this is not the case. It might be noted, however, that Krauss et al.'s (2000) claims are restricted to a specific kind of gesture that they call lexical gestures, defined as gestures that resemble some part of their referent iconically. Although in the present study gestures were not classified into different types, a majority of the gestures produced in this context depicted spatial configurations. Hence, they would be iconically representing aspects of the referents such as the size or configuration of a room. Moreover, previous research has shown that gestures occur predominantly in searches for words with spatial content than for words with other contents and that when gesturing is restricted, speech production is especially impaired when talking about spatial content (Rauscher et. al., 1996).

The Krauss et al.'s (2000) model furthermore cannot account for the results of the timing analysis. This is because the mechanism proposed for gesture termination is reactive to events in speech, whereas the current results suggest that suspensions in gesture may foreshadow events in speech. Specifically, gesture termination is assumed to occur via a feedback loop from the auditory monitor to the gestural motor planner. This would not seem to allow for gesture termination to occur prior to the corresponding speech event. For this model to explain the temporal patterning of gesture and speech suspension, it would seem necessary to postulate an additional mechanism for gesture interruption that would operate internally, without mediation by auditory feedback.

The current results also appear problematic for the theory of gesture-speech coordination put forth by McNeill (1992, 2000). He proposes that gesture and speech form a single system and interact throughout their production. Thus, gestural activity and speech activity should always mirror each other. What this would predict, then, is that during periods of lower speech activity, such as during a disfluency, there should also be reduced gestural activity. Furthermore, there should be a higher suspension rate during disfluent compared to fluent periods. However, the analysis did not reveal any difference in the rate of gesture phases or in the rate of stroke phases during disfluent speech, or in the likelihood of suspensions.

In addition, McNeill's assumption of a unified system would not appear to predict the observed temporal patterning of speech and gesture suspension. Based on the assumption of full interaction throughout production, gesture and speech suspension should be simultaneous. However, the timing results do not fit this model, since gesture suspension tended to occur before speech suspension in cases of after-word suspensions.

The results can be explained by models that assume a speech-gesture link at the conceptualizer level. Such a model is the one proposed by De Ruiter (1998, 2000). In his model, a decision is made at the conceptualizer level regarding the distribution of the information to be communicated over the two modalities. The system is therefore maximally flexible relative to the communicative goals of the speaker. For example, when a speaker encounters a lexical retrieval problem, the speaker could decide to

continue the ongoing gesture so as to elicit assistance from the addressee. Alternatively, the speaker could choose to suspend the gesture until the corresponding speech is resumed. Thus, whether or not a speech disfluency results in a change in gestural behavior depends upon the speaker's goals in a given situation. Because of this flexibility, it is not clear what this model would predict for the gestural activity in disfluent utterances. With regard to the timing of gesture suspensions relative to speech suspensions, De Ruiter's model assumes that the conceptualizer is the locus of error detection as well as stop signal generation. For synchronous speech and gesture suspension he suggests that upon error detection stop signals are sent to the formulator and to the gesture planner module, which pass the signal on to the lower modules. Although we have come to the conclusion that synchronous gesture and speech suspension is more likely triggered by repair readiness than error detection, the mechanism of simultaneously sending stop-signals to the formulator and the gesture planner as suggested by De Ruiter is still possible. With respect to the observed gesture-speech asynchrony in cases of after-word suspensions, De Ruiter's model would be compatible with both the MIR and the DIP account. Under the MIR account, stop signals might be sent to the respective modality at different points in time. Under the DIP account, there would be no stop signals at all, since in both modalities suspension would be the result of the respective buffers running out of material.

In sum, De Ruiter's (1998) model is maximally flexible and seems to be able to accommodate all findings of the gesture study, whereas the models proposed by Krauss et al. (2000) and McNeill (1992, 2000) have problems to account for some of the findings.

Up to this point, all of the discussion has focused on the evidence for gestural sensitivity to speech disfluency. This evidence was based on observed differences between disfluent utterances and fluent baseline utterances. Note, however, that the distribution of gesture suspension latencies in disfluent and fluent baseline utterances suggests that in a considerable number of disfluent utterances the gesture suspensions did not differ from fluent baseline utterances. This implies that gesture may be suspended in some cases but not in others. We will now discuss why this might be the case.

One approach that may prove fruitful toward explaining some of this variability is the idea that gesture and speech are parallel channels of a multimodal signaling system (Kendon, 2004; Clark, 1996). Under this view, the temporal and semantic integration of the *speech-gesture ensemble* is a speaker achievement (Kendon, 2004). The integration of gesture and speech is not fixed, but varies according to communicative purpose, the semiotic affordances of the channels and the material to be expressed. An optimal distribution of information is sought in order to maximize communicative efficacy and minimize effort (Clark, 1996). To ensure that the intended message is properly conveyed, the speaker must coordinate the information presented in the two modalities. A disruption in speech can potentially disrupt this coordination and thereby threaten the integrity of the message that the speaker is attempting to convey by the gesture-speech ensemble. Under this view, speakers would take this perceived threat into account when determining whether and when to stop their gesture. The extent to which a speech disruption poses a threat will depend upon the nature and degree of temporal and semantic coordination that is required to successfully convey the intended message. For example, consider the difference in gesture-speech coordination required for an expression such as *over there*, accompanied by a pointing gesture. Normally in such instances, the pointing gesture should be in place as the speaker utters the deictic word *there*, since the deictic expression cannot be comprehended without the direction indicating gesture. However, a delay in speech may pose very little threat to the integrity of the message. If the pointing gesture is in place before the speech is ready, it can simply be held in place for as long as is necessary for the repair to be effectuated, without any loss to the integrity of the message. In contrast, consider an iconic gesture displaying the spiraling downward movement of an object accompanied by the phrase *and then it went down*. Here, the semantic and temporal coordination of the gesture is crucial, because the non-redundant iconic information presented in gesture is dynamic and cannot be visually stabilized through a hold. So in this case, if speakers anticipate a delay in speech then they may choose to withhold the gesture until the point at which both can be articulated simultaneously.

A related point is that when speakers anticipate a disfluency, they are likely to have some estimate of the length of an impending cut-off-to-repair interval

(Clark, 2002; Clark & Fox Tree, 2002). If they are expecting a brief hiatus, then the gesture may be allowed to run to completion because it would not threaten the integrity of the composite signal. In support of this idea, it was informally observed in the current data set that gesture tended to overrun speech disfluencies in cases of very fast closed class item repetitions, like *und auf der der linken Seite* ('and on the the left side'), where the definite article is repeated. Also Kita (1993) did not find any evidence that repetitions affect gesture execution.

Future research is needed to explore these possibilities. Such an investigation would need to operationalize variables related to the degree of integration between gesture and speech such as information distribution, level of dependencies between gestural and speech information, visual stability of gesture, and the level of threat posed by a speech disfluency. Further systematicities in gesture-speech coordination may be revealed.

Concluding statement

It is customary in the psycholinguistic tradition of research on error monitoring and repair to focus on the speech channel alone. Furthermore it is generally assumed that the aim of the speaker is to produce accurate speech. Our results have shown that disfluencies are not only a phenomenon of the speech channel, but are also reflected in gesture. Furthermore, our results on the relationships between the timing of speech suspensions, the types of speech suspensions and the repair complexity, show that speakers deploy a strategy that favors fluency over accuracy. The conversational environment to which the language production system is adapted places broader demands on speakers than accuracy of expression alone. Because the nature of the processes underlying self-monitoring is determined by these demands, it is important to take them into account when constructing theories of self-monitoring in language production.

Often, cognitivist approaches to language abstract away from interaction, while interactional approaches abstract away from cognition. However, these approaches should not be seen as mutually exclusive; indeed, they both seem necessary for the

elaboration of a comprehensive theory of language use. On the one hand, the interactional approach provides a rich descriptive framework within which it is possible to evaluate the functionality of processes observed in the laboratory. Moreover, it provides a specification of the basic communicative goals of speakers and the relevant variables of face-to-face communication. These serve as useful guideposts for research on language use. On the other hand, the cognitivist approach provides a processing framework for understanding behavior, with an emphasis on detailed, testable predictions. It seems that a synthesis of these approaches offers the best chance of deepening our understanding of language use.

Samenvatting

In alledaagse conversaties komt het tijdens het spreken regelmatig tot haperingen (*disfluencies*) en gebeurt het dat de spreker een woord of zin afbreekt of verbeterd. Zulke spreekfouten zijn interessant, omdat zij de complexe processen bloot leggen die aan het plannen van gesproken taal ten grondslag liggen. De oorzaaken van deze disfluencies moeten met name worden gezocht in de semantische, morfologische, fonologische of articulatorische planning. Op ieder van deze planningsniveaus kunnen zich problemen voor doen, met als gevolg een hapering of fout in de gesproken boodschap. Omdat de spreker er naar streeft dat de informatie die hij geeft correct en goed te begrijpen is, volgt hij zijn eigen spraak (zowel gepland als uitgesproken) nauwlettend en toetst hij voortdurend of wat hij zegt toepasselijk en correct is. Als hij een fout tegen komt moet hij vervolgens beslissen hoe hij daarmee omgaat. De vraag is dan *of*, *wanneer*, en *op welke manier* hij zijn eigen spraak onderbreekt om de fout te herstellen.

Eén mogelijkheid is dat de spreker direkt na het ontdekken van een fout zijn spraak onderbreekt. Op die manier zorgt hij ervoor dat de luisteraar zo weinig mogelijk foutieve informatie te horen krijgt. Het streven naar correctheid wordt echter gecomplementeerd door de wens om de conversatie zo soepel mogelijk te laten verlopen. Het gaat bij communicatie immers om een interactief proces waarbij spreken en luisteren elkaar afwisselen, en waarbij de spreker rekening moet houden met de reacties van de luisteraar. De spreker wil zijn verhaal kunnen doen zonder door de andere spreker te worden onderbroken of het woord te verliezen. Dit maakt het noodzakelijk om haperingen en pauzes zo veel mogelijk te voorkomen. Als hij hier niet in slaagt loopt hij gevaar te worden onderbroken, of om ongeconcentreerd en niet

welsprekend over te komen (Clark & Wasow, 1998). Sprekers moeten daarom voortdurend de eisen van *fluency* en correctheid in balans zien te houden. Daarom kan het gebeuren dat zij na het ontdekken van een fout gewoon verder spreken terwijl zij tegelijkertijd een correctie plannen. Op deze manier kunnen zij haperingen zo veel mogelijk voorkomen en tegelijkertijd snel de fout corrigeren.

Bij conversaties waarbij de beide partners elkaar kunnen zien, wordt de situatie verder gecompliceerd door het feit dat de interactie niet tot de verbale modaliteit beperkt is. Sprekers luisteren niet alleen naar elkaar, ze letten ook op elkaars non-verbale gedrag. Vaak maakt de spreker daarbij gebruik van gebaren (vooral bewegingen van handen en armen), bijvoorbeeld om de grootte, de locatie, het bewegingstraject, of de vorm van een referent over te brengen.

Spraakbegeleidende gebaren (gesticulatie) zijn zowel in het tijdsdomein als ook in het betekenisdomein nauw aan de gesproken woorden verbonden. Het onderdeel van de beweging dat de betekenis draagt valt tijdens het spreken in de tijd samen met de woorden die deze betekenis uitdrukken (Kendon, 1983; McNeill, 1992). Als sprekers bij het spreken moeilijkheden ondervinden, wordt dit verband mogelijk verstoord. Dit leidt tot de vraag of spraakbegeleidende gebaren beïnvloed worden door disfluencies tijdens spraak.

In dit proefschrift wordt onderzocht hoe sprekers met de tegenstrijdige principes van *vloeiendheid* en *correctheid* om gaan als zij tijdens het spreken problemen ondervinden. Het onderzoek werd daarbij gestuurd door twee vragen:

1. Onderbreken sprekers hun spraak direct op het moment dat zij een fout ontdekken, of op het moment dat zij klaar zijn om de fout te herstellen?

2. Wat is het effect van versprekingen op de spraakbegeleidende gebaren? Kunnen gebaren inzicht verschaffen in de processen die aan het onderbreken van spraak ten grondslag liggen?

Het merendeel van eerder onderzoek naar zelfonderbreking en correctie is gebaseerd om data die met behulp van experimenten verkregen zijn waarin sprekers op elkaar lijkende patronen herhaald beschrijven (Levelt, 1983; Oomen, 2001). De proefpersonen beschikken in een dergelijke setting niet over een gesprekspartner, maar weten alleen dat hun beschrijvingen zullen worden opgenomen en afgeluisterd. Het interactionele karakter van conversatie is daarmee in dit soort onderzoeken afwezig.

In het hier beschreven onderzoek is daarom voor een alternatieve opzet gekozen. Er werd een corpus van zogenaamde *living space descriptions* samengesteld. Daarin beschrijven sprekers in een interactionele setting (de inrichting van) een huis dat voor de luisteraar niet bekend is. Deze taak leidt tot veel gebaren, maar ook haperingen en versprekingen. In de Hoofdstukken 3 en 4 worden de analyses van deze data beschreven.

In Hoofdstuk 3 worden twee hypotheses over de processen die ten grondslag liggen aan het onderbreken van spraak getest. De zogenaamde *main-interruption-rule* (MIR) hypothese gaat ervan uit dat sprekers ernaar streven de hoeveelheid foutieve informatie die de luisteraar te horen krijgt zo gering mogelijk te houden (Levelt, 1983, 1989). Volgens de MIR-hypothese wordt spraak direct onderbroken als hetgeen dat gecorrigeerd moet gaan worden (het *reparandum*), als foutief herkend wordt. Als de fout pas op een relatief laat tijdstip gedetecteerd wordt, zodat een directe onderbreking zou leiden tot een onderbreking binnen een correct woord dat het reparandum opvolgt, dan wordt de onderbreking opgeschort. Het resultaat is dan een pauze na het woord (*after-word suspension*). Ook als het reparandum niet foutief, maar slechts ontoepasselijk is, wordt de onderbreking tot na het woord opgeschort.

In tegenstelling hierop stelt de *delayed-interruption-for-planning* (DIP) hypothese dat sprekers er naar streven om de lengte van een aarzeling die ontstaat bij het corrigeren van een fout zo kort mogelijk te houden. Spraak wordt daarom pas onderbroken op het moment dat de correctie reeds gepland is, en klaar is om geuit te worden. Als een spreker een fout ontdekt zal hij volgens deze hypothese door blijven spreken en tegelijkertijd een nieuw planningsproces opstarten. Als de *monitor* (een interne bewakingsmodule die het spraakproces controleert) aangeeft dat het systeem

klaar staat om de fout te corrigeren, zal de spraak onderbroken en een correctie uitgevoerd worden. De spraak wordt volgens deze hypothese echter ook onderbroken als er in het systeem geen voorbereid spraakmateriaal meer te beschikking staat.

De twee hypotheses onderscheiden zich met betrekking tot de hoeveelheid correctie-planning die verborgen (tijdens het spreken) en openlijk (na afbraak van de spraak, herkenbaar als pauze) plaats kan vinden. De MIR-hypothese stelt dat sprekers direct na het ontdekken van de fout het spraakproces onderbreken door het hele spraakproductie systeem stop te zetten. Als deze hypothese klopt, dan kan de correctie-planning alleen *na* de spraakonderbreking plaats vinden, tijdens het zogenaamde *cut-off-to-repair interval*. Voordat spraak weer door kan gaan moet er, tijdens een pauze, eerst een nieuwe uiting gepland zijn. Onderzoek van Blackmer en Mitton (1991) heeft echter uitgewezen dat een correctie (*repair*) zonder enige pauze op een spraakonderbreking van volgen. Dit suggereert dat de correctie-planning nog *voor* de onderbreking heeft plaats gevonden. De MIR-hypothese kan dit resultaat verklaren door aan te nemen dat (a) niet het volledige spraakproductie systeem wordt gestopt, maar alleen de articulatie, (b) correctie-planning parallel met de onderbreking gestart wordt, en (c) correctie-planning afgesloten is op het moment dat de onderbreking inzet (Hartsuiker & Kolk, 2001). De tijd die ter beschikking staat voor een correctie-planning is daarmee gelijk aan de tijd die nodig is om een onderbreking uit te voeren (ca. 150-200 ms). Dit interval is vrij kort. Een correctie die zonder enige vertraging op de onderbreking volgt zou daarom een redelijk kort stukje spraak moeten betreffen (bijvoorbeeld een foneemvervanging of eventueel de vervanging van een enkel woord). Omvangrijkere correcties (zoals een *fresh start*, bijvoorbeeld "*Wenn man links in eh... Vorm Haus war eine Garage*", 'When one left into uh... In front of the house was a garage') vereisen echter een complexere lexicale en syntactische planning. Het is onwaarschijnlijk dat dergelijke complexe correctie-planningen binnen de beschikbare tijd kunnen worden uitgevoerd.

Het belangrijkste verschil tussen de twee hypotheses ligt dus in de hoeveelheid tijd die ter beschikking staat voor verborgen correctie-planning. Volgens de MIR-hypothese is dit interval beperkt tot de tijd die nodig is om de spraak te onderbreken (of een woord dat zojuist wordt gearticuleerd af te maken). Als er meer tijd nodig is zal de

planning voortgezet worden in de tijd tussen het afbreken van spraak en het begin van de correctie (*cut-off-to-repair interval*). Volgens de DIP-hypothese gelden dergelijke beperkingen echter niet. De spreker kan verborgen correctie-planningen uitvoeren zolang er voldoende materiaal beschikbaar is voor de *formulator* en de *articulatory buffer*.

Deze hypothesen zijn in het hier beschreven onderzoek getest met behulp van een corpus analyse waarin de manier waarop spraak afgebroken wordt (binnen het woord of na het woord) in verband werd gebracht met de lengte van het volgende *cut-off-to-repair interval* en met de complexiteit van de correctie. Gevallen van *within-word suspension* zijn óf het gevolg van een directe onderbreking na het ontdekken van de fout (MIR-hypothese), óf zij geven het moment aan waarop de correctie klaar was (DIP-hypothese).

Voor de MIR-hypothese betekent dit dat verborgen planningsprocessen hooguit 150-200 ms hebben kunnen geduurd (dwz., de tijd die nodig is om te onderbreken). Daarentegen zijn gevallen van onderbrekingen na het woord minder eenduidig. Zij kunnen aan de ene kant het resultaat zijn van een directe onderbreking, waarbij het tijdstip van onderbreking toevallig tussen twee woorden valt. Aan de andere kant kan het ook om gevallen van vertraagde onderbreking gaan, waarbij de spreker eerst het woord af maakt alvorens aan een correctie te beginnen. In het laatste geval zou de lengte van het correctie-interval aanzienlijk langer kunnen zijn, namelijk even lang als het duurde het woord te beëindigen, plus de tijd die het kost om het proces te onderbreken (*interruption latency*).

Omdat in gevallen van onderbreking *na* het woord meer verborgen correctie-planning uitgevoerd kan zijn voordat de spraak daadwerkelijk wordt afgebroken zouden de erop volgende *cut-off-to-repair* intervallen korter moeten zijn dan in het geval van *within-word suspension*. In het laatste geval zou er maar weinig correctie-planning plaats kunnen vinden *voor* afbraak, en zou een groot deel van de planning *na* afbraak plaats moeten vinden. Volgens de DIP-hypothese vindt een *within-word suspension* plaats op het moment dat de correctie-planning is afgerond. Een afbraak aan het woordeinde kan daarentegen ook het gevolg zijn van onvoldoende ter beschikking

staand spraakmateriaal. Daarmee voorspelt de DIP-hypothese in vergelijking met de MIR-hypothese het tegenovergestelde patroon: langere *cut-off-to-repair* intervallen na afbraak aan het woordeinde dan na afbraak binnen het woord.

De eerste analyse in Hoofdstuk 3 laat zien dat sprekers na een afbraak binnen het woord minder lang pauzeren dan na een afbraak aan het einde van een woord. Dit resultaat spreekt voor de DIP-hypothese en suggereert dat er meer verborgen correctie-planning plaats vindt in het geval van afbraak binnen het woord dan in het geval van afbraak aan het woordeinde. Het resultaat is echter ook verklaarbaar vanuit de MIR-hypothese, als men ervan uit gaat dat de correcties in de gevallen van afbraak binnen het woord gering van omvang waren en daarom minder planningtijd nodig hadden dan de correcties die na het woordeinde plaats vonden. De verdeling van de gemeten correcties steunt deze laatste interpretatie.

In een tweede analyse werden alle correcties als omvangrijk (zoals het eerder gegeven voorbeeld) of als gering van omvang (bijvoorbeeld "*Rechts eh links war das Bad*", 'right uh left was the bathroom') geclassificeerd. De MIR-hypothese voorspelt dat het *cut-off-to-repair interval* afhankelijk is van de complexiteit van de correctie, zodat complexe correcties meer tijd nodig hebben dan simpele. De DIP-voorspelt daarentegen dat het *cut-off-to-repair interval* onafhankelijk zou moeten zijn van de complexiteit (in het geval van afbraak binnen het woord), omdat het tijdstip van onderbreking puur afhankelijk is van het moment wordt waarop het systeem klaar is voor de correctie. Een analyse van de *disfluencies* in het corpus laat zien dat in het geval van *within-word suspension* de tijd tot het herstarten van spraak inderdaad onafhankelijk is van het type correctie. Dit resultaat bevestigt de DIP-hypothese en kan niet worden verklaard met behulp van de MIR-hypothese.

Een verdere analyse laat zien dat sprekers een omvangrijke correctie zodanig konden afronden dat zij konden beginnen met articuleren zonder enige voorafgaande pauze. Omdat de verwerkingstijd die nodig is voor het verwerken van een omvangrijke correctie langer is dan de tijd die het kost om spraak te onderbreken, moet de herberekening van de uiting al tijdens het spreken zijn gestart.

De resultaten suggereren dat de sprekers in eerste instantie er naar streefden zo vloeiend mogelijk te blijven spreken en dat zij spraak niet direct afbraken bij ontdekking van een fout, maar pas op het moment dat zij de correctie klaar hadden liggen.

Het grootste probleem bij het onderzoek naar het tijdstip van initialisering van onderbreking is dat het moment van de interne fout-detectie niet bekend is. Een dergelijke maat zou nodig zijn om te bepalen of de tijd tussen ontdekking van de fout en het onderbreken van spraak even lang is als de tijd die nodig is om spraak af te breken (zoals voorspeld door de MIR-hypothese), of dat het langer duurt (zoals voorspeld door de DIP-hypothese). Een onafhankelijke maat voor het tijdstip waarop de fout ontdekt wordt, zou hier verder uitsluitsel over kunnen geven. In Hoofdstuk 4 werd onderzocht in hoeverre spraakbegeleidende gebaren een dergelijke maat kunnen leveren. In een aantal studies is aangetoond dat er tussen spraak en spraakbegeleidende gebaren een hecht verband bestaat, en dat beide processen zowel in het tijdsdomein als ook met betrekking tot de semantiek aan elkaar gekoppeld zijn (Kendon, 2004, McNeill, 1992). In een aantal onderzoeken is bovendien aangetoond dat spraak begeleidende gebaren beïnvloedt worden door haperingen (*disfluencies*) (Christenfeld, Schachter & Bilous, 1991; De Ruiter, 1998; Kita, 1993; Ragsdale & Silvia, 1982).

In Hoofdstuk 4 wordt getracht aanvullende evidentie te leveren voor een invloed van *disfluencies* op spraakbegeleidende gebaren. Daarvoor werden gebaren uit het corpus die tijdens vloeiende en niet-vloeiende spraakproductie uitgevoerd werden met elkaar vergeleken. Daarbij werden vooral zogenaamde *gesture suspensions* (het stoppen of afbreken van een gebaar) onderzocht. Specifiek werden er drie verschillende maten onderzocht: (1) de frequentie van *gesture suspensions*, (2) de positie van deze onderbrekingen in het tijdsverloop (ten opzichte van de spraak-onderbreking), en (3) de positie van de onderbrekingen in het verloop van het gebaar.

Het corpus onderzoek steunt de hypothese dat gebaren beïnvloedt worden door spraak *disfluencies*. Terwijl de frequentie van *gesture suspensions* in vloeiende en niet-vloeiende spraak even hoog is, laten de analyses zien dat in het geval van disfluencies het tijdstip van *gesture suspension* vóór het tijdstip van *speech suspension* ligt. Ook

liggen de *gesture suspensions* bij disfluencies in een vroegere fase van het gebaar dan bij suspensions geassocieerd met vloeiende spraak. De resultaten ondersteunen het idee dat de twee modaliteiten - spraak en gebaren - nauw met elkaar verbonden zijn. Het feit dat de relatieve positie van de *gesture suspensions* in het geval van niet-vloeiende spraak verschuift, kan geïnterpreteerd worden als een structurele reorganisatie van het gebaar als reactie op een *disfluency* in spraak.

In een volgende stap werd onderzocht of spraakbegeleidende gebaren verdere inzicht kunnen verschaffen in de vraag naar het tijdstip van spraakonderbreking. Zoals boven beschreven stelt de MIR-hypothese dat een dergelijke onderbreking plaats vindt zodra een fout opgemerkt wordt, en stelt de DIP hypothese dat de onderbreking pas plaats vindt als de correctie voldoende is voorbereid (*repair readiness*). Als gebaren in het geval van een *disfluency* eerder onderbroken worden dan de spraak zelf, kan men concluderen dat het tijdstip van *gesture suspension* dichter bij de oorspronkelijke fout-detectie zit dan het tijdstip van *speech suspension*. Noch de MIR-hypothese, noch de DIP-hypothese maken expliciete predicties met betrekking tot het gedrag van spraakbegeleidende gebaren. Er zijn echter twee kritieke gevallen waarin een vroege gebarenonderbreking gebruikt kan worden om deze hypotheses te toetsen:

1. Gevallen van expliciete correcties (*overt repairs*) die een within-word suspension opvolgen. Volgens de MIR-hypothese zouden deze gevallen het gevolg moeten zijn van een directe onderbreking na het ontdekken van een fout. Als de onderbreking van het gebaar echter aan de onderbreking van de spraak vooraf gaat, is dit kennelijk niet het geval.

2. In gevallen van afbraak na het woord (*after-word suspensions*) kan volgens de MIR-hypothese een gebaar eerder worden afgebroken dan spraak, omdat het spraakproductiesysteem de afbraak tot het woordeinde heeft uitgesteld. Is het interval tussen *gesture suspension* en *speech suspension* echter langer dan de tijd die nodig is om het woord af te maken, dan kan dit door de MIR-hypothese niet meer verklaard worden.

Beide scenario's konden niet volledig worden bevestigd. In het geval van *within-word suspension* werd gevonden dat spraak en gebaren op hetzelfde moment worden

afgebroken. In het geval van *after-word suspensions* werden gebaren eerder gestopt dan spraak, maar het interval tussen deze twee tijdstippen was niet langer dan de tijd die nodig was om het woord te beëindigen. De resultaten van het gebarenonderzoek zijn daarmee niet voldoende eenduidig om een van beide hypotheses te ondersteunen.

Een punt dat nog aandacht verdient is de interpretatie dat het afbreken van een gebaar eerder inzet dan het afbreken van spraak. Deze interpretatie is gebaseerd op de assumptie dat beide stop-processen even veel tijd nodig hebben, ongeacht de modaliteit. Het is echter ook mogelijk dat de onderbreking weliswaar synchroon geïnitieerd wordt, maar dat de geobserveerde asynchronie tot stand komt doordat het afbreken van een gebaar sneller gaat dan het afbreken van spraak. In Hoofdstuk 4 wordt een controle-experiment beschreven waarmee deze mogelijkheid is onderzocht. De experimentele setting leek sterk op de setting waarin de corpus data waren verzameld. De proefpersonen moesten huizen en appartementen aan een gesprekspartner beschrijven en werden daarbij gefilmd. Op verschillende momenten kregen zij een toon te horen, waarop zij moesten stoppen met spreken en gesticuleren. De tijd die nodig was om te stoppen verschilde niet tussen de modaliteiten. Dit resultaat ondersteunt de bovengenoemde assumptie dat een asynchronie tussen *gesture suspension* en *speech suspension* niet het gevolg is van een verschil in afbreek-latenties voor de twee modaliteiten.

Het in dit proefschrift beschreven onderzoek wijst uit dat het onderbreken van spraak beïnvloedt wordt door de interactionele setting waarin de spraak plaats vindt. De beslissing wanneer spraak onderbroken wordt, wordt niet alleen beïnvloedt door de intentie om zich zo correct mogelijk te uiten, maar ook door het streven naar een zo vloeiend mogelijke uiting. Dit kan ertoe leiden dat sprekers een onderbreking van hun eigen spraak uitstellen in plaats van direct een fout te herstellen. Het multimodale karakter van het spreken in een interactionele setting betekent ook dat zowel spraak als ook gesticulatie betrokken zijn als een fout gemaakt en vervolgens gecorrigeerd wordt. Conversatie drukt daarmee niet alleen een stempel op spraak en gebaar, maar ook op de manier waarop deze processen onderbroken worden.

Appendix

Overview of general speech and disfluency characteristics in the corpus

Table 6.1. Overview of speaking time, number of uttered words, and the number of words per minute.

Participant	Speaking time in min	Number of words	Words per min
AN	6.08	1098	180.69
AR	5.80	838	144.50
BI	7.86	1435	182.52
FA	9.04	1694	187.36
KA	8.21	1265	154.06
MA	8.40	1312	156.20
NI	8.68	780	89.85
NA	8.59	1619	188.46
SE	8.41	1283	152.62
SI	8.34	1374	164.83
SM	8.36	977	116.83
TO	8.53	1403	164.47
Total	96.3	15078	1882.39

Table 6.2. Overview of the number of disfluencies, disfluency rates per word, and per minute.

Participant	Disfluencies	Disfluencies per word	Disfluencies per min
AN	95	0.087	15.63
AR	61	0.073	10.52
BI	85	0.059	10.81
FA	99	0.058	10.95
KA	106	0.084	12.91
MA	110	0.084	13.10
NI	112	0.144	12.90
NA	111	0.069	12.92
SE	125	0.097	14.86
SI	124	0.090	14.87
SM	91	0.093	10.89
TO	83	0.059	9.73
Total	1202	0.997	150.09

Overview of general gesture characteristics in the corpus

Table 6.3. Overview of the number of gesture phases and stroke phases for each hand.

Participant	Gesture phases Left hand	Gesture phases Right hand	Stroke phases Left hand	Stroke phases Right hand
AN	356	505	137	187
AR	173	468	57	147
BI	453	859	148	304
FA	223	477	83	184
KA	333	610	93	182
MA	283	337	87	118
NI	59	420	21	157
NA	300	572	98	201
SE	220	434	61	154
SI	416	413	140	143
SM	217	270	84	99
TO	158	228	63	88
Total	3191	5593	1072	1964

Table 6.4. Overview of the summed duration of gesture phases for each hand and the summed duration of overlapping gesture phases.

Participant	Left hand (min)	Right hand (min)	Overlap (min)
AN	4.82	4.40	4.14
AR	3.83	4.11	3.36
BI	3.46	5.98	3.03
FA	1.60	3.66	1.33
KA	2.78	5.65	2.62
MA	3.76	3.58	2.93
NI	0.41	2.97	0.33
NA	2.10	3.95	1.83
SE	2.54	3.38	1.07
SI	4.01	3.74	3.05
SM	1.69	2.12	1.44
TO	1.12	1.63	0.46
Total	32.12	45.17	25.59

Table 6.5. Overview of the duration of gesture strokes in minutes by participant for each hand.

Participant	Left hand (min)	Right hand (min)
AN	0.70	0.99
AR	0.38	1.22
BI	0.97	1.91
FA	0.60	1.39
KA	0.62	1.41
MA	1.05	1.66
NI	0.17	1.22
NA	0.72	1.39
SE	0.38	0.93
SI	0.96	1.10
SM	0.70	0.92
TO	0.44	0.68

Below an overview is given of the gesture characteristics of the corpus, averaged over participants. Included are the number of gestural movement units, gesture phases, strokes, stroke rate, and gesture time. Note that results are for right and left hand independently as well as together, because some gestures include both hands, while others include only a single hand.

Table 6.6. Average number of gesture phases by hands (right and left).

Phases	mean	stdev	min	max	total
right hand	466.08	166.71	228	859	5593
left hand	265.92	113.23	59	453	3191
right & left hand	732	251.95	386	1312	8784

Table 6.7. Average number of strokes (expressive phases) by hands (right and left).

Strokes	mean	stdev	min	max
right hand	163.67	56.58	88	304
left hand	89.33	37.72	21	148
right & left hand	253	82.80	151	452

Table 6.8. Stroke rate.

Stroke rate	mean	stdev	min	max	total
strokes per minute	32.32	12.33	17.70	57.49	387.83

Table 6.9. Gesturally active time by hands (right and left).

Gesturing time in min	mean	stdev	min	max	total
right hand	3.76	1.25	1.63	5.98	45.17
left hand	2.68	1.33	0.41	4.82	32.12
right & left hand	6.44	2.28	2.75	9.44	77.29

References

Aboudan, R., & Beattie, G. (1996). Cross-cultural similarities in gestures. The deep relationship between gestures and speech which transcends language barriers. *Semiotica, 111*(3/4), 269-294.

Bavelas, J., Kenwood, C., Johnson, T., & Phillips, B. (2002). An experimental investigation when and how speakers use gestures to communicate. *Gesture, 2*(1), 117.

Bavelas, J. B., Chovil, N., Lawrie, D. A., & Wade, A. (1992). Interactive gestures. *Discourse Processes, 15*(4), 469-489.

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

Beattie, G., & Aboudan, R. (1994). Gestures, pauses and speech: An experimental investigation of the effects and changing social context on their precise temporal relationships. *Semiotica, 99*(3/4), 239-272.

Beattie, G., & Shovelton, H. (1999). Mapping the range of information contained in the iconic hand gestures that accompany spontaneous speech. *Journal of Language and Social Psychology, 18*(4), 438-462.

Bell, A., Daniel, J., Fosler-Lussier, E., Girand, C., & Gildea, D. (1999). *Forms of English function words – effects of disfluencies, turn position, age and sex, and predictability.* Paper presented at the 14th International Congress of Phonetic Sciences, San Francisco.

Bell, A., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustic Society of America, 113*(2), 1001-1024.

Berg, T. (1986). The aftermath of error occurrence: Psycholinguistic evidence from cut-offs. *Language & Communication, 6*(3), 195-213.

Berg, T. (1992). Productive and perceptual constraints on speech-error correction. *Psychological Research, 54*, 114-126.

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

Brédart, S. (1991). Word interruption in self-repairing. *Journal of Psycholinguistic Research, 20*(2), 123-138.

Brotherton, P. (1979). Speaking and not speaking: Processes for translating ideas into speech. In A. W. Siegman & S. Feldstein (Eds.), *Of speech and time: Temporal speech patterns in interpersonal contexts* (pp.179-209). Hillsdale, N.J.: Lawrence Erlbaum.

Brugman, H., & Kita, S. (1995). Impact of digital video technology on transcription: a case of spontaneous gesture transcription. *KODIKAS/CODE: Ars Semiotica, An international journal of semiotics, 18*, 95-112.

Cassell, J., McNeill, D., & McCullough, K.-E. (1999). Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics & Cognition, 7*(1), 1-33.

Catchpole, C., Hartsuiker, R. J., & Pickering, M. (2003). Self-corrections in speech: Evidence against Levelt's Main Interruption Rule. Poster presented at the Ninth Conference on Architectures and Mechanisms for Language Processing, Glasgow, Scotland.

Christenfeld, N., Schachter, S., & Bilous, F. (1991). Filled pauses and gestures: It's not a coincidence. *Journal of Psycholinguistic Research, 20*(1), 1-10.

Chui, K. (2005). Temporal patterning of speech and iconic gestures in conversational discourse. *Journal of Pragmatics, 37*(6), 871-887.

Cicone, M., Wapner, W., Foldi, N., Zurif, E., & Gardner, H. (1979). The relation between gesture and language in aphasic communication. *Brain and Language, 8*, 324-349.

Clark, H. H. (1994). Managing problems in speaking. *Speech Communication, 15*. 243-250.

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

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

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition, 84*(1), 73-111.

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

Cohen, A. A., & Harrison, R. P. (1973). Intentionality in the use of hand illustrators in face to-face communication situations. *Journal of Personal and Social Psychology, 28*(2), 276-279.

De Ruiter, J. P. A. (1998). *Gesture and speech production.* Unpublished Dissertation, Radboud University Nijmegen, Nijmegen.

De Ruiter, J. P. A. (2000). The production of gesture and speech. In D. McNeill (Ed.), *Language and gesture: Window into thought and action* (pp. 284-311). Cambridge: Cambridge University Press.

De Smedt, K., & Kempen, G. (1987). Incremental sentence production, self-correction, and coordination, *Natural language generation: recent advances in artificial intelligence, psychology, and linguistics* (pp. 365-376). Dordrecht: Kluwer.

Dell, G. S. (1986). A spreading activation theory of retrieval in sentence production. *Psychological Review, 93*, 283-321.

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

Efron, D. (1972). *Gesture, race and culture.* The Hague: Mouton.

Feyereisen, P. (1997). The competition between gesture and speech production in dual-task paradigms. *Journal of Memory and Language, 36*(1), 13-33.

Feyereisen, P., & Lannoy, J.-D. (1991). *Gestures and speech: Psychological investigations.* Cambridge: Cambridge University Press.

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

Goldman-Eisler, G. (1968). *Psycholinguistics: Experiments in spontaneous speech.* London: Academic Press.

Gerwing, J., & Bavelas, J. (2005). Linguistic influences on gesture's form. *Gesture, 4*(2), 157–195.

Graham, J. A., & Heywood, S. (1975). The effects of elimination of hand gestures and of verbal codability on speech performance. *European Journal of Social Psychology, 5*(2), 189-195.

Gullberg, M. (1998). *Gesture as a communication strategy in second language discourse. A study of learners of French and Swedish.* Lund: Lund University Press.

Hadar, U., & Butterworth, B. (1997). Iconic gestures, imagery, and word retrieval in speech. *Semiotica, 115*(1/2), 147-172.

Hadar, U., Wenkert-Olenik, D., & Soroker, N. (1996). Gesture and the processing of speech in aphasia. *Brain and Language, 55*(1), 141-187.

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

Hieke, A., E. (1981). A content processing view of hesitation phenomena. *Language and Speech, 24*(2), 147-160.

Indefrey, P., & Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition, 92*, 101-144.

Jefferson, G. (1974). Error correction as an interactional resource. *Language in Society, 2*, 181-199.

Kelly, S. D., Barr, D. J., Breckinridge Church, R., & Lynch, K. (1999). Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of Memory and Language, 40*(4), 577-592.

Kempen, G., & Hoenkamp, E. (1987). An incremental procedural grammar for sentence formulation. *Cognitive Science, 11*, 201-258.

Kendon, A. (1972). Some relationships between body motion and speech: An analysis of an example. In A. W. Siegman & B. Pope (Eds.), *Studies in dyadic communication* (pp. 177-210). New York: Pergamon.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In M. R. Key (Ed.), *The relationship of verbal and nonverbal communication* (pp. 207-227). The Hague: Mouton.

Kendon, A. (1983). Gesture and speech. How they interact. In J. M. Wiemann & R. P. Harrison (Eds.), *Nonverbal interaction* (pp. 13-45). Beverly Hills: Sage Publications.

Kendon, A. (1993). Human gesture. In K. R. Gibson & T. Ingold (Eds.), *Tools, language and cognition in human evolution* (pp. 43-62). Cambridge: Cambridge University Press.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press.

Kita, S. (1993). *Language and thought interface: A study of spontaneous gestures and Japanese mimetics*. Unpublished Dissertation, University of Chicago, Chicago.

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal? Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language, 48*, 16-32.

Kita, S., van Gijn, I., & van der Hulst, H. (1998). Movement phases in signs and co-speech gestures, and their transcription by human coders. In I. Wachsmuth & M. Fröhlich (Eds.), Gesture and Sign Language in Human-Computer Interaction, International Gesture Workshop, Bielefeld, Germany, September 17-19, 1997, Proceedings. Lecture Notes in Artificial Intelligence, Volume 1371 (pp.23-35) Berlin: Springer-Verlag.

Klapp, S. T., Anderson, W. G., & Berrian, R. W. (1973). Implicit speech in reading revisited, reconsidered. *Journal of Experimental Psychology, 100*, 368-374.

Klapp, S. T., & Erwin, C. I. (1976). Relation between programming time and the duration of the responses being programmed. *Journal of Experimental Psychology: Human Perception and Performance, 2*, 591-598.

Kormos, J. (2000). The timing of self-repairs in second language speech production. *Studies in Second Language Acquisition, 22*(2), 145-167.

Krauss, R. M., Chen, Y., & Chawla, P. (1996). Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? *Advances in Experimental Social Psychology, 28*, 389-450.

Krauss, R. M., Chen, Y., & Gottesman, R. F. (2000). Lexical gestures and lexical access: a process model. In D. McNeill (Ed.), *Language and gesture* (pp. 261-283). Cambridge: Cambridge University Press.

Krauss, R. M., & Hadar, U. (1999). The role of speech-related arm/hand gestures in word retrieval. In L. S. Messing & R. Campbell (Eds.), *Gesture, speech, and sign* (pp. 93-116). New York: Oxford University Press.

Lackner, J. R., & Tuller, B. H. (1979). Role of efference monitoring in the detection of self-produced speech errors. In W. E. Cooper & E. C. T. Walker (Eds.), *Sentence processing* (pp. 281-294). Hillsdale, N. J.: Lawrence Erlbaum.

Ladefoged, P., Silverstein, R., & Papcun, G. (1973). Interruptibility of speech. *Journal of the Acoustic Society of America, 54*, 1105-1108.

Laver, J. D. M. (1980). Monitoring systems in the neurolinguistic control of speech production. In V. A. Fromkin (Ed.), *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand* (pp. 287-305). New York: Academic Press.

Levelt, W. J. M. (1981). The speaker's linearization problem. *Philosophical Transactions of the Royal Society,* London, *B 295*, 305-315.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition, 14*, 41-104.

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

Levelt, W. J. M. (1996). Perspective taking and ellipsis in spatial descriptions. In P. Bloom & M. A. Peterson & L. Nadel & M. F. Garrett (Eds.), *Language and space* (pp. 77-107). Cambridge, MA: MIT Press.

Levelt, W. J. M., Richardson, G., & La Heij, W. (1985). Pointing and voicing in deictic expressions. *Journal of Memory and Language, 24*(2), 133-164.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*, 1-37.

Lickley, R. J. (1994). *Detecting disfluencies in speech.* Unpublished Dissertation, University of Edinburgh, Edinburgh.

Linde, C., & Labov, W. (1975). A site for the study of language and thought. *Language, 51*(4), 924-939.

Logan, G. D. (1982). On the ability to inhibit complex movements. A stop-signal study of type writing. *Journal of Experimental Psychology: Human Perception and Performance, 8*, 778-792.

Logan, G. D., & Cowan, W. B. (1984). On the ability to inhibit thought and action: A theory of an act of control. *Psychological Review, 9*(3), 295-327.

MacKay, D. G. (1982). The problems of flexibilty, fluency and speed-accuracy trade-off in skilled behavior. *Psychological Review, 89*(5), 483-506.

MacKay, D. G. (1987). *The organization of perception and action: A theory for language and other cognitive skills*. New York: Springer-Verlag.

MacKay, D. G. (1990). Perception, action, and awareness: A three-body problem. In O. Neumann & W. Prinz (Eds.), *Relationships between perception and action*. Berlin: Springer.

MacKay, D. G. (1992a). Awareness and error detection: new theories and research paradigms. *Consciousness and Cognition, 1*, 199-225.

MacKay, D. G. (1992b). Errors, ambiguity, and awareness in language perception and production. In B. J. Baars (Ed.), *Experimental slips and human error: exploring the architecture of volition* (pp. 39-69). New York: Plenum Press.

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

Marslen-Wilson, W., & Tyler, L. (1981). Central processes in speech understanding. *Philosophical Transactions of the Royal Society in London, B295*, 317-332.

Mayberry, R., Jaques, J., & DeDe, G. (1998). What stuttering reveals about the development of the gesture-speech relationship. *New Directions for Child Development, 79*, 77-87.

Mayberry, R. I., & Jaques, J. (2000). Gesture production during stuttered speech: insights into the nature of gesture-speech integration. In D. McNeill (Ed.), *Language and gesture* (pp. 199-214). Cambridge: Cambridge University Press.

McGarry, T., & Franks, I. M. (1997). A horse race between independent processes: Evidence for a phantom point of no return in the preparation of a speeded motor response. *Journal of Experimental Psychology: Human Perception and Performance, 23*(5), 1533-1542.

McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review, 92*(3), 271-295.

McNeill, D. (1992). *Hand and mind. What the hands reveal about thought*. Chicago: Chicago University Press.

McNeill, D., & Duncan, S. D. (2000). Growth points in thinking-for-speaking. In D. McNeill (Ed.), *Language and gesture* (pp. 141-161). Cambridge: Cambridge University Press.

Melinger, A., & Kita, S. (2001). *Does gesture help processes of speech production? Evidence for conceptual level facilitation.* Proceedings of the Berkeley Linguistics Society, USA, 27.

Melinger, A., & Levelt, W. J. M. (2004). Gesture and the communicative intention of the speaker. *Gesture, 4*(2), 119–141.

Meringer, R. (1908). *Aus dem Leben der Sprache.* Berlin: Behr.

Meyer, A. S., Sleiderdink, A. M., & Levelt, W. J. M. (1998). Viewing and naming objects: eye movements during noun phrase production. *Cognition, 66*, B25-B33.

Morrel-Samuels, P., & Krauss, R. M. (1992). Word familiarity predicts the temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory and Cognition, 18*(3), 615-623.

Motley, M. T., Camden, C. T., & Baars, B. J. (1982). Covert formulation and editing of anomalies in speech production: Evidence from experimentally elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior, 21*, 578-594.

Nakatani, C. H., & Hirschberg, J. (1994). A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustic Society of America, 95*(3), 1603-1616.

Nobe, S. (2000). Where do most spontaneous representational gestures occur with respect to speech? In D. McNeill (Ed.), *Language and Gesture.* Cambridge: Cambridge University Press.

Nooteboom, S. G. (1980). Speaking and unspeaking: Detection and correction of phonological and lexical errors in spontaneous speech. In V. A. Fromkin (Ed.), *Errors in linguistic performance* (pp. 87-95). New York: Academic Press.

Oomen, C. C. E. (2001). *Self-monitoring in normal and aphasic speech.* Unpublished Dissertation, University of Utrecht, Utrecht.

Oomen, C. C. E., & Postma, A. (2001). Effects of time pressure on mechanisms of speech production and self-monitoring. *Journal of Psycholinguistic Research, 30*(2), 163-184.

Oomen, C. C. E., & Postma, A. (2002). Limitations in processing resources and speech monitoring. *Language and Cognitive processes, 17*(2), 163-184.

Özyürek, A. (2002). Do speakers design their co-speech gestures for their addressees? The effects of addressee location on representational gestures. *Journal of Memory and Language, 46*, 688-704.

Pedelty, L. (1987). *Gesture in aphasia.* Unpublished Dissertation, University of Chicago, Chicago.

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

Postma, A., & Kolk, H. H. J. (1992). The effects of noise masking and required accuracy on speech errors, disfluencies, and self-repairs. *Journal of Speech and Hearing Research, 35*, 472-487.

Postma, A., & Kolk, H. H. J. (1993). The covert repair hypothesis: prearticulatory repair processes in normal and stuttered speech. *Journal of Speech and Hearing Research, 36*, 472-487.

Postma, A., & Noordanus, C. (1996). Production and detection of speech errors in silent, mouthed, noise-masked, and normal auditory feedback speech. *Language and Speech, 39*(4), 375-392.

Ragsdale, J. D., & Silvia, C. F. (1982). Distribution of kinesic hesitation phenomena in spontaneous speech. *Language and Speech, 25*(2), 185-190.

Rauscher, F., H.,, Krauss, R., M., , & Chen, Y. (1996). Gesture, speech, and lexical access: The role of lexical movements in speech production. *Psychological Science, 7*(4), 226-231.

Sacks, H. (1992). *Lectures on conversation.* Oxford: Blackwell.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*(4), 696-735.

Schegloff, E. A. (1979). The relevance of repair to syntax-for-conversation. In T. Givon (Ed.), *Syntax and semantics* (Vol. 12, pp. 261-286). New York: Academic Press.

Schegloff, E. A. (1984). On some gestures' relation to talk. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action* (pp. 266-296). Cambridge: Cambridge University Press.

Schegloff, E. A. (1992). Repair after next turn: The last structurally provided defense of intersubjectivity in conversation. *American Journal of Sociology, 97*(5), 1295-1345.

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language, 53*(2), 361-382.

Seyfeddinipur, M. (2002). Repair at hand: Monitoring for errors in gesture. Paper presented at the first conference of the International Society of Gesture Studies: Gesture: The Living Medium, Austin, TX, June 5-8, 2002.

Shriberg, E. E. (1994). *Preliminaries to a theory of speech disfluencies.* Unpublished Dissertation, University of California at Berkeley, Berkeley.

Smith, V. L., & Clark, H. H. (1993). On the course of answering questions. *Journal of Memory and Language, 32*, 25-38.

Svartvik, J., & Quirk, R. (1980). *A corpus of English conversation.* Lund, Sweden: Gleerup.

Tyler, L., & Marslen-Wilson, W. (1986). The effects of context on the recognition of polymorphemic words. *Journal of Memory and Language, 25*, 741-752.

Ulmer-Ehrich, V. (1982). The structure of living space descriptions. In R. J. Jarvella & W. Klein (Eds.), *Speech, place, and action* (pp. 219-249). New York: Wiley.

Van Hest, E. (1996). *Self-repair in L1 and L2 production.* Tilburg: Tilburg University Press.

Van Wijk, C., & Kempen, G. (1987). A dual system for producing self-repairs in spontaneous speech: Evidence from experimentally elicited corrections. *Cognitive Psychology, 19*, 403-440.

Wheeldon, L. R., Levelt, W. J. M. (1995). Monitoring the time course of phonological encoding. *Journal of Memory and Language, 34*, 311-334.

Curriculum Vitae

Mandana Seyfeddinipur (1967) studied Linguistics, German Studies, Persian Studies and German as a Foreign Language at the Freie Universität Berlin, Germany. She was granted a PhD stipend from the Max Planck Gesellschaft and joined the Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands. She worked in the Gesture Project within the Language and Cognition group. She has been granted a Marie Curie Stipend to pursue postdoctoral studies at Stanford University and at MPI, Nijmegen.

MPI Series in Psycholinguistics

1. The electrophysiology of speaking: Investigations on the time course of semantic, syntactic, and phonological processing
Miranda van Turennout

2. The role of the syllable in speech production: Evidence from lexical statistics, metalinguistics, masked priming, and electromagnetic midsaggital articulography
Niels O. Schiller

3. Lexical access in the production of ellipsis and pronouns
Bernadette M. Schmitt

4. The open-/closed-class distinction in spoken-word recognition
Alette Haveman

5. The acquisition of phonetic categories in young infants: A self-organising artificial neural network approach
Kay Behnke

6. Gesture and speech production
Jan-Peter de Ruiter

7. Comparative intonational phonology: English and German
Esther Grabe

8. Finiteness in adult and child German
Ingeborg Lasser

9. Language input for word discovery
Joost van de Weijer

10. Inherent complement verbs revisited: Towards an understanding of argument structure in Ewe
James Essegbey

11. Producing past and plural inflections
Dirk Janssen

12. Valence and transitivity in Saliba: An Oceanic language of Papua New Guinea
Anna Margetts

13. From speech to words
Arie van der Lugt

14. Simple and complex verbs in Jaminjung: A study of event categorisation in an Australian language
Eva Schultze-Berndt

15. Interpreting indefinites: An experimental study of children's language comprehension
Irene Krämer

16. Language-specific listening: The case of phonetic sequences
Andrea Weber

17. Moving eyes and naming objects
Femke van der Meulen

18. Analogy in morphology: The selection of linking elements in Dutch compounds
Andrea Krott

19. Morphology in speech comprehension
Kerstin Mauth

20. Morphological families in the mental lexicon
Nivja H. de Jong

21. Fixed expressions and the production of idioms
Simone A. Sprenger

22. The grammatical coding of postural semantics in Goemai (a West Chadic language of Nigeria)
Birgit Hellwig

23. Paradigmatic structures in morphological processing: Computational and cross-linguistic experimental studies
Fermín Moscoso del Prado Martín

24. Contextual influences on spoken-word processing: An electrophysiological approach
Daniëlle van den Brink

25. Perceptual relevance of prevoicing in Dutch
Petra M. van Alphen

26. Syllables in speech production: Effects of syllable preparation and syllable frequency
Joana Cholin

27. Producing complex spoken numerals for time and space
Marjolein Meeuwissen

28. Morphology in auditory lexical processing: Sensitivity to fine phonetic detail and insensitivity to suffix reduction
Rachèl J. J. K. Kemps

29. At the same time...: The expression of simultaneity in learner varieties
Barbara Schmiedtová

30. A grammar of Jalonke argument structure
Friederike Lüpke

31. Agrammatic comprehension: An electrophysiological approach
Marlies Wassenaar

32. The structure and use of shape-based noun classes in Miraña (North West Amazon)
Frank Seifart

33. Prosodically-conditioned detail in the recognition of spoken words
Anne Pier Salverda

34. Phonetic and lexical processing in a second language.
Miriam Broersma

35. Retrieving semantic and syntactic properties: ERP studies on the time course in language comprehension.
Oliver Müller

36. Lexically-guided perceptual learning in speech processing.
Frank Eisner

37. Sensitivity to detailed acoustic information in word recognition.
Keren Shatzman

38. The acquisition of phonological structure: Distinguishing contrastive from non-contrastive variation.
Christiane Dietrich

39. Disfluency: Interrupting speech and gesture.
Mandana Seyfeddinipur