Production and Perception of Pauses in Speech

Kristina Lundholm Fors

UNIVERSITY OF GOTHENBURG

Production and Perception of Pauses in Speech

Production and Perception of Pauses in Speech

Kristina Lundholm Fors

UNIVERSITY OF GOTHENBURG
PHILOSOPHY, LINGUISTICS & THEORY OF SCIENCE

For my father, Gunnar Lundholm (1941-2004)

Doctoral dissertation in linguistics, University of Gothenburg
August 17, 2015

Disputationsupplaga

Printed by Ale Tryckteam, Bohus, 2015

Distribution:
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
Box 200, SE-405 30 Gothenburg, Sweden

Abstract

Ph.D dissertation at University of Gothenburg, Sweden, 2015
Title: *Production and Perception of Pauses in Speech*
Author: Kristina Lundholm Fors
Language: English
Department: Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, Box 200, SE-405 30 Göteborg
Year: 2015

Silences can make or break the conversation: if two persons involved in a conversation have different ideas about the typical length of pauses, they will face problems with turn taking. Pauses occur in conversation for a number of reasons, for example for breathing, thinking, word-searching and turn taking management. In this dissertation, we explore the production and perception of pauses in speech. Our aim consists of three main parts: to describe and analyse the production of pauses, to investigate the perception of pauses, and to examine the role of pauses in turn-taking. Our hypothesis is that pauses fill varying functions, and that these functions depend on the context of the pauses. We believe that the duration of pauses may be linked to the pause type, and that we adapt the our pause lengths to the persons we are speaking to. Further, we suggest that pauses occur regularly throughout dialogues. We also hypothesise that the duration of pauses in speech affect the processing of speech.

Pauses are tied to the process of turn taking, and as we learn more about the nature of pauses we may also be able to further develop our understanding of the process of turn holding and turn yielding. We will also be able to use the information about pause production and perception when modelling turn taking in dialogue systems.

Our results show that pause lengths vary greatly across speakers, pause types and dialogues. Pauses tend to be entrained by speakers involved in dialogues, and pauses occur regularly throughout conversations. We also found evidence that pauses have a positive impact on memorising spoken utterances. While speakers adapt their pause lengths to the other speaker in the conversation, they are inclined to keep a consistent ratio between pause types, and this is not dependent on the conversational partner. While it is interesting to look at pauses separately, we need to put them into context to really understand their functions. To highlight the role of pauses in conversation, we proposed an updated turn taking model, where the results from our studies are integrated.

Keywords: pauses, silences, turn taking, dialogue, entrainment

Acknowledgments

There are many people whose support and encouragement have been invaluable to me during my time as a PhD student and while I have been writing my dissertation. First and foremost, I would like to thank my supervisor Staffan Larsson, who has been nothing but brilliant. His patience and enthusiasm have been consistent throughout the years, and he has encouraged me at every step. My second supervisor Mattias Heldner has also been very helpful in responding to my various questions about pauses and phonetics. All errors and mistakes in the dissertation are of course my own.

I would like to thank my fellow PhD students at FLoV (ingen nämnd, ingen glömd!), and the rest of the research staff, especially the my linguistics colleagues, who have been supportive and kind. Thank you also to the administrative staff. In the SIMSI project I had the pleasure of working with Ellen, Jessica, Simon, Stina and Chris, and I enjoyed that very much. I've had the opportunity to teach a lot during my PhD studies, and I am very grateful to all the students who have helped me learn how to explain things in new ways.

Without the people who have allowed themselves to be recorded and participated in my experiments, there would be no dissertation, and therefore I am in their debt.

I would also like to thank Kungliga och Hvitfeldtska stiftelsen for the scholarship I received.

When you get stuck on something, there is always great relief to get help from someone knowledgable in that area. Robert

Adesam and Sebastian Berlin have helped me with technical and programming issues. Marcus Nyström has been a great resource in the area of eye tracking. Andreas Jakobsson pointed me towards the Bølviken test, which I would never have found otherwise. Chris Howes patiently answered my statistics questions, and additionally served as opponent at my final seminar, and did a fantastic job!

My friends and family have been supportive throughout. Thank you Karin Ö, Sofia and Elsa for always being there. Thank you Mimmi, Amy, and my sister Maria. Thank you Karin C for the walks and the constant encouragement! I am lucky in that I have the loveliest in-laws, and I would especially like to thank my parents-in-law Mats and Yvonne for their love and kindness. My mother Ulla has done whatever she has been able to do to help me, and for that I am grateful. My father Gunnar, to whom I have dedicated this dissertation, has been gone now for over ten years. Still, I do not believe I would have pursued a doctorate if it had not been for him. His eagerness to learn new things and his enjoyment of his work as a professor is what awakened my passion for research. His love and kindness are forever with me.

Finally: Anders, my husband. Thank you, my love, from the bottom of my heart.

Contents

Chapter 1

Introduction

The right word may be effective, but no word was ever as effective as a rightly timed pause.

Mark Twain

Silences can make or break the conversation: if two persons involved in a conversation have different ideas about the typical length of pauses, they will face problems with turn taking. One person might feel that the silences are long and awkward, while the other person might feel that there is never a silence long enough for them to take the turn. Silences occur in conversation for a number of reasons, for example for breathing, thinking, word-searching and turn taking management. Of course, one may also be silent because the other person is talking. Before we go any further, we need to establish what we mean by silences, and what we mean by pauses. A person can be silent in all sorts of situations, for example when sitting on the bus, when cooking food, or when listening to someone else speaking. Silence can be defined as "complete absence of sound" or "the fact or state of abstaining from speech" [1], whereas a pause is defined as "a tem-

[1] New Oxford American Dictionary, Third edition

porary stop in action or speech" [2]. We define a pause as follows: *a pause is a silence that occurs during an ongoing conversation, and during a speaker's turn or at a turn change*. From this follows that when a person in a conversation is silent because the other person is speaking, we do not call that silence a pause. We will expand on the concept of pauses in section 2.2.1.

Example 1below is an excerpt from the material used in this dissertation (the material will be described in section 3.3.1). Pauses are annotated on separate lines by the length of the pause in milliseconds within parenthesis. Each line with transcribed speech begins with the name of the speaker. Translation into English is added where relevant.

(1)

```
01 Cilla: bli du stressad liksom
         do you get like stressed
02 (223 ms)
03 Cilla: såna saker
          things like that
04 (821 ms)
05 Cilla: men- men hur- hur
          but- but how- how
06 (127 ms)
07 Cilla: hur ska de få honom å inse
          how is that going to make him realize
```

In the example above there are three pauses: one on line 2, one on line 4 and one on line 6. They differ with regards to duration, but they are similar in that they all occur within a speaker's turn; the person that speaks before the pause also speaks after the pause. At first the difference between the pause on line 2 and on line 6 may seem small, but the 223 ms pause on line 2 is almost double the length of the 127 ms pause on line 6. What, then,

[2] New Oxford American Dictionary, Third edition

could be the reason that they differ in duration? If we look closer at what is being said, we see that the very short pause on line 06 is preceded by words that suggest the speaker intends to continue, whereas the two longer pauses seem to follow utterances that are more complete. This means that at line 2 and 4, the turn could have shifted to another speaker. That would not be as likely to happen at line 6. If we hypothesise that the duration of the pause is significant when it comes to turn taking, we might be able to differentiate different types of pauses by looking at their duration, and this is one of the matters that we will investigate in this thesis.

If the duration of the pause is important, this gives rise to further questions: for example, how do people agree on what is a short pause and what is a long pause? Do we adapt our pause lengths when speaking to different people? Another question related to turn taking is how often we pause, and if pauses occur regularly over the course of dialogues.

Pauses are tied to the process of turn taking, and as we learn more about the nature of pauses we may also be able to further develop our understanding of the process of turn holding and turn yielding. We will also be able to use the information about pause production and perception when creating computer software that enables computers to communicate in a more human-like way (such systems are often referred to as dialogue systems). Knowledge about how pauses are realised in spontaneous interaction can also provide a basis for comparison when evaluating persons that have speech and language impairments.

1.1 Aim

The aim of this dissertation is to pin down the silences in conversation, and have a long, hard look at them. What are their characteristics? Where are they most often found? What happens to the surrounding speech when a pause is introduced? Are there different types of pauses, or do they all look and behave the same?

And what happens when the pauses misbehave, becoming too long?

The aim can be divided into three main parts: to describe and analyse the production of pauses, to investigate the perception of pauses, and to examine the role of pauses in turn-taking. Our hypothesis is that pauses fill varying functions, and that these functions depend on the context of the pauses. We believe that the duration of pauses may be linked to the pause type, and also that we adapt the duration of our pauses to the persons we are speaking to. Further, we suggest that pauses occur regularly throughout dialogues. We also hypothesise that the duration of pauses in speech affect the processing of speech.

1.2 Structure

Below is an outline of the thesis, with short descriptions of each chapter.

- **Chapter 2: Talk and silence.** In this chapter, previous research on conversations and pauses is presented and discussed. Key concepts such as pauses and turns are defined.

- **Chapter 3: Methodology and material.** In this chapter methodological issues are presented and discussed, and the material that the studies are based on is introduced.

- **Chapter 4: Production of pauses.** Pauses have several characteristics, such as duration, place and context, and the four studies presented in this chapter explore these different aspects of pauses. The first two studies are concerned with length of pauses; how long pauses are and if we adapt our pause lengths to the person we are involved in conversation with. In the third study, the focus is on the distribution of pauses over the course of dialogues, and if they occur with a certain regularity. The fourth and final study concentrates on the syntactic context of pauses, and the goal is to find out how the context differs between different types of pauses.

- **Chapter 5: Perception of pauses.** The length of pauses influences both how pauses are perceived and also how the surrounding speech is perceived. In this chapter we present a study on how pauses of different length influence the processing of spoken language and the ability to remember spoken sentences.

- **Chapter 6: An updated turn taking model.** Based on the studies in previous chapter, we propose an update to the prevailing model of turn taking. Primarily, we suggest that places for turn taking should viewed on a continuum rather than as binary.

- **Chapter 7: Summary and future work.** In this chapter, the findings of the studies are discussed in relation to previous research, and areas of future research are highlighted.

Chapter 2

Talk and silence

> *Silence remains, inescapably, a form of speech (. . .) and an element in a dialogue.*
>
> Susan Sontag

The backdrop to pauses in speech is speech itself. Speech has been studied from differing viewpoints over time, and with different purposes. Speech can be read or spontaneous, and in monologue or in conversation with others. These are however not dichotomies, but rather continuums: For example, would you place an actor's lines in a play into the "read" or "spontaneous" category? How about someone retelling a story? Is a lecture a monologue, or is it a conversation where one speaker has the majority of the speaker time?

We need to be aware of the fact that the context and the conditions will shape the conversation, and that this will lead to different types of conversations even if the same people are involved. A conversation can involve two or more people. In this dissertation the focus is on spontaneous two-party conversations, which will be referred to as dialogues[1].

[1] The word "dialogue" is based on the Greek "dialogos", stemming from "dia" (through) and "logos" (speech, reason). Hence, there is nothing in the word dialogue that suggests that it can only be used for two-party conversa-

2.1 Studying conversations

There are several problems that need to be resolved when doing research on pauses in conversations. The most basic problem is defining what a pause is, and what it is not. We can look at pauses from different perspectives and this might give us conflicting ideas about what a pause is. For example: should pauses be defined acoustically, perceptually, or should a combination of the two be used? This will be further discussed in section 3.1. First, we will put the study of conversation in a historical context, and have a look at where it started.

Speech and conversation have been studied for a long time: rhetoric, for example, can be traced back to ancient Greece. During the later part of the 19th century, the interest in local history and dialects led to research into spoken language. With it arose a need to be able to write down, or transcribe, not only what was being said, but how it was said. In Sweden, this led to the creation of "Landsmålsalfabetet", a phonetic alphabet which can be used to transcribe Swedish and Swedish dialects (Norrby, 2004). The International Phonetic Alphabet was also created in the late 19th century, and is now used worldwide to transcribe spoken language. [2] When studying spoken language, speech is routinely transcribed and subsequently turned into written language. Through this conversion, we risk losing significant information. It is therefore vital that transcriptions of spoken language follow agreed upon conventions as to what is to be transcribed, and how.

The study of spoken language lead to new research paradigms, such as Conversation Analysis (commonly abbreviated as CA). Conversation analysis grew out of sociology, and more specifically ethnomethodology. CA was developed in the 1960s by Harvey Sacks, Emanuel Schegloff and Gail Jefferson. The then quite recent easy access to audio recording equipment

tions, even if we use it to mean two-party conversations in this dissertation.

[2] However, the International Phonetic Alphabet was originally developed as a pedagogical tool.

was vital to the emergence of the field (ten Have, 2007). In CA, the material is to be approached without preconceived notions and ideas. Instead of working from a hypothesis and trying to prove it, CA encourages the researcher to look for repeating patterns in the material and to build conclusions on that (Norrby, 2004). In Conversation Analysis, prosody is seen as an integral part of language. CA is used in many different fields, such as sociology, anthropology and linguistics. The study of talk in interaction within the field of linguistics is sometimes known as "interactional linguistics".

In 1974, Sacks et al. published a key paper on the systematics of turn taking, suggesting that speaker changes follow an orderly structure which can be described by a set of rules. Different types of pauses can be derived from these rules, and this will be discussed in more detail in Section 2.2.1.

How speakers adapt to each other has been of great interest to speech researchers during the last two decades. This is referred to as entrainment, alignment, mirroring, accommodation or coordination. Used in the context of conversation it means that two speakers become more similar when speaking to each other. Entrainment has been found to manifest itself in many ways in language, from the words we use to the pitch of our voices. It is believed that entrainment is present at all levels of communication, and that this process helps us understand each other (Pickering and Garrod, 2004). It may help us with turn taking and coordination in conversations. In section 4.4 we will explore entrainment in pause duration.

2.2 Pauses in conversations

It has been suggested by Linell (2004) that a bias towards written language has been pervasive in linguistics. Linell argues that methods and models used to study spoken language are based on the methods used to standardise and explore written language. The idea of the ideal delivery can be traced back to the written language bias. Linell describes the written language bias as follows:

> "An 'ideal delivery' of an utterance or text is free from (unplanned) pauses (filled or unfilled), pleonasms[3], restarts, structure shifts, and other errors. When disfluency problems occur in speech, they are not pertinent to language *per se*." Linell (2004, p. 105)

Because of the written language bias, pauses and disfluencies were seen as noise in the signal. Traces of the written language bias can be found in studies on speech. For example, when analyzing the predictability of words in speech, Goldman-Eisler (1958) excluded sentences that were not "grammatically correct and well constructed". In this dissertation, we base our work on the underlying assumption that speech and writing are two different expressions of language, and that pauses and disfluencies are an intrinsic part of spoken language.

When we speak, we inevitably produce pauses — we cannot speak without pausing (Zellner, 1994). The simplest way to try to explain the reason for pauses would be to suggest that we pause to inhale, speak for as long as our lung capacity allows, and then pause to inhale again. Pausing to breathe is a physiological necessity, but we also pause due to cognitive needs. A study by Howell and Sackin (2001) further underlines the cognitive aspect of pauses by demonstrating that when speakers are conditioned to avoid silent pauses, they instead increase function word repetitions. Based on this, we can argue that we not only pause to breathe, but we pause to gain time to for example plan what we are going to say (and when we can't use pauses, we use other strategies to get planning time, such as repeating function words). Goldman-Eisler was one of the first researchers focusing on pauses, and especially the cognitive functions of pauses,

[3] The term pleonasm refers to the use of more words or parts of words than is necessary for clear expression.

and carried out a number of studies investigating this. The findings that pauses correlate with an increase in information and are more common in conjunction with less frequent words further highlight the connection between speech planning and pausing (Goldman-Eisler, 1958). Pause durations were also observed to vary depending on the situation and the speaker (Goldman-Eisler, 1961).

So, we have pauses for breathing and for planning what to say; we speak until we run out of air, or run out of planned things to say. Again, if we ponder the issue in the simplest way, we would propose that a pause occurs at the very moment the speaker is not able to produce another syllable, either because more air is needed, or because she has nothing at all left to say. This simplistic view is not supported by empirical data: the placement of pauses seems to be more complex. It seems that we do not pause haphazardly, but rather we plan where to pause, following certain constraints such as speech rhythm (Szczepek Reed, 2010).

Why do we need to plan the placement of our pauses? Why do we not pause at the moment we feel the need to do so? We have to remember the basic use of speech, which is to communicate with others. With that taken into consideration, we find an additional reason for pauses: we pause, to allow the person we are speaking with take the turn, if they wish. So, sometimes we pause to check whether someone else wants to speak, sometimes we pause to plan, and sometimes we pause to breathe. Most likely these acts are not mutually exclusive, but rather we may for example pause both to breathe and to plan what we are going to say.

Somehow we need to prevent misunderstandings, such as the person we are speaking to interpreting our planning pauses as checking whether they want to speak, and this is part of the reason why we do not just pause anywhere. An overview of the different types of pauses and their placements follows in section 2.2.1.

Pauses are also constrained with regards to length. When we are speaking to someone, we most often mutually agree to follow some conversational rules. One of these (unspoken) rules tells us that very long pauses should be avoided, if they can not be explained by some activity or a spoken preface (Levinson, 1983; Newman, 1982). It is quite alright to be quiet for a long time if the person I'm speaking to can see that I'm looking for something that I want to show them in a book, or if I tell them I need to think about something before I give my answer. But if in the middle of speaking to someone, I fall silent and don't speak for more than 3 seconds or so, the person I'm speaking to will wonder what has caused my silence, and might interpret this as if the communication has broken down. Therefore, we limit the duration of our pauses. To complicate matters, duration may depend on the context: some locations allow for longer pauses, whereas in other positions shorter pauses are more common. We also have to take into account individual differences, and how people affect each other when speaking. A more in depth exploration of the different constraints on pause durations is given in section 2.2.2.

2.2.1 Pause types

Pauses occur in specific contexts, and by context we do not only refer to the immediate words surrounding the silence, but also the larger linguistic and cultural context. Further, pauses are not one homogenous group, but can be divided into different subgroups.

In Conversation Analysis, conversations are divided into turns. A turn is everything a speaker says from when she takes the floor until another speaker takes over. Turns consist of turn-constructional units (TCUs), delimited by transition-relevance places. A TCU can consist of a clause, a phrase or a single word, and the definition of aTCU is that it may constitute a turn, which means that a turn change can take place after a turn-constructional unit is completed. A transition relevance place (often referred to as TRP) occurs when a turn-constructional unit has been completed, and it marks a place in the conversation where the turn may be transferred to another speaker. The TRP does not necessitate a turn change, but indicates the possibility of such an event.

Sacks et al. (1974) separate silences in conversation into gaps, lapses and pauses. The categorisation of silence is dependent on its place and context, and this is governed by turn-taking rules. As a speaker reaches a transition-relevance place (TRP), the following rules apply:

- if the current speaker has nominated another speaker, speaker change takes place (rule 1a)
- if the current speaker has not nominated another speaker, any participant in the conversation may take the turn, and speaker change can take place (rule 1b)
- if the current speaker has not nominated another speaker, and no other person has self-nominated, the current speaker may continue (rule 1c)
- these rules apply at every transition-relevance place (rule 2)

This system is presented as a simplest rule system capable of describing turn taking in conversations. The different types of pauses described by Sacks et al. are derived from the turn taking rules:

- a pause is a silence that occurs inside a speaker's turn. This includes the silence at a TRP, when a speaker has been nominated but has not begun to speak. It also includes the silence at a TRP, when a speaker has stopped, but following rule 1c then continues to speak after the TRP.
- a gap is the silence that occurs at a TRP when the first speaker has not nominated another speaker, but another speaker self-nominates and there is a turn change

- a lapse is the silence at a TRP, when the first speaker has stopped speaking, has not nominated a new speaker, and does not continue speaking. No other speaker takes the turn. A lapse is in part defined by the perceived length: thus, a lapse should be perceived as longer than a gap and as a discontinuity in the flow of conversation.

In this dissertation, pauses, gaps and lapses will all be referred to as different types of pauses, and an overview of pause types based on the categorisation above will be given in the sections below.

Pauses at transition relevance places

Two types of pauses can occur at transition relevance places. The first type of pause that occurs at transition relevance places is a result of rule 1b (described in 2.2.1): a speaker has reached a TRP, and another speaker self-nominates and takes the turn. This results in the TRP coinciding with a turn change. Typically, this type of pause is longer than when a second speaker has been nominated (Lundholm, 2000). An example of pause at a TRP that coincides with a turn change can be seen in example 2: Anna pauses on line 4 without clearly signalling who is to speak next, and after a 468 ms pause, Cilla self-nominates and takes the turn. These pauses will be referred to as **TRP-pauses with turn change**. (Sacks et al. refer to these as "gaps".)

(2)

```
01 Anna: ha din intuition med dig
         have your intuition with you
02 Anna: hela tiden liksom
         all the time kind of
03 (649 ms)
04 Anna: på nå vis
         in some way
05 (468 ms)
```

```
06 Cilla: asså vi måste tänka rätt också
          like we have to think correctly too
07 Cilla: sådär vi måste hela tiden
          we have to all the time
```

The second type arises when a speaker reaches a TRP, pauses, and then, following rule 1c, continues to speak. During such a pause, the listener often provides the speaker with feedback to make it clear that she does not want to take the turn. It is not evident which speaker this pause belongs to. It ends up being incorporated into one speaker's turn, but at a TRP, it could be argued that all participants in the conversation are equally responsible for making sure that the flow of conversation is not disturbed by an uncomfortably long pause. Even in a supposedly equal conversation, the person who seems to take the most responsibility for keeping the conversation going is also the person that is perceived by external listeners as feeling bad when the conversation comes to a halt (Newman, 1982).

Example 3 shows an example of Anna pausing after a TRP, and then continuing to speak after the TRP (we see the same type of pause on line 2 in example 2). This type of pause will be referred to as **TRP-pause without turn change**. (Sacks et al. refer to these as "pauses".)

(3)

```
01 Anna: inte så länge de tror jag inte
         not for long I don't think
02 (1374 ms)
03 Anna: men frågan e ju
         but the question is
```

According to Sacks, lapses may occur at TRPs. A lapse is qualitatively different than a gap or a pause, as it is said to be longer and disruptive. The definition will then be highly dependent on the perception of the transcriber/analyst. In this dissertation, pauses and gaps are not differentiated from lapses.

It is important to remember that speaker change does not necessarily always take place at a TRP. It is also possible for a speaker to be interrupted while in the middle of an utterance, or that another participant in the conversation misinterprets a planning pause and thinks that the speaker is relinquishing her turn.

Pauses in places other than Transition Relevance Places

As Sacks et al. (1974) states, pauses may occur in other places than at transition relevance places, and when a pause occurs within a speaker's turn but not at a TRP this is often referred to as a planning pause or a hesitation pause. This type of pause belongs to the speaker that speaks before and after the pause. It is common that this type of pause occurs before a content word, or after a discourse marker such as "and" or "but" (van Donzel and Koopmans - van Beinum, 1996). During, or in conjunction with, this type of pause, the speaker may inhale, swallow, exhale etc (Zellner, 1994).

When pauses occur within a person's turn, they can still occur in different places. Some pause locations will correlate with breathing, but not all. It has been suggested that pause locations are planned, so that we don't cease speaking the moment we realize that we need time to plan, but rather we continue until we reach a possible pause location (Szczepek Reed, 2010). This would lead to pauses being more common in some contexts than others. In three word sentences (noun phrase - verb phrase - noun phrase) with varying focus, two thirds of the pauses were found at the boundary between the first noun phrase and the verb phrase (Strangert, 2003). Example 4 shows the speaker Cilla pausing in a location that is not a TRP, and continuing to speak after the pause. These pauses will be referred to as **planning pauses**. (Sacks et al. refer to these as "pauses".)

(4)

```
01 Cilla: som du sa liksom va- var
          like you said kind of whe- where
02 (400 ms)
03 Cilla: var går min gräns liksom för de
          where is my boundary sort of for what
04 Cilla:  ja har kompetens för
           I am qualified for
```

The final type of pause occurs when the speaker who speaks before the TRP, nominates another speaker, for example by asking the second speaker a question. The second speaker starts speaking after the TRP, and the pause belongs to the second speaker. It may sound odd to categorise these pauses as not occurring at a TRP, but we argue that since a speaker has been nominated, the turn has already transitioned to the second speaker at the start of the pause, and therefore the pause occurs after the TRP. It is of course also possible for the second speaker to start his turn without a pause: he/she might overlap slightly with the first speaker, or start speaking exactly as the first speaker finishes. However, the most common situation, is that a short pause of about 200 ms occurs at the turn change (Heldner and Edlund, 2010). In example 5 speaker A poses a question to speaker B, and speaker B answers after a short pause. This kind of pause will be referred to as **initial pause**. (Sacks et al. refer to these as "pauses".)

(5)

```
01 Anna: å hur länge kunde de va då
         and how long could that be then
02 (2250 ms)
03 Emma: SUCK mellan kanske en å fyra veckor
         SIGH between maybe one and four weeks
04 Emma: eller nåt sånt
         or something like that
```

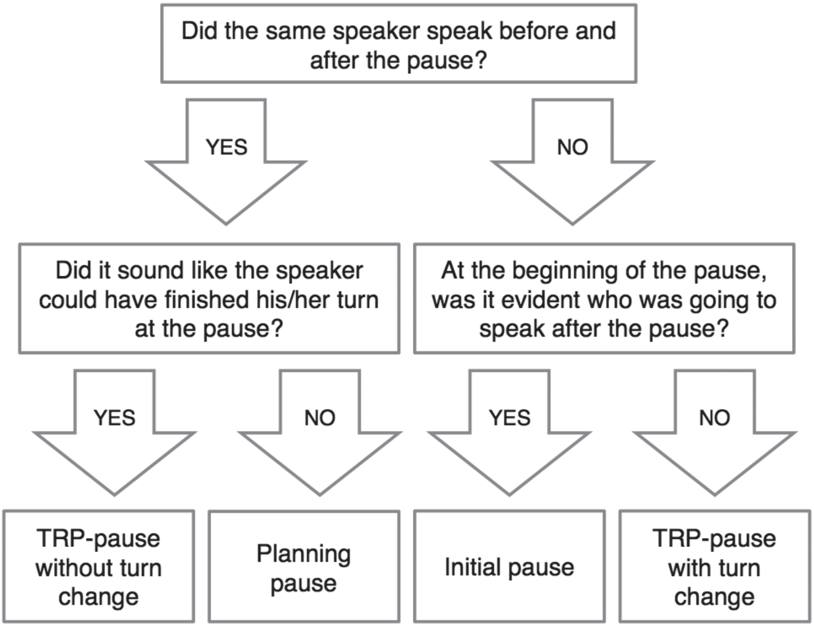
Did the same speaker speak before and after the pause?
YES
NO
Did it sound like the speaker could have finished his/her turn at the pause?
At the beginning of the pause, was it evident who was going to speak after the pause?
YES
NO
YES
NO
TRP-pause without turn change
Planning pause
Initial pause
TRP-pause with turn change

Figure 2.1: Categorisation of pauses in speech

Categorisation of pauses in speech

To simplify the categorisation of pauses in speech, we can use the flowchart in figure 2.1.

The flowchart will aid categorisation into the four pause categories described in Sections 2.2.1 and 2.2.1, by asking a set of questions. To begin with, we need to ascertain whether the speaker that spoke before the pause also spoke after the pause. If the answer is yes to that question, we need to judge if the speaker could have finished her turn at the beginning of the pause. If the same speaker spoke before and after the pause, and the turn could be regarded as complete by the beginning of the pause, the pause will be categorised as a TRP-pause without turn change. If it did not sound like the speaker that spoke before and after the pause could have finished his/her turn at the beginning of the pause, the pause will be categorised as a planning pause.

If we determine that the speaker that spoke before the pause is not the same as the speaker that speaks after the pause, we move on to find out if it was evident at the beginning of the pause who was going to speak after the pause. If it was evident who was going to speak after the pause, we will categorise the pause as an initial pause. If the speaker before and after the pause was not the same, and it was not evident at the beginning of the pause who was going to speak after the pause, we will categorise the pause as a TRP-pause with turn change.

There is at least one situation that the flow chart does not cover. We can imagine that a speaker is in the middle of an utterance, and pauses to for example find a word. If the speaker is interrupted by another speaker who begins speaking during the pause, this situation will not fit any of the categories in the model. While we find the categorisation sufficient for the studies in this dissertation, the flowchart could be extended to cover interruptions and other phenomena not currently included.

2.2.2 Individual, social, cultural and linguistic differences

Pauses exist in all human languages and cultures, but the meaning of pauses and silences vary, ranging from uncomfortable and undesirable to useful and even obligatory. This means that the use of silence differs between cultures and languages. The tolerance for silence also varies: what is perceived as, for example, a long pause is dependent on the context and the language spoken. Consequently, while it is possible to create a general description of conversational structure, it is still important to be aware of the influence of contextual factors.

Kendall (2009) examined the impact of different factors on pause length, and a mixed model analysis showed that region, gender and ethnicity are significant influences on pause duration. Age and year of birth were not found to be significant influences. However, the actual differences in mean length between different ethnic groups and males and females are relatively small: 50

ms (between two most differing groups) and 34 ms respectively. Kendall concludes:

> ″We have established, I think, that pause duration is not only an outcome of processing activity and so forth, but that it does in fact appear to be impacted socially by such categories as regional affiliation, ethnicity, and gender, even though the effects of these social categories are far from straightforward.″ (Kendall, 2009, p. 118)

The meaning of silence

Pauses exist within the larger context of silences and their function in communication. Pauses are specifically the silences interspersing an ongoing conversation, but silence may carry illocutionary force and have perlocutionary effects in itself (Sifianou, 2011). The obligation to produce or avoid silences when meeting another person depends on the cultural, situational and individual context. We conform to the norms of the dominant culture, adapt to the situation, while also displaying individual preferences (Sifianou, 2011).

In Akan (spoken in Ghana and Côte d'Ivoire), silences are used to convey for example grief support, and silence is so important that it is explicitly taught in which situations silence should be used instead of words. Silence is therefore not seen as "lack of words", but rather there is a sense of words not being adequate to express the feelings in the situation (Agyekum, 2002).

Politeness and silence are intimately linked in many cultures, and silence is the most polite strategy for handling face-threatening acts (Brown and Levinson, 1978). Nakane (2006) shows that silence can be used as a face-saving act in some cultures, for example in Japan. When investigating communication in university seminars, she found that Japanese students tended to be quiet and not respond when they didn't know the answer to a question. In Japan, this is the unmarked alternative, and the teachers then give more clues and help the students answer. However, when Japanese students used this strategy in Australia, the teachers perceived the same behaviour as negative and they might ask again or wait for the student to respond, causing the student to lose face.

Even in cultures that have an extensive use of silence in common, the connotations of the silence may be different: silence can be perceived as a way of creating distance or as a means of generating the right kind of atmosphere (Sajavaara and Lehtonen, 2011).

Pause length

When a pause in a conversation appears in an unexpected context or has an unexpected length, the pause gains further significance (Krahmer et al., 2002). Levinson (1983, p. 302) describes it as follows:

> [A] speaker addresses a recipient and with the first part of a pair and receives no immediate response. Strong inferences are immediately drawn, either of the sort 'no response means no channel contact', or, if that is clearly not the case, then 'no response means there's a problem'.

As Levinson states, a longer pause than expected will lead the listener to believe that there is trouble in the conversation: either that the other person has not heard or understood what was said, or that the other person is delaying her answer because she is about to say something that is negative or uncomfortable. Roberts et al. (2006) showed that the duration of pauses clearly affects the perception of willingness to comply with requests and agree with assessments. A pause can be a sign of a request being denied — a non-preferred answer. However, the ability to discern pauses that are too long is based on a common understanding of how long a pause should be, and this perception of pause length differs depending on a person's individual, social, cultural and linguistic background.

There are several studies that show how different cultural and language backgrounds entail different pause patterns. A by now classic study describes the difficulties in communication between Anglo-Americans and Athabaska Indians, which is caused by different pause tolerance (Scollon et al., 1981). The Athabaskan Indians commonly use longer pauses than the Anglo-Americans, so when the Anglo-American stops speaking at a TRP, the Athabaska Indians wait 500 ms more than the Anglo-Americans are used to before speaking. This results in the Anglo-Americans resuming speaking before the Athabaska Indians feel that there has been enough of a pause for them to take the turn. Consequently, the conversation becomes dominated by the Anglo-Americans, while the Athabaska Indians do not get many words in edgewise. It is easy to imagine how this might influence the perception of the other person: the person with a shorter pause tolerance might be seen as rudely monopolising the conversation, whereas the person with the longer pause tolerance might be seen as uncommunicative and distant.

Both Fujio (2004) and Mushin and Gardner (2009) show that pause tolerance can vary distinctly between different cultures, and Fujio further points to how pragmatic transfer of pause patterns can cause problems in intercultural communication. Still, there are studies that point out the commonalities of pause timings as well. Stivers et al. (2009) investigated turn transitions with regards to pauses and overlaps in 10 structurally and culturally different languages, and found that turn transitions were quite similar in timing. The mean response time across the analysed languages was 208 ms, with language-specific means falling within approximately 250 ms of the cross-language mean.

2.2.3 Cognitive aspects of pauses

To understand why there is a cognitive need for pauses, we can look toward the underlying processes in the brain. Kircher et al.

(2004) investigated the neural correlates of pausing through the use of functional magnetic resonance imaging (fMRI). Subjects talked spontaneously about Rorschach inkblots, while neural activity was measured. Pauses were defined as an interval of non-speech lasting between 550ms and 3000ms (no silent intervals shorter than 550 ms were found, and therefore no possible pauses were excluded because of falling below that threshold). The 85 longest pauses per subject were chosen for analysis (6 subjects took part in the study). Pauses were divided into pauses at grammatical boundaries, and pauses within clauses. Pauses at grammatical boundaries are equal to pauses at transition relevance places, whereas pauses within clauses are pauses that do not occur at TRPs (however, since the speech studied was monological and not dialogical, no turn changes took place).

The pauses at grammatical boundaries made up 55% and the pauses within clauses 45%. The two types of pauses examined were associated with different activation patterns in the brain, which supports the assumption that pauses serve separate functions. Pauses at grammatical boundaries were associated with activation in the right inferior frontal gyrus, which may be related to conceptual organisation, such as sentence planning and memory retrieval. Pauses within clauses were associated with activation in the left superior temporal, superior frontal, middle temporal and middle frontal gyri. These areas are located within what is known as Wernicke's area, which is involved in the understanding of spoken and written language. Lesions in Wernicke's area lead to impaired planning on the sentence level.

The study of pauses in subjects with language disabilities may give information about the function of pauses in speech production. When examining speech production in adults with aphasia, Butterworth (1979) found that neologisms tended to be preceded by pauses. The subject did not pause significantly more than normal speakers in comparable tasks. If a word was preceded by a pause of 250ms or more, it was considered hesitant. 44,6% of pauses were at grammatical boundaries, which the author suggests is in line with unimpaired speakers. Neologisms were significantly more likely to follow hesitations. Verbal paraphasias (confusing one word for another) are less likely than neologisms to follow pauses. Existing nouns were preceded by pauses in a quarter of the cases, whereas half of the noun neologisms were preceded by a pause. This suggests that the pauses may be connected to word-finding difficulties.

In summary, pauses seem to be related to language planning on various levels, and different types of pauses may be connected to different types of cognitive activities.

2.3 Summary

In this chapter, we have given a brief overview of the history of research on speech and conversation, with a focus on research relating to pauses. We have discussed the need for pauses for breathing, speech planning and turn taking management, and we have introduced four different types of pauses, based on the turn taking rules described by Sacks et al. (1974). Furthermore, we have discussed how pauses vary in different contexts, concluding that pause length is contextually sensitive and is dependent on the social, cultural and linguistic context of the conversation.

Chapter 3

Methodology and material

There is no such thing as an empty space or an empty time. There is always something to see, something to hear. In fact, try as we may to make a silence, we cannot.

John Cage

We have tentatively defined pauses as silences that occur during an ongoing conversation, and during a speaker's turn or at a turn change. In the previous chapter we introduced different types of pauses, and in this chapter we will explore some methodological matters, such as whether a pause has to be of a certain length to be viewed as a proper pause. We will also present the material that the studies in this dissertation are based on.

3.1 Defining pauses

I'm sitting in a cottage garden in the forest, listening to birds chirping and insects buzzing. I also hear the occasional bark from the dogs and a cockerel boasting his strength. There are almost no manmade sounds, and I think to myself: 'how quiet it is here'. Isn't that strange? I just described the sounds I'm hearing, and still I think it's quiet! This experience highlights the relativeness of silence: silence is not something absolute. Rather, it is something that is defined in relationship to sounds. Silence is different for different people: this depends both on their personal definition of silence, but also on their hearing.

Sound level is measured in decibels. The decibel value is expressed as the relationship between the measured sound pressure and a reference sound pressure, which is .0002 microbar (20 microPascal). When the sound pressure level is 0 dB, it means that the measured sound pressure is equivalent to the reference sound pressure. 0 dB is also approximately the quietest sound a healthy, young person can hear (depending on the frequency and duration of the sound, and the environment).

Silence, therefore, cannot be defined as an absolute decibel value. Rather, silence may be defined as "no audible sound" or no sounds above a certain chosen sound pressure level. What a person perceives as silence will depend on that person's auditory capacity, and also on that persons's own idea of what silence is. However, when we are interested in pauses in speech, we may safely limit the concept of pauses to lack of acoustic energy emanating from the person that currently has the turn. This means that when the speaker that has the turn pauses, and another speaker provides feedback, we still regard this as a pause, since the speaker that has the turn is silent.

3.1.1 Shortest perceivable pause

When researching pauses, one important aspect is to decide whether a minimum silence duration should be set, and in that case how that minimum duration should be decided. One way to decide a reasonable minimum duration would be to consider how long a pause has to be to be perceived. Heldner (2011) found substantial individual variation when it comes to detecting pauses at speaker changes, but reported that the mean length that is necessary for a pause in speech to be detected is approximately 120 milliseconds. Perception of these pauses seems to depend on length, but can also be affected by other factors. 120 milliseconds might give an indication of a relevant minimum duration, but the perception of silence will vary a great deal. It should also be noted that Heldner only analysed pauses at speaker changes, and not pauses within speakers' turns.

Different studies use different thresholds when deciding what constitutes a pause. A minimum time is often set, primarily to exclude occlusion intervals in voiceless stops (the occlusion interval is the pressure build-up before the burst). During the occlusion interval no sound is produced, and these intervals can therefore be mistakenly identified as pauses by automatic silence detection methods. Occlusion intervals in voiceless stops in English last 90–140 ms, with an average of 120 ms (Lisker, 1957), but this may vary slightly depending on language, whether the stop is in a stressed syllable etc. Heldner and Edlund (2010) report that in a study of approximately 13000 voiceless stops, almost all (99.2 %) stop closures are shorter than 180 ms. If automatic pause detection methods are used, a minimum duration tends to be required. Table 3.1 shows some silence thresholds for a number of studies.

As shown in table 3.1, maximum silence durations are sometimes set, to exclude silences that are significantly longer than average. This can also be done later in the process, by removing outliers.

When pauses are excluded based on duration, this will naturally have an effect on reported average values. Wlodarczak and Wagner (2013) show that silence thresholds significantly impact research results when exploring silences and talk-spurts, both with regards to frequency and duration. Setting pause thresholds will also influence the interpretation of other conversational phenomena, such as overlaps.

3.1.2 Perceived pauses

One way to define pauses that does not entail the problems with duration thresholds, discussed in 3.1.1, would be to ask some listeners to listen to the dialogues and indicate where they heard

Table 3.1: Pause length cutoff values in milliseconds in various pause studies

Minimum duration	Maximum duration	Reference
10 ms	-	Oehmen et al. 2010
10 ms (gaps)	-	Heldner and Edlund 2010
50 ms	-	Norwine and Murphy 1938
50 ms	-	Kousidis et al. 2013
50 ms	-	Gravano and Vidal 2014
60 ms	5000 ms	Kendall 2009
150 ms	-	van Donzel and Koopmans - van Beinum 1996
180 ms (pauses)	-	Heldner and Edlund 2010
200 ms	-	Yanushevskaya et al. 2014
250 ms	-	Goldman-Eisler 1958
550 ms	3000 ms	Kircher et al. 2004

pauses. If several listeners agree that there is a pause at a certain place, one might assume that this would be a reliable indicator of an actual pause. However, when listeners are asked to annotate where they perceive pauses, the subjective perception of pauses do not always correlate with acoustic silences (Hansson, 1998). Instead, listeners may be relying on other cues, such as inserted vowels (fillers), preboundary lengthening, specific F0-patterns and drops in intensity (Strangert, 1993). Duez (1993) found that segment lengthening, especially lengthening of vowels, has a significant effect on the number of pauses perceived that do not correspond with acoustic silences. Therefore, while perceptually defined pauses may be interesting in their own right, they do not necessarily coincide with acoustically defined pauses.

3.1.3 Pauses and breathing

When we are speaking we breathe differently than when we are breathing quietly. During quiet breathing inhalations and exhalations are of equal length, but when speaking we inhale rapidly, and exhale slowly. We speak primarily when exhaling, and we control the flow of our air to allow a consistent air stream. When studying speech and breathing, utterances are often divided into breath groups. A breath group is a chunk of speech, delimited on each side by pauses necessitated by breathing. A breath group does not have to be identical to a sentence or an utterance, since breathing can occur at different places in speech. One of the main reasons for pausing in speech is breathing, but not all pauses are related to breathing. It has been shown that breathing during speech is affected by cognitive and linguistic planning: for example, subjects tend to inhale more deeply before producing loud or long utterances (McFarland, 2001; Torreira et al., 2015). It is quite possible that the different types of breathing, such as an audible inhalation versus a quiet intake of air, are used as signals in turn taking in conversation. Indeed, Rochet-Capellan et al. (2014) found that inhalations are shorter inside a turn than at the start of a new turn, which shows that we adapt our breathing according to the current state of the conversation, and Torreira et al. (2015) have demonstrated that audible in-breaths can be used for turn taking management.

It seems natural that breathing is affected by speaking, but breathing can also be influenced by listening to someone else speak. Breathing when listening, especially in spontaneous conversation, is different from quiet breathing: inhalations are quicker and more similar to speech breathing (McFarland, 2001).

When transcribing recordings of dialogue, it is possible to transcribe some breathing, as breath sounds sometimes are loud enough to be audible. However, without specialised equipment, it is not possible to annotate all instances of breathing in a dialogue. This leads to some methodological questions: should pauses in which there is audible breathing be treated as different than pauses where no breaths can be heard? Kendall (2009) suggests that, based on previous research, it might not be necessary to differentiate between hesitation pauses and breathing pauses. He also raises the point that it might be too simplistic to say that pauses are either for breathing or for planning, as they may very well fall into both categories. In this dissertation we do not distinguish between pauses that include breathing and pauses where no breathing is heard, but this is something that could be explored in future studies.

3.2 Measuring pauses

3.2.1 Measuring duration

Jefferson (1989) describes how she used to time pauses with a stopwatch when transcribing, but when it broke, she instead started counting silently "one one thousand two one thousand. . . " to get a rough duration measure for pauses. In conversation analysis, pauses are not always measured in seconds: pauses shorter than 0.5 seconds may be referred to as micro-pauses and are transcribed as a period within parenthesis. This of course leads to problems when making comparisons between different studies.

In studies specifically related to pauses, pauses are often reported in seconds or milliseconds, and this leads to easier comparison between studies. Pause length distributions are typically positively skewed (Edlund et al., 2009), which means that using statistic methods that assume that the values are normally distributed (parametric statistics) will not produce reliable results. Image 3.1 shows a histogram of a typical pause distribution, and the histogram is overlaid with a normal distribution curve (the pause distribution presented is of one of the dialogues in the material presented in section 3.3.1. The skewness value for this pause distribution is 2.925 (a non-skewed distribution will have a skewness value of 0) and this is also reflected in the mean being greater than the median: the median for this distribution is 0.527 seconds, whereas the mean is 0.610 seconds.

One way to address the problem of the skewed distribution of pause lengths is to logarithmize the values, which will diminish the skewness. In image 3.2 we see a histogram of the same distribution as in 3.1, but the pause lengths values have been logarithmized. As can be seen in the image, the histogram now follows the normal distribution curve somewhat closer. The skewness has changed to –0.356, which means that the distribution is now slightly negatively skewed instead of the significant positive skewness seen previously. Median and mean values are now closer to each other: the median is naturally still 0.527 seconds, but the mean has changed to 0.488 log seconds.

While transforming the data into the log domain will normalise the distribution, it is not without problems. Log transforming data may introduce artefacts, especially when examining pauses and overlaps combined (Heldner and Edlund, 2010). An alternative to log transforming data and using parametric statistics is to use generalized linear mixed models (GLMMs), which do not require normally distributed data. However, in this dissertation we use log transforms where necessary.

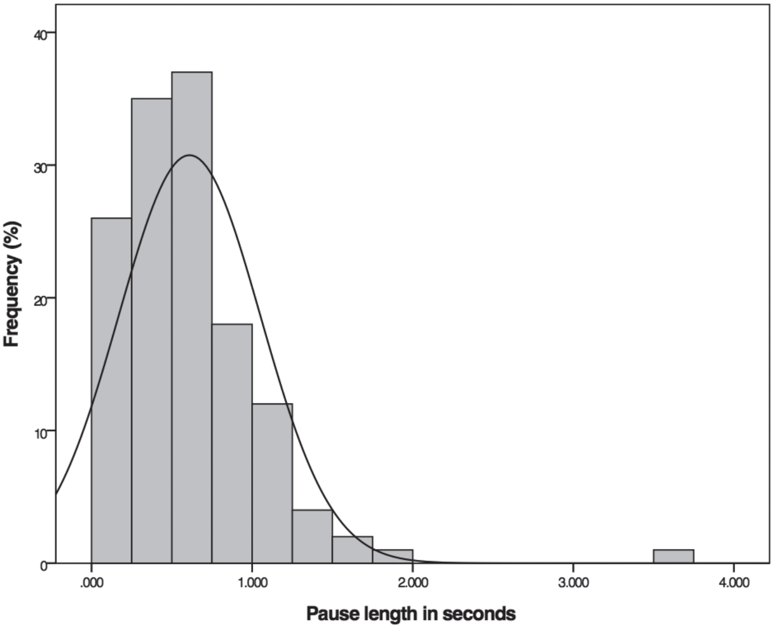

Figure 3.1: Histogram of pause distribution overlaid with normal distribution curve

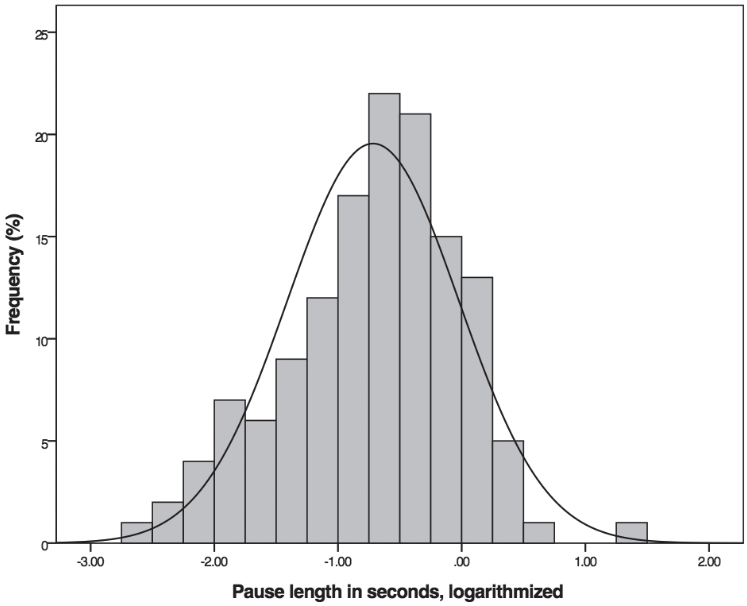

Figure 3.2: Histogram of logarithmized pause distribution overlaid with normal distribution curve

3.2.2 Absolute length or relative length: speech rate

The rate with which we articulate speech sounds varies both intra- and interindividually, and speech tempo can be measured in differing ways (Szczepek Reed, 2010). Articulation rate is measured in syllables per second excluding pauses, whereas speech rate includes pauses. The perception of speech rate is dependent on the conversational context, as the speech rate is perceived in relation to surrounding utterances by the same speaker, and to the other speakers involved in the conversation. It also depends on the number and length of pauses produced. Therefore, one might argue that absolute pause length is not interesting, as it does not tell us anything about how the length of the silence was perceived.

Kendall (2009) examined articulation rate and pauses from a sociolinguistic perspective, and found that there is a tendency for pauses to become shorter when articulation rate increases (note that although Kendall refers to it as speech rate, the definition corresponds with articulation rate). Articulation rate is measured in syllables per second, omitting any silent intervals longer than approximately 60 ms. Articulation rate also varies based on different sociolinguistic variables, such as gender and dialect. This provides evidence that articulation rate is not perceived as faster solely because pauses are shorter, but because speech sounds are articulated more quickly.

3.2.3 Automatic versus manual identification of pauses

When researching spoken language, it is necessary to process the data in various ways. This often includes segmenting the material into for example speech and pauses, and transcribing the material, which is a time-consuming task. Transcription can be done orthographically or phonetically, depending on what the material is going to be used for. When automatic segmentations are possible, the amount of data that can be processed increases greatly. However, automatic segmentations may lead to a decrease in the reliability of the annotations.

Automatic identification and annotation of pauses are typically dependent on cutoff values for the lengths of silent and sounding intervals, and dB limit. This means that the researcher must decide how long a pause should be to be considered a pause, and how much quieter than the speech it should be (some examples of cutoff values for duration are given in 3.1.1). Possible problems that arise are:

- Overinclusion with regards to length: if the length cutoff value is too short, occlusion intervals from stop consonants may be mistakenly identified as pauses

- Overinclusion with regards to decibel level: if the decibel cutoff value is to high, softer speech sounds may be mistakenly identified as pauses

- Underinclusion with regards to decibel level: if the decibel level is set too low, some pauses will not be identified as pauses due to for example background noise

Underinclusion with regards to length will always be present when a cutoff value for length is set, which means shorter pauses will not be included in the selection. Often overinclusion and underinclusion with regards to decibel will occur in the same sample, due to variance in speech loudness.

On the other hand, manual identification and annotation may also result in errors. If the transcribers do not have access to a spectrum or spectrogram representation of the material, they may perceive pauses that do not exist in the material (Duez, 1993).

Oehmen et al. (2010) investigated inter-rater reliability in manual identification of pauses. Four raters segmented sound files into stretches of speech and pauses, and one of the sound files was segmented again by each rater a week later. The boundaries that were annotated marked speech transitioning to silence and vice versa. Some sounds were excluded, for example coughing and audible movements of the muscles of articulation. The minimum pause duration threshold was 10msec. When comparing the transcriptions, it was found that intra-analyst reliability was high and the analysts were consistent in their ratings. Inter-analyst reliability was lower, and was inversely proportionate to the quality of the recording. The main difficulty seemed to be that the analysts were not in agreement on how to classify audible breaths and sounds of the articulators moving. Oehmen et al. (2010) proposed that improved segmentation guidelines would resolve the inter-rater differences.

In section 4.1.2 a comparison between automatic and manual methods of silence detection is presented.

3.3 Audio material

The studies on the production of pauses in chapter 4 is based on a corpus of consisting of human-human dialogues. Further details about the corpus follows below.

3.3.1 PauDia corpus

The spontaneous conversation material in this study consisted of 6 dialogues, and the recordings were made in a recording studio at the Department of Linguistics at Lund University. Five persons, all female, took part in the recordings, and each person was involved in two or three dialogues. The speakers ranged in age from 21 years to 28 years. All speakers were native Swedish speakers without language or hearing impairments, and the dialogues were carried out in Swedish. The speakers all knew each other from the university program they were taking part in. Altogether, 6 dialogues were recorded, each lasting approximately 10 minutes. The speakers have been given pseudonyms, and their pseudonyms and ages are presented in table 3.2.

In table 3.3, the speakers' speaking time per dialogue is displayed, together with the total duration of each dialogue, and the number of turns for each speaker per dialogue. We can see that

Table 3.2: Speakers: pseudonyms and ages

Speaker	**Pseudonym**	**Age**
A	Anna	22 years
B	Beatrice	23 years
C	Cilla	28 years
D	Doris	26 years
E	Emma	21 years

the number of turns varies widely, from 18 turns per speaker in dialogue 3 to 45 turns per speaker in dialogue 1 and 5.

To get the speakers started they received a question to discuss, but they were informed that they did not need to constrain themselves to that subject. The recordings were made on analogue audio tapes, and were digitized with a sampling frequency of 44100 Hz. The material was transcribed ortographically in Praat (Boersma and Weenink, 2014), and pauses were identified manually based on the acoustic signal.

The PauDia corpus is the basis of the analyses in sections 4.3, 4.4, 4.5 and 4.6.

3.4 Summary

In the current chapter several methodological issues have been discussed. We have argued that it is necessary to provide a thorough definition of what constitutes a pause, and how pauses are to be detected, before undertaking a study on pauses. We have also introduced the PauDia corpus which is the basis the production studies in this dissertation are based.

Table 3.3: Dialogues 1-6: speakers and speaking time (number of turns are given within parenthesis)

Dialogue	1	2	3	4	5	6
Anna (speaking time)	229 s (45)	-	202 s (18)	-	274 s (43)	-
Beatrice (speaking time)	-	153 s (29)	-	-	188 s (43)	-
Cilla (speaking time)	-	-	241 s (18)	-	-	257 s (21)
Doris (speaking time)	-	250 s (30)	-	272 s (25)	-	150 s (22)
Emma (speaking time)	216 s (45)	-	-	212 s (25)	-	-
Total duration	574 s	558 s	609 s	608 s	625 s	537 s

Chapter 4

Production of pauses

There was a terribly ghastly silence.
There was a terribly ghastly noise.
There was a terribly ghastly silence.

Douglas Adams (*The Hitchhiker's Guide to the Galaxy*)

In this chapter, various characteristics of pauses are investigated. The aim is to determine the typical length of pauses, how pauses vary depending on the conversational partner, where pauses ordinarily occur syntactically, and how pauses are distributed over the course of a dialogue. However, first we present two methodological studies: the first study examines inter-rater reliability with regards to pause categorisation, and in the second study we compare automatic pause annotations and manual pause annotations.

When we know how pauses typically vary over the course of a dialogue, we can use this information for example when planning speech output in dialogue systems. It will also be beneficial to have a baseline with which to compare speech of persons with different types of communication disorders. Knowing more about the function of pauses in typical conversations, and how they affect communication as a whole, will give us further understanding of the dynamics of communication and turn taking.

4.1 Reliability

In the previous chapter, we discussed several methodological problems concerning pause research. Both annotation and categorisation of pauses will vary depending on the criteria used in a study: for example, if pauses are defined as having a minimum duration, all pauses shorter than that duration will be excluded from pause annotations. In order to investigate the reliability of our methods, we have undertaken two methodological studies. In the first study, we analyse the reliability of our pause categorisation guidelines, by calculating inter-rater reliability. In the second study, we investigate automatic pause detection and how this compares to manual annotations of pauses.

4.1.1 Inter-rater reliability

The robustness of the pause taxonomy described in Section 2.2.1 was tested with inter-rater reliability measures. Two linguists, who had not previously been involved in the project, were asked to categorise the pauses in two one-minute dialogue samples. Guidelines for categorisation were based on the flowchart in Section 2.2.1.

The dialogue samples provided for the raters were accompanied by orthographic transcriptions, and pauses were labelled 'pause'. The first dialogue sample contained 18 pauses and the second 23 pauses.

Inter-rater reliability measured using Cohen's kappa was 0.674, which equals substantial agreement (Landis and Koch, 1977)[1] We interpret this to mean that the flowchart in section 2.2.1

[1] It should be noted that the interpretation guidelines provided by Landis

captures the differences between pause types relatively well, and that a categorisation based on this flowchart will be reliable. It would likely be possible to achieve a higher degree of agreement, if the raters had been trained on examples from dialogues instead of just having access to the flowchart.

4.1.2 Reliability of automatic pause annotations

In Section 3.2.3 we discussed some of the advantages and disadvantages of using automatic pause detections. Our hypothesis is that automatic pause detection will be less successful than manual pause annotation, both since a minimum duration threshold for pauses needs to be set with automatic detection, and because existing automatic pause detection methods are based on sound intensity level, and this may vary greatly during a dialogue. In order to test the hypothesis that automatic identification of pauses may not be good enough for pause research, an experiment was carried out. The material used was a 563 second sample from a dialogue from the Spontal corpus. We used the Spontal corpus instead of the PauDia corpus to evaluate automatic pause annotations, since the SponTal corpus has better separation between the two speakers' audio channels. The Spontal corpus was recorded at the Royal Institute of Technology in Stockholm (Edlund et al., 2010).

In this outtake, pauses were identified manually and automatically. The automatic identification of pauses was done in the Praat phonetic analysis software, which has a built-in function for doing silence detection (Boersma and Weenink, 2014). This function, which is based on intensity analysis, finds the intervals in the sound file which falls below a certain decibel threshold. The silence threshold is set with respect to the maximum intensity in the current sound file, which means that if the silence threshold is set to for example 35 dB, anything that is 35 dB lower than

and Koch (1977) are somewhat arbitrary, and it is important to recognise that the threshold for what is a "good enough" agreement will vary between scientific disciplines.

the maximum intensity will be regarded as silent. The decibel threshold can be set by the user, as well as other parameters such as minimum time for something to count as silence or sound.

Minimum silent intervals and minimum sounding intervals were set to 200 ms. This was done to exclude occlusion intervals in consonants. The analysis was run 7 times with different decibel thresholds: 30db, 35db, 40db, 45db, 50db, 55 db and 60db.

The automatically annotated pauses were checked against the manually transcribed pauses. We considered the manual annotations our "gold standard" (for more on the manual annotation process, see Section 4.2). If the start of a manually annotated pause overlapped with a silent interval in the automatic annotation, the automatic annotation was considered correct if the interval started within +/- 100 ms of the manual annotation. The same principle was used for the end of pauses. If an automatically annotated pause both starts and ends within +/- 100 ms of the manually annotated pause, this was considered a match. The 100 ms interval was chosen since we know that many pauses are around 200 ms in length, and therefore even a deviation of 100 ms would be considerable. Image 4.1 presents an overview of the comparison process. In the manual annotation, 171 pauses were found.

The results of the automatic pause annotations are presented in table 4.1.

To evaluate the automatic pause annotations, precision and recall-values are used. High precision would mean that out of all intervals annotated as silent pauses by the automatic method, the majority are actual pauses that are consistent with the manual annotation, whereas high recall would mean that a majority of all pauses identified in the manual annotation are labelled as pauses by the automatic method. The F-measure is the harmonic mean of precision and recall.

As can be seen in table 4.1, the number of silent intervals detected decrease as the distance between the maximum intensity level of the dialogue and the threshold for silence increases.

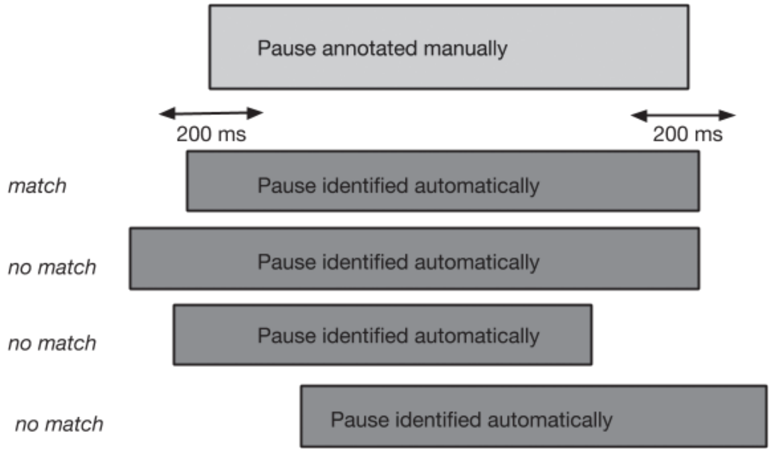
Pause annotated manually
200 ms
200 ms
match
Pause identified automatically
no match
Pause identified automatically
no match
Pause identified automatically
no match
Pause identified automatically

Figure 4.1: Comparing manual annotations of pauses with automatic identifications of pauses

Table 4.1: Automatic pause annotations

	30 dB	35 dB	40 dB	45 dB	50 dB	55 dB	60 dB
Number of silent intervals identified	351	363	322	268	193	134	79
Number of hits	20	30	39	55	45	35	14
Precision	5.7 %	8.3 %	12.1 %	20.5 %	23.3 %	26.1 %	17.7%
Recall	11.7 %	17.5 %	22.8 %	32.2 %	26.3 %	20.5 %	8.2 %
F1-measure	0.077	0.113	0.158	0.250	0.247	0.230	0.112

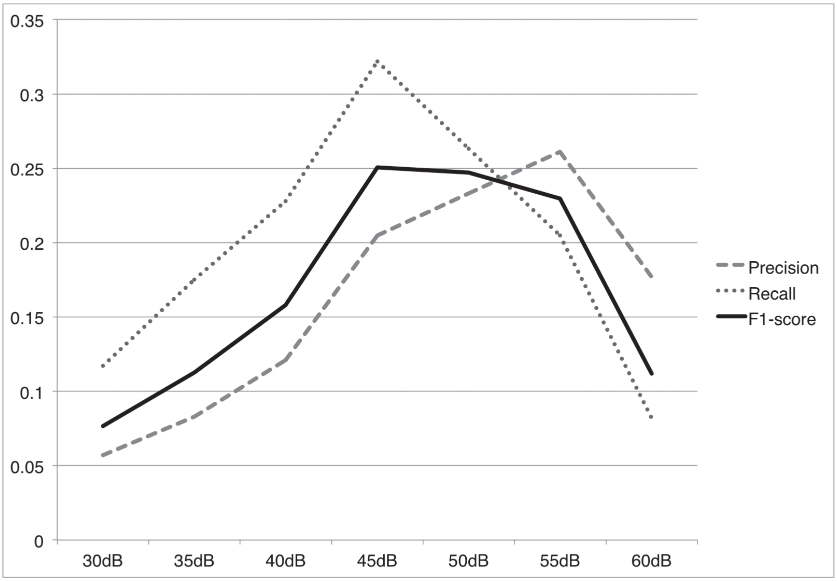

Figure 4.2: Recall, precision and F-measure for automatic pause annotation, with 200 ms cutoff

When looking at precision and recall, it is evident that setting the silence threshold to 45 dB gives the best results (see figure 4.2 for an overview of precision, recall and F-measure at different levels). However, at this level, precision is still only 20.5 % and recall 32.2 %, which means that a majority of pauses are not identified by the automatic process.

In the manual annotations, around 50 of the pauses were shorter than 200 ms, and will therefore not be found by the automatic analysis. This limits the rate of recall. Based on the data presented in table 4.1 , two more automatic pause annotations were tested. In this trial, the shortest pause allowed was 100 ms, so half of that in the previous trials. Based on the F-measure in the previous trials, 45 db and 55 db were chosen as the silence thresholds. For the automatic annotation with silence threshold 45 db, 351 silent intervals were identified. Precision, recall and F-measure were 25.4 %, 52 % and 0.341 respectively. With the 55 db silent threshold, 172 silent intervals were identified. Precision

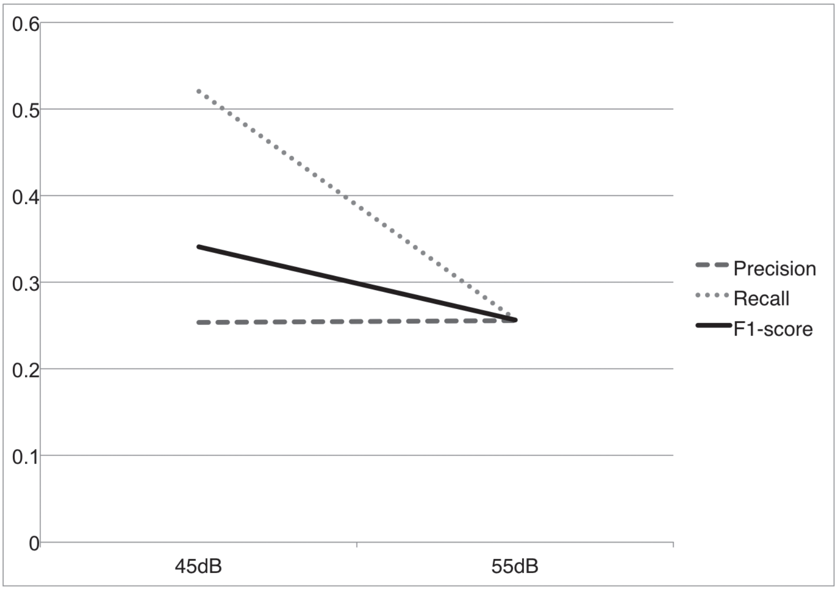

Figure 4.3: Recall, precision and F-measure for automatic pause annotation, with 100 ms cutoff

was 25.6%, recall was 25.7% and F-measure was 0.257. The results are presented in figure 4.3 .

Clearly, decreasing the shortest interval allowed improves the result, and at best we reached an F-measure of 0.341.

To improve automatic annotation of pauses, a first step should be to merge results from different silence thresholds. This could be done with some sort of weighting. It should also be possible to use machine learning to train an algorithm to recognise stops, so that these can be excluded from the data. The cutoff for shortest pause duration could then be lowered without the risk of including stops.

In upcoming studies it may be well worth to evaluate automatic pause annotations anew, since technology is constantly evolving. It is also possible to combine automatic and manual methods, by starting with automatic annotations and correcting them manually. Still, the evaluation made here clearly supports the decision to annotate pauses manually to get reliable results.

4.2 Identification and annotation of pauses

In the present dissertation, pauses have been identified and annotated manually, since it has been shown that automatic identification of pauses is not very reliable (see section 4.1.2). The audio files were analysed and transcribed in the software Praat (Boersma and Weenink, 2014), where it is possible to view the sound both as a spectrum and a spectrogram. An example of an outtake of a sound file is presented in figure 4.4. The figure shows a spectrogram of the sound at the top, and below that a spectrum, with the transcription and boundaries between speech and pauses at the bottom. Intensity is shown in the spectrogram as a gradient from white to black, where black is the highest intensity. In the spectrum, intensity is shown on the y-axis. In figure 4.4, we can see that during the interval labelled "pause", sound intensity is low. All sound files were inspected manually, and pauses were identified through auditory and visual examination of the sound. When a pause has been labelled, the accuracy is controlled through playback of the interval, making sure that no sound was included. The preceding and following sounding intervals are also checked near the boundary, to make sure that no part of the pause has been included in the interval labelled as speech. This eliminates the need for a cutoff value for the duration of pauses (see section 3.1.1). Pauses were categorised into different pause types, based on the descriptions in 2.2.1.

Since breathing was not monitored during the recording of the material, pauses that include audible breaths are not treated differently in the analysis than pauses where no breathing can be heard (see section 3.1.3). As Kendall (2009) suggests, pauses that include breathing may very well serve other functions as well.

The relationship between articulation rate and pauses is an interesting area which merits further study, but in this dissertation articulation rate has not been taken into consideration when analysing pauses. Pause duration is consistently measured in milliseconds, and log-transforms of pause lengths is used to normalise the data when calculating and comparing mean lengths

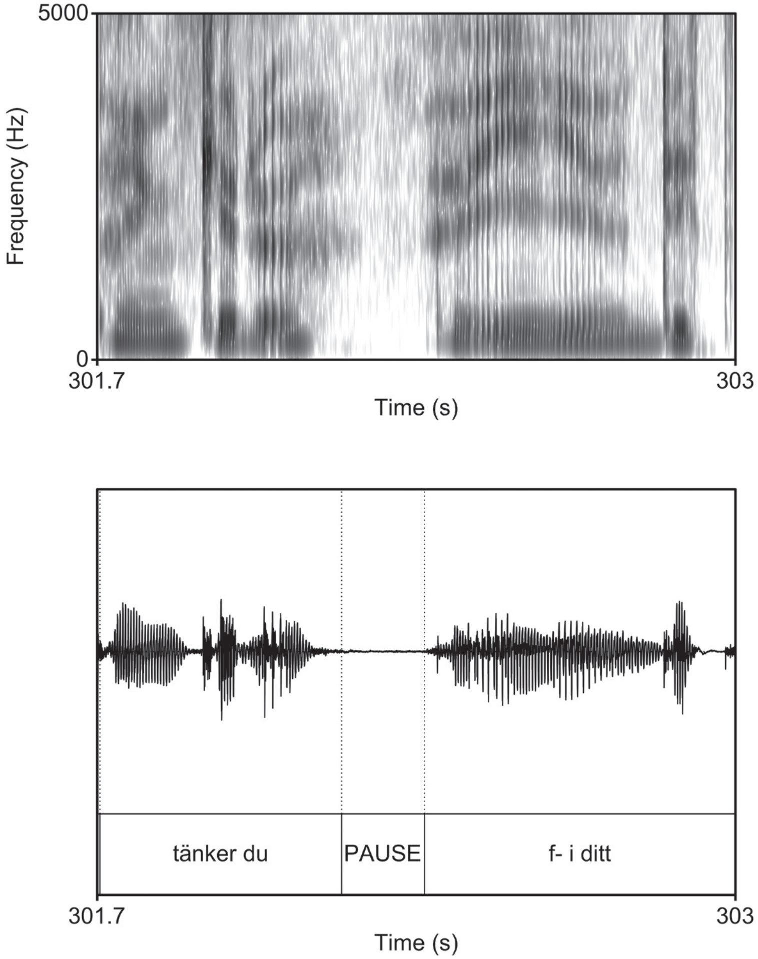

Figure 4.4: An example of a pause annotated in Praat. The image shows a spectrogram, a spectrum, annotation of a pause, and transcription of the surrounding speech.

(for more on this, see section 3.2.1).

4.3 Mean pause lengths

4.3.1 Introduction

The majority of pauses are shorter than 1000 ms, and the most common gap length is 200ms (Heldner and Edlund, 2010). When analysing pauses in English, French, German, Italian, and Spanish, Campione and Véronis (2002) found that pauses seem to be trimodally distributed, supporting a categorisation into brief (<200 ms), medium (200–1000 ms) and long (>1000 ms) pauses. The group of long pauses are only present in spontaneous speech, and not in read speech. When there is a stronger prosodic boundary, the pause tends to be longer (Horne et al., 1995), and grammatical pauses, i. e. pauses at transition relevance places, are significantly longer than non-grammatical pauses (median length 384 vs 293 ms) (Kendall, 2009). There are differences between how individual speakers use pauses, and van Donzel and Koopmans - van Beinum (1996) found that the use of pauses, filled pauses (also known as fillers) and lengthenings of words vary between individuals, but that each speaker seemed to consistently prefer a particular strategy. For example, some speakers used a lot of fillers, whereas other speakers hardly used any.

4.3.2 Aim

We established in section 2.2.1 that pauses can be divided into different categories, depending on their relation to transition relevance points (TRPs) and turn changes. Our aim is to analyse whether different types of pauses differ significantly with regards to mean durations, and if speakers employ different pause strategies, for example preferring one type of pause over another. The pause types investigated in this section are the four different types of pauses presented in Section 2.2.1.

Our hypothesis is that TRP-pauses without turn change will be longer than planning pauses, as during TRP-pauses without turn change, the speaker is trying to deduce whether the other speaker wants to take over the turn, while also planning her own turn, whereas during planning pauses the speaker is mainly concerned with speech planning, and not with managing turn changes. Further, we hypothesise that speakers will differ with regards to the ratio of planning pauses and TRP-pauses without turn changes.

4.3.3 Method

The material described in section 3.3.1 was analysed with regards to the duration of pauses. Pauses have been identified manually, and therefore no cutoff value has been used. The methods used for analysis of the material are summarised in Section 3.4. Mean durations for each pause type per dialogue were calculated in the log domain (see section 3.2.1)

4.3.4 Results

The results will be presented separately for each dialogue. Initial pauses were quite rare, with each speaker producing around 5 such pauses per dialogue, and they are therefore not included in the presentation of the results. Planning pauses and TRP-pauses without turn change are categorised as belonging to the speaker that speaks before and after the pause. TRP-pauses that occur at turn changes are not categorised as belonging to any speaker (although one might argue that they could be seen as belonging to the speaker that speaks after the pause). Therefore, they are presented as a separate group.

For each dialogue, the variation in duration of planning pauses, TRP-pauses without turn change, and TRP-pauses with turn change is presented graphically in box plots. Since box plots do not assume normally distributed data, the pause durations are plotted in milliseconds and not transformed into the log domain.

Dialogue 1

Anna and Emma are the speakers in dialogue 1 . The duration of the different pauses and the distribution of pause categories in dialogue 1 are presented in table 4.4 and figure 4.5[2].

Table 4.2: Mean pause lengths in ms per speaker in D1 (based on log-transformed values)

Speaker	Planning pauses	No of pauses (%)	TRP-pauses w/o turn change	No of pauses (%)
Anna	302 ms	43 (46%)	575 ms	50 (54%)
Emma	503 ms	33 (19%)	470 ms	143 (81%)

40 TRP-pauses with turn change were produced in dialogue 1, with a mean length of 338 milliseconds (log-transformed). Analysis of the variance shows that the differences in mean pause duration between the three groups of pauses are significant (F=5.843, p=0.003), with Post Hoc tests showing statistically significant differences between all three groups.

Dialogue 2

Beatrice and Doris are the speakers in dialogue 2. The duration of the different pauses and the distribution of pause categories in dialogue 2 are presented in table 4.3 and figure 4.6 In dialogue 2, 34 TRP-pauses with turn change were found, with a mean length of 438 milliseconds (log-transformed).

An ANOVA test showed no statistically significant differences between mean pause lengths (F=0.513, p=0.5999).

[2] One TRP-pause at turn change, which is 4410 ms long, falls outside the plot area due to it being considerably longer than the other pauses.

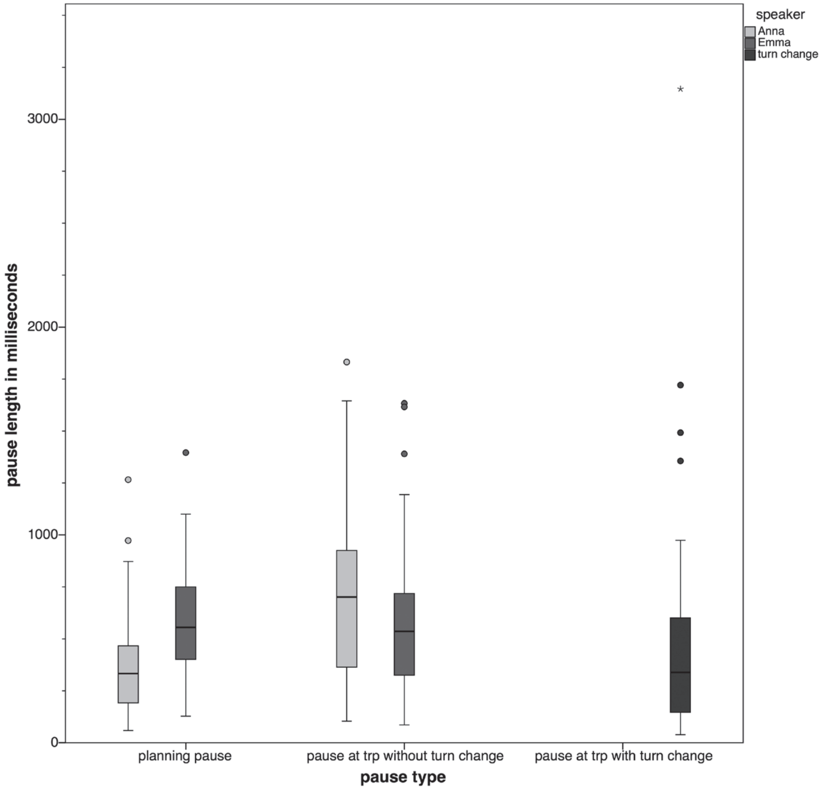

Figure 4.5: Pause lengths per pause type and speaker in dialogue 1

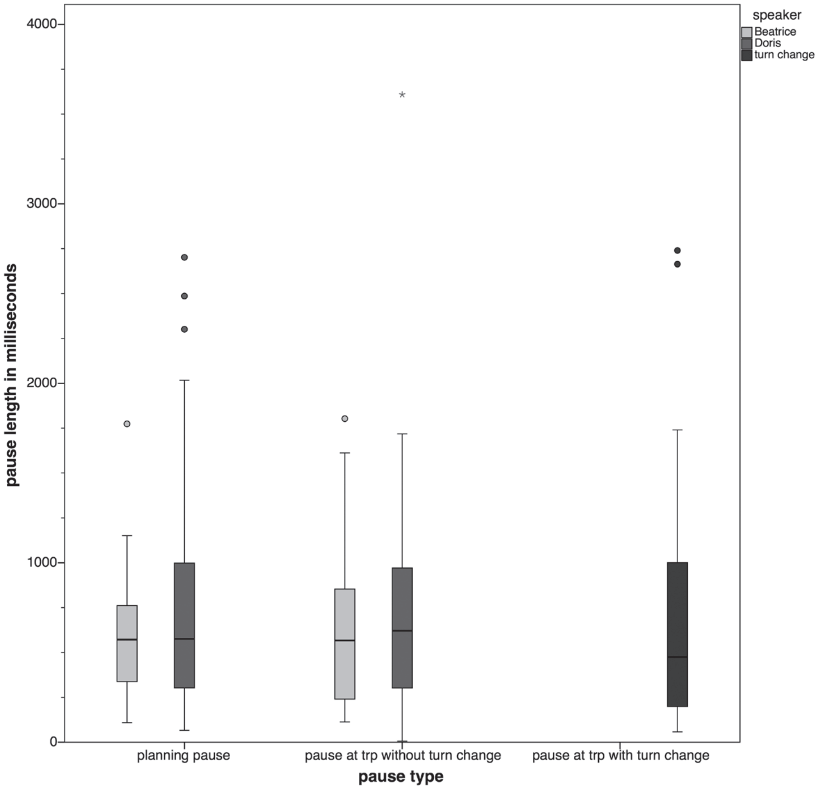

Figure 4.6: Pause lengths per pause type and speaker in dialogue 2

Table 4.3: Mean pause lengths in ms per speaker in D2 (based on log-transformed values)

Speaker	Planning pauses	No of pauses (%)	TRP-pauses w/o turn change	No of pauses (%)
Beatrice	509 ms	31 (52%)	471 ms	29 (48%)
Doris	528 ms	76 (63%)	497 ms	45 (37%)

Dialogue 3

Speakers in dialogue 3 are Anna and Cilla. In figure 4.7, the planning pauses, TRP-pauses without turn change, and TRP-pauses with turn change are shown (not log-transformed). One outlier falls outside the plot area: Cilla produced one TRP-pause without turn change that was 4949 milliseconds long, and that pause is not visible in the figure.

Table 4.4: Mean pause lengths in ms per speaker in D3 (based on log-transformed values)

Speaker	Planning pauses	No of pauses (%)	TRP-pauses w/o turn change (SD)	No of pauses (%)
Anna	418 ms	46 (49%)	700ms	47 (51%)
Cilla	571 ms	57 (57%)	643 ms	42 (42%)

In dialogue 3, we found 19 TRP-pauses with turn change, with a mean length of 512 milliseconds (log-transformed).

An ANOVA shows that the means of the log-transformed pauses differ significantly (F=4.029, p = 0.019). A Post Hoc test verifies that the mean duration of pauses at TRP without turn change differ significantly from planning pauses. The mean duration of the TRP-pauses at turn change does not differ from the other pause groups in a statistically significant way.

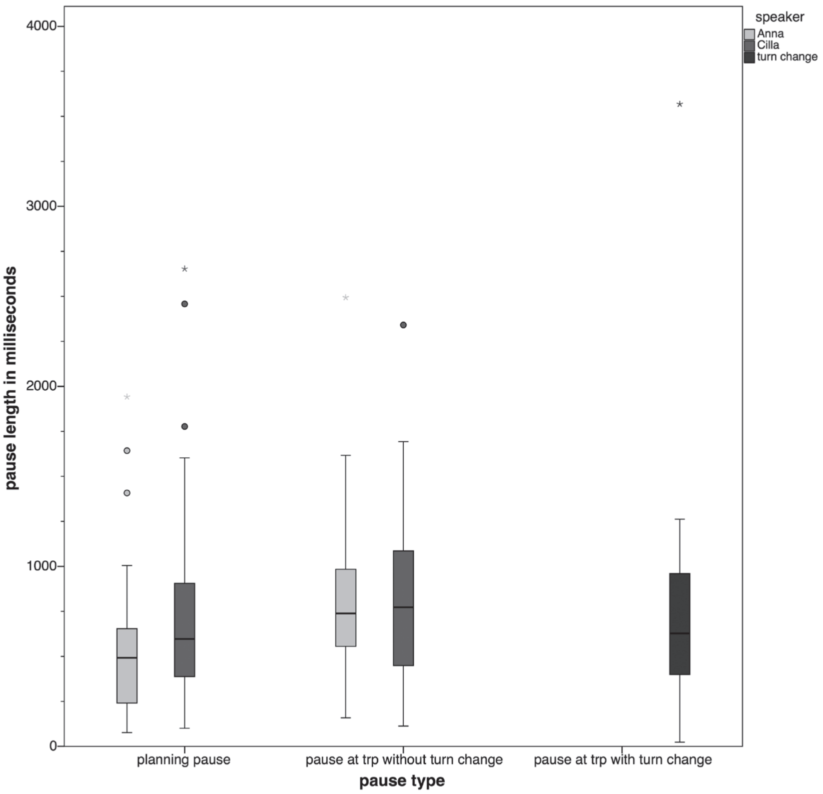

Figure 4.7: Pause lengths per pause type and speaker in dialogue 3

Dialogue 4

The speakers in dialogue 4 are Doris and Emma. Results from the analysis of the pauses in dialogue 4 are shown in table 4.5 and figure 4.8.

Table 4.5: Mean pause lengths in ms per speaker in D4 (based on log-transformed values)

Speaker	Planning pauses (SD)	No of pauses (%)	TRP-pauses w/o turn change (SD)	No of pauses (%)
Doris	534 ms	73 (66%)	476 ms	37 (34%)
Emma	545 ms	36 (55%)	398 ms	30 (45%)

39 TRP-pauses at turn change were found in the material, and the mean duration of those pauses (calculated in the log domain) was 242 ms.

The difference in mean duration between the pause types was statistically significant (F=8.891, p<0.0001), and Post Hoc tests showed that mean durations of planning pauses and TRP-pauses with turn changes were significantly different.

Dialogue 5

Speakers in dialogue 5 are Anna and Beatrice. The results from the analysis of pauses in dialogue 5 is shown in table 4.6 and figure 4.9.

49 TRP-pauses with turn change occurred in this dialogue, and the mean length of these was 404 ms (log-transformed).

The differences in mean pause durations between the three pause types were found to be statistically significant (F=5.224, p=0.006), and Post Hoc tests showed that the difference between mean durations of planning pauses and TRP-pauses without turn change was significant.

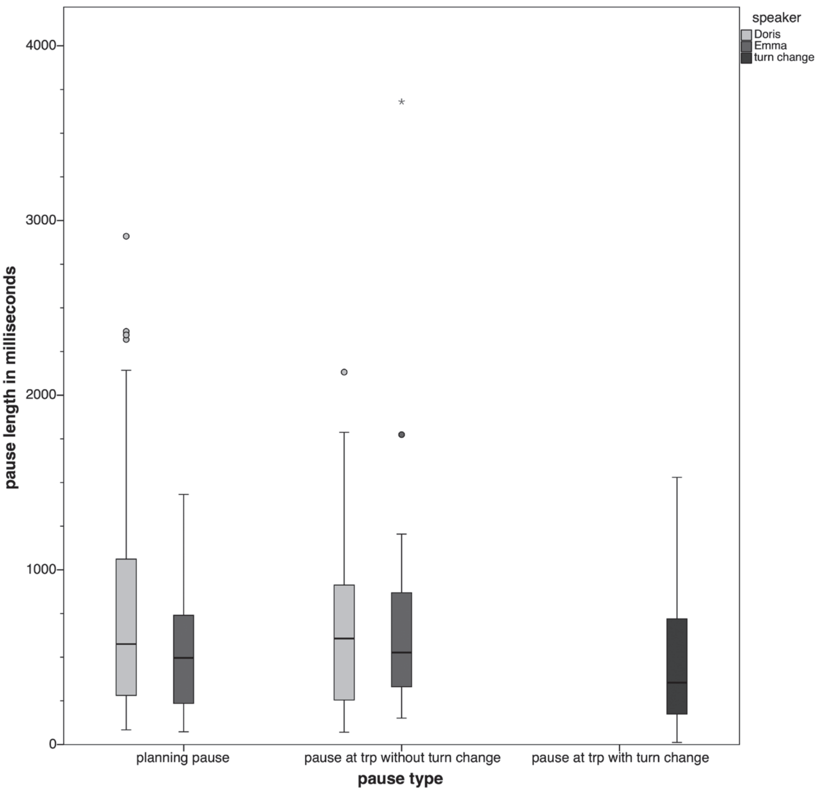

Figure 4.8: Pause lengths per pause type and speaker in dialogue 4

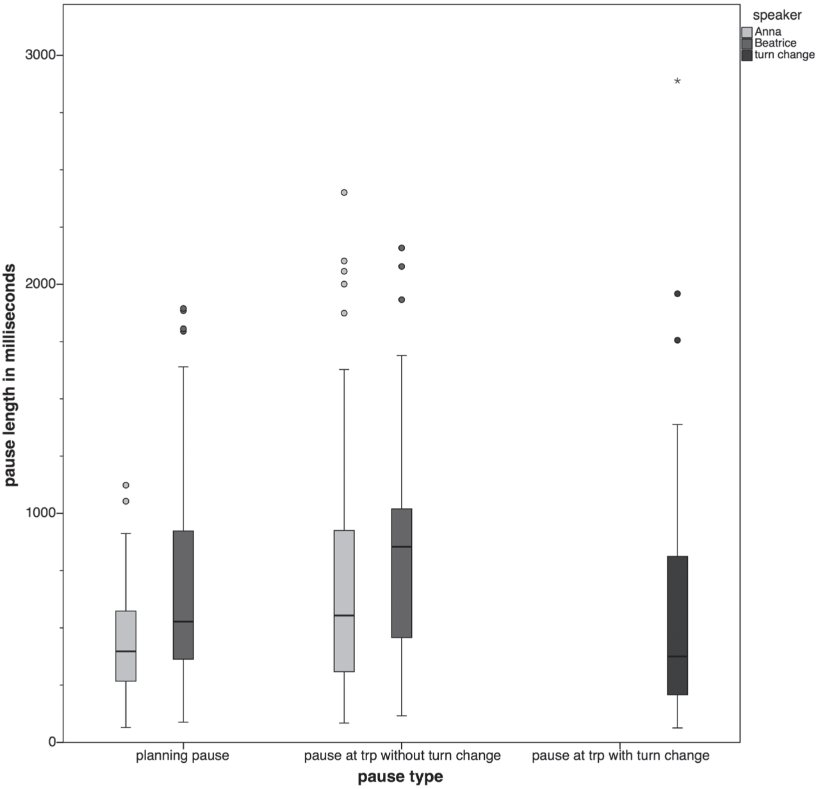

Figure 4.9: Pause lengths per pause type and speaker in dialogue 5

Table 4.6: Mean pause lengths in ms per speaker in D5 (based on log-transformed values)

Speaker	Planning pauses (SD)	No of pauses (%)	TRP-pauses w/o turn change (SD)	No of pauses (%)
Anna	359 ms (242 ms)	50 (50%)	540 ms (586 ms)	51 (50%)
Beatrice	549 ms (532 ms)	41 (55%)	720 ms (517 ms)	34 (45%)

Dialogue 6

Speakers in dialogue 6 are Cilla and Doris. Results from the analysis of dialogue 6 are presented in table 4.7 and figure 4.10.

Table 4.7: Mean pause lengths in ms per speaker in D6 (based on log-transformed values)

Speaker	Planning pauses	No of pauses (%)	TRP-pauses w/o turn change	No of pauses (%)
Cilla	491 ms	77 (67%)	543 ms	38 (33%)
Doris	528 ms	58 (75%)	611 ms	19 (25%)

10 TRP-pauses with turn change occurred in dialogue 6, and the mean duration was 785 ms (log-transformed).

The difference in duration between different pause types was not statistically significant (F=1.654, p=0.194).

4.3.5 Discussion

Anna's and Cilla's mean durations of planning pauses are shorter than their mean durations of TRP-pauses without turn change in all dialogues. They are also quite consistent in their pause preferences: Anna uses approximately the same amount of planning

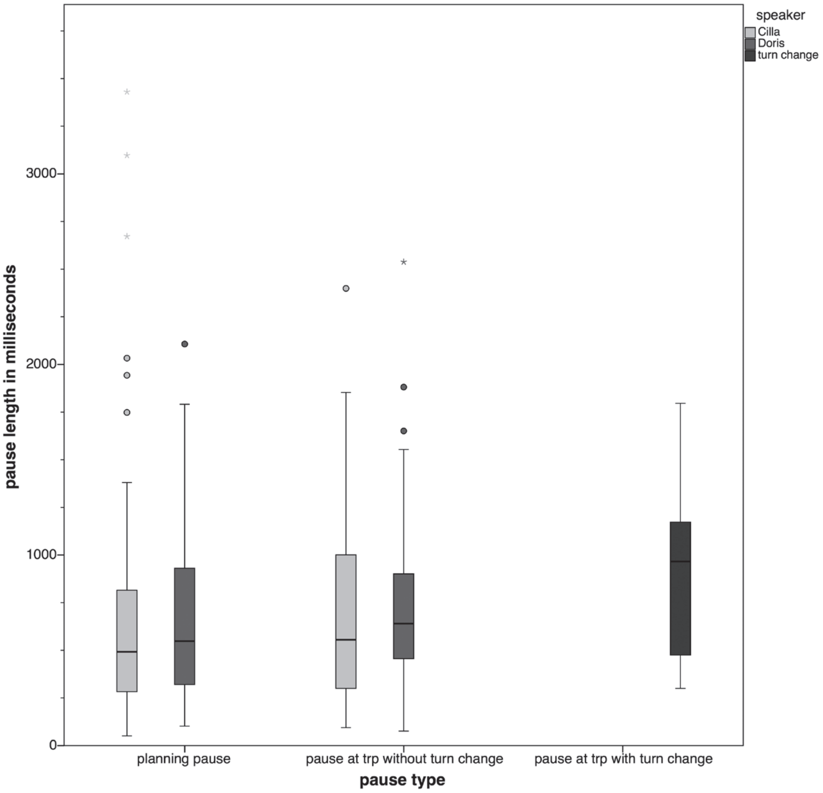

Figure 4.10: Pause lengths per pause type and speaker in dialogue 6

pauses and TRP-pauses, whereas Cilla seems to have a slight preference for planning pauses. So far, this is in line with our hypotheses regarding pause lengths and preference for pause types.

However, the speaker Emma acts completely opposite to what we predicted. The mean duration of her TRP-pauses without turn change is longer than the mean duration of her planning pauses, and in one dialogue only a fifth of her pauses are planning pauses, while in another dialogue over half of her pauses are planning pauses. Speakers Beatrice and Doris are consistent in their preference of pause types. Doris prefers planning pauses over TRP-pauses without turn change, whereas Beatrice uses approximately an equal amount of each type.

van Donzel and Koopmans - van Beinum (1996) has shown that speakers tend to prefer different types of pauses, and the majority of the speakers (four out of five) in our study are consistent in the amount of pauses of a certain type that they use. We expected speakers to produce longer TRP-pauses without turn changes than planning pauses, but this hypothesis did not hold. Pause lengths vary greatly across speakers and dialogues, and there is no evidence that each planning pause is shorter than each TRP-pause without turn change. However, in dialogues 1, 3, 4 and 5, we found that the difference in mean duration between pause types was statistically significant between two or more groups of pauses.

4.4 Pause length entrainment

4.4.1 Introduction

In section 2.1, we briefly mentioned entrainment. To entrain can be defined as “cause (another) gradually to fall into synchronism with it” [3]. When we speak about entrainment in the context of communication, we mean the process in which two (or more) persons speaking to each other tend to become more similar, and

[3] New Oxford American Dictionary, Third edition

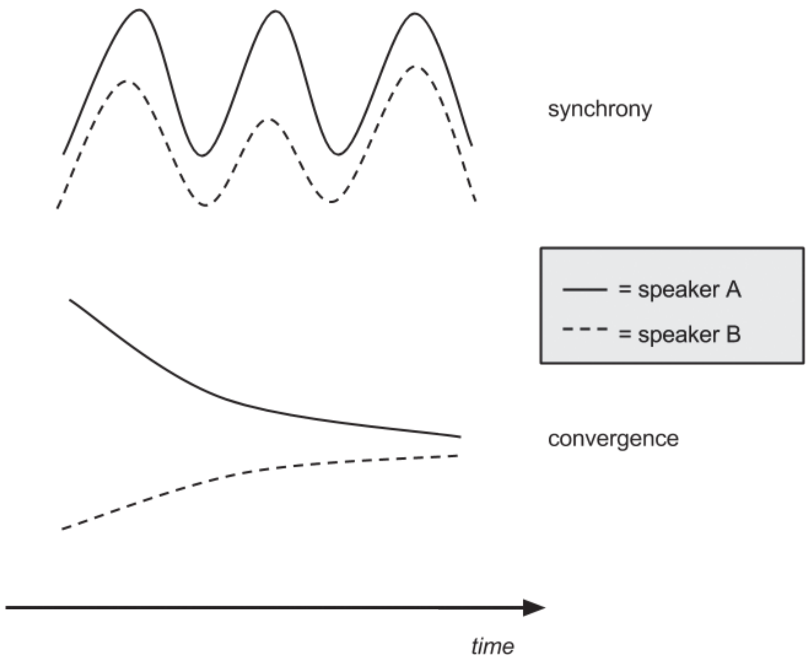

Figure 4.11: Depiction of synchrony and convergence

also that the persons speaking to each other will be more similar to each other than to other persons. Entrainment has been investigated in numerous studies, and evidence has been found for for example lexical entrainment (Brennan and Brook, 1996), phonetic entrainment (Pardo, 2006) and acoustic-prosodic entrainment (Levitan, 2011). Edlund et al. (2009) analyzed pause and gap lengths in dialogues, and found indications of entrainment, albeit the results were not consistent across all dialogues.

Researchers have used various methods to capture entrainment in speech, and two basic aspects of entrainment are synchrony and convergence, which are visualised in image 4.11 (image inspired by Edlund et al. (2009)).

Synchrony means that speakers are consistently behaving in a similar way, whereas convergence indicates that speakers become progressively more and more similar over a period of time. Both processes can be present at the same time: two speakers may exhibit synchronous behaviour that deviates from what we would expect would happen by chance, and at the same time become more similar. For example, two speakers involved in a conversation may be found to vary the loudness of their voices in synchrony, so that they are speaking loudly or quietly at the same time, which would be evidence of synchrony. Simultaneously, the measured intensity of their voices could approach the same values.

4.4.2 Aim

The aim of this study, is to investigate if entrainment affects speakers' pause lengths. To do this, we needed to examine how pause lengths vary over time, and if the variations found in one speaker are mirrored in the variations found in the other speaker in the conversation. By looking at the variation in pause lengths over the course of the dialogue, we capture a dynamic that can not be seen from merely looking at overall mean pause lengths, as in previous section.

Our hypothesis is that we will find entrainment both in the form of synchrony and convergence, which means that we expect to find that when one speaker increases or decreases the length of her pauses, the other speaker follows, and that speakers are producing pauses of similar lengths. The basis for this assumption is that speakers need to adapt their pause lengths to each other for turn taking to function easily. We also propose that the pauses at TRPs (Transition Relevance Places) will be the most important ones to adapt to the other speaker. If speakers have very differing expectations regarding pause lengths, it will impact the conversation in a negative way (see Section 2.2.2).

If we find evidence of entrainment with regards to pause lengths, this will be an important detail to add to what we know about turn taking, and how speakers time their utterances in conversation.

4.4.3 Method

Edlund et al. (2009) outline a method to quantitatively analyse dialogues dynamically, and a similar method has been applied by Kousidis and Dorran (2009), who refer to it as the TAMA-method (TAMA stands for Time-Aligned Moving Average). These methods are based on a moving window average, which means that for a number of data points (in the current study, pauses), an average is calculated. The window is then moved one step, overlapping all but one data points in the previous calculation, and a new average is calculated. This will provide a smoother curve, where smaller variations are evened out. The smoothness of the curve can be controlled by the number of data points included in the window. The more data points included, the smoother the curve will be. In this case, a window of 9 data points was used. The material used in this section was the data described in 3.3.1. Pauses included in this analysis were planning pauses, initial pauses and TRP-pauses without turn change.

This averaging method is applied to each of the speakers' pause lengths, so that these curves can be compared. This raises a problem relevant to many studies of communication: most of the time, speakers do not speak simultaneously, which means that we will only have data for one speaker at a time. How do we compare the speakers, if their data is not overlapping? Edlund et al. (2009) suggest that this problem can be solved by interpolating data. This means, that based on the data available for the speakers, interpolated data points are added so that speakers can be compared. Using interpolated curves, average pause lengths were calculated for both speakers in each dialogue; the pause lengths were measured at the data points of one speaker.

When pause length variation over time was analysed, a moving, Gaussian-shaped window of 9 data points was used. The Gaussian shape of the window gives more weight to the central values in the window, and provides a smoother curve.

To compare the pause patterns of the speakers, the correlation between the pause lengths was calculated. This was done by calculating the Pearson product-moment correlation coefficient (Pearson's r). Pearson's r outputs a value between –1 and 1, with 0 meaning that there was no correlation at all between the values. –1 means that the values are in "reverse synchrony", which means that when one is low the other is high and vice versa. 1 means that the values are in complete synchrony, changing consistently in the same way. Each Pearson value comes with a significance measure, so to be sure that two sets of data are in synchrony, the output of Pearson's r should be positive (the closer to 1 the better) and significant ($p<0.05$).

To examine if there was any convergence or divergence between pauses, we investigated whether the difference between the mean pause lengths of the two speakers grew larger or smaller over time. This was done by analysing the correlation between the pause length difference and time. If there is a positive correlation, it means that the difference becomes larger with time, which would suggest that the pause lengths are diverging. A negative correlation signifies that pause length differences become smaller over time, which indicates convergence.

4.4.4 Results

The results of the analysis of synchrony and convergence are presented in table 4.8.

Four of the dialogues (D1, D3, D4 and D6) exhibited a significant positive correlation between the pause lengths of the speakers in the dialogue, which is evidence for entrainment in the form of synchrony. Synchrony in this case means that the speakers are varying their pause lengths in the same way: their pauses becomes longer at the same time, and shorter at the same time. In dialogue 1 (see figure 4.12) we found a strong positive correlation between the pause lengths of the speakers, but we found no significant evidence of convergence.

In dialogues 3 (figure 4.13), 4 (figure 4.14)and 6 (figure 4.15) there was a moderately positive relationship between the pause lengths of the speakers. In these three dialogues we also see that

Table 4.8: Synchrony and convergence of pause lengths in dialogues 1-6

	Synchrony	Convergence
D1	0.621**	-0.223 (p=0.087)
D2	-0.635**	0.717**
D3	0.327*	-0.613**
D4	0.391**	-0.306*
D5	-0.175 p=0.132	0.059 (p=0.666)
D6	0.405**	-0.317**

*Significant correlation at the .05 level; **Significant correlation at the .01 level

there is convergence between the speakers' mean pause lengths, which tells us that the speakers' pauses are more alike in length at the end of the dialogue than at the start of the dialogue.

Dialogue 2 (figure 4.16) differs from the other dialogues in that there is strong evidence of asynchrony and divergence, which suggests that when one speakers is making longer pauses, the other is making shorter pauses, and that the mean pause lengths of the speakers differ more at the end of the dialogue than at the beginning. We will return to dialogue 2 in the discussion, and propose possible explanations for this pattern.

In dialogue 5 (figure 4.17) we found no significant correlation between the pause lengths of the speakers. Since we have proposed that pause length entrainment will be more important at TRPs (Transition Relevance Places), we decided to explore whether grouping the pauses based on whether they coincided with TRPs would affect the outcome of the analysis of this dialogue. For the pauses that coincided with TRPs, we found a significant positive correlation (r=0.466, p=0.039), whereas the pauses that did not occur at TRPs showed a significant negative correlation (r=–0.530, p=0.001).

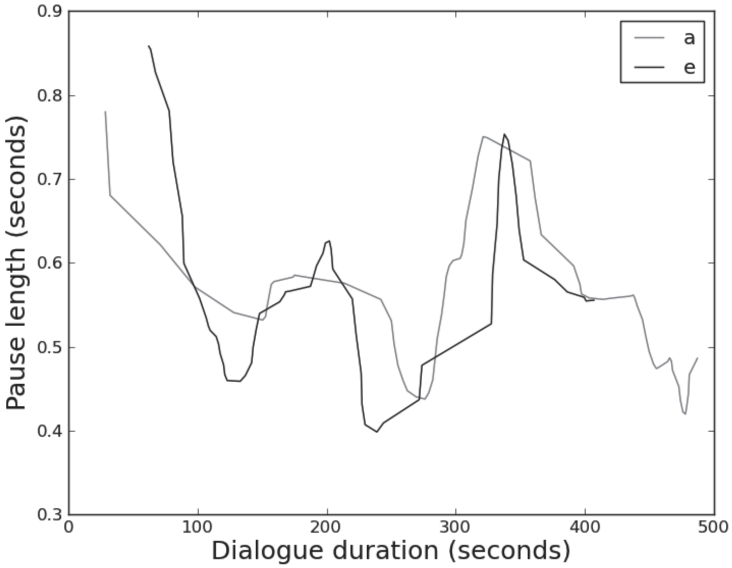

Figure 4.12: Pause length variations in dialogue 1

4.4.5 Discussion

Our hypothesis was that speakers adapt the duration of pauses to each other, and we did find support for this in our analyses. In four of the six dialogues, there were clear indications of entrainment over the course of the dialogue, and in one additional dialogue pauses that occurred at TRPs (Transition Relevance Places) showed a positive correlation. This supports the hypothesis that speakers need to adjust their pause lengths to their conversation partners, to ensure smooth turn taking. We can also conclude that in 3 of the dialogues the speakers are more similar with regards to pause lengths at the end of the dialogue than at the beginning, which shows that entrainment is an ongoing process. In dialogue 1 we did not find evidence of convergence. One possible reason for this could be that the speakers in dialogue 1 show a high degree of synchrony throughout the dialogue, and that convergence happened quickly at the beginning of the dialogue.

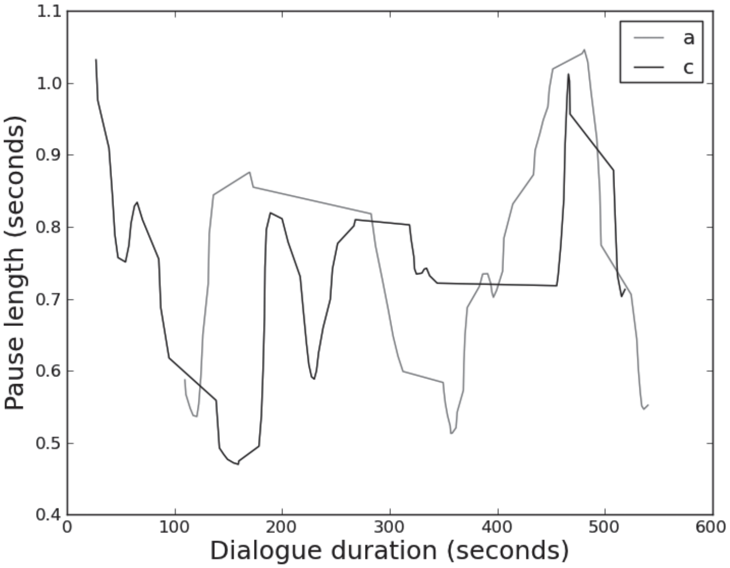

Figure 4.13: Pause length variations in dialogue 3

In dialogue 2 we found that there was evidence of reverse synchrony, which means that when one of the speakers were producing long pauses, the other speaker was producing short pauses, and vice versa. When we examined dialogue 2, we found that it differed from the other dialogues with regards to the emotional content. Beatrice and Doris spent about a third of the conversation talking about a boy who was ridiculed by his father because he had a stutter, and they both expressed anger and disgust at this. None of the other dialogues contained such high degree of emotional involvement. Therefore, we suggest that it might be the emotional arousal that negatively impacts entrainment. The connection between emotional state and entrainment is something that should be explored in future studies.

As discussed in Section 4.4, entrainment has been confirmed at different levels of communication, and this study underlines the ubiquity of this trait.

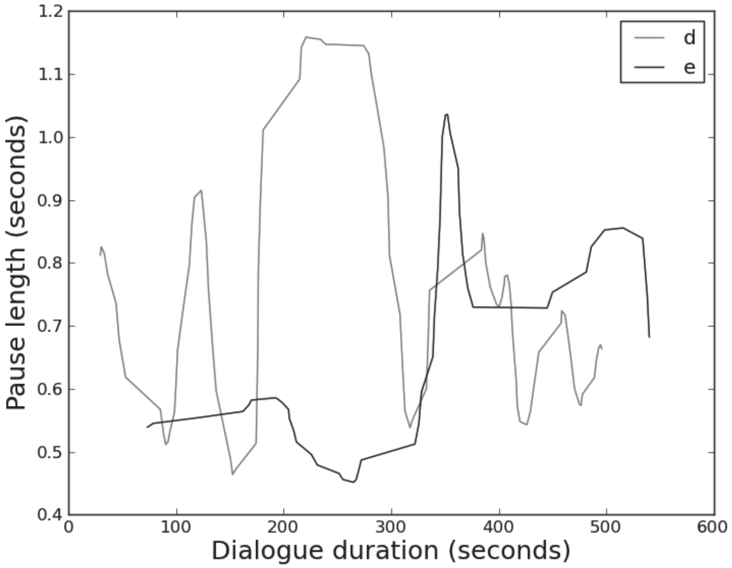

Figure 4.14: Pause length variations in dialogue 4

4.5 Periodicity of pauses

4.5.1 Introduction

Most speech-related activities seem to follow a rhythm. Syllables tend to be of somewhat equal length, and it has been suggested that we adapt to an inner "biologic clock" (Zellner, 1994). This inner sense of timing has been linked to the existence of endogenous oscillators, groups of neurons that show periodicity in their activity, in the brain (Wilson and Wilson, 2005). Speech is typically produced with 3–8 syllables per second, and this oscillation rate is seen in the theta frequency band in the brain (4–10 Hz) (Wilson and Wilson, 2005; Menenti et al., 2012). The mandible oscillates at a mean frequency of 5 Hz, which is consistent with the syllable rate (Lindblad et al., 1991).

Wilson and Wilson (2005) suggest that when speakers are engaged in conversation with each other, their internal oscillators

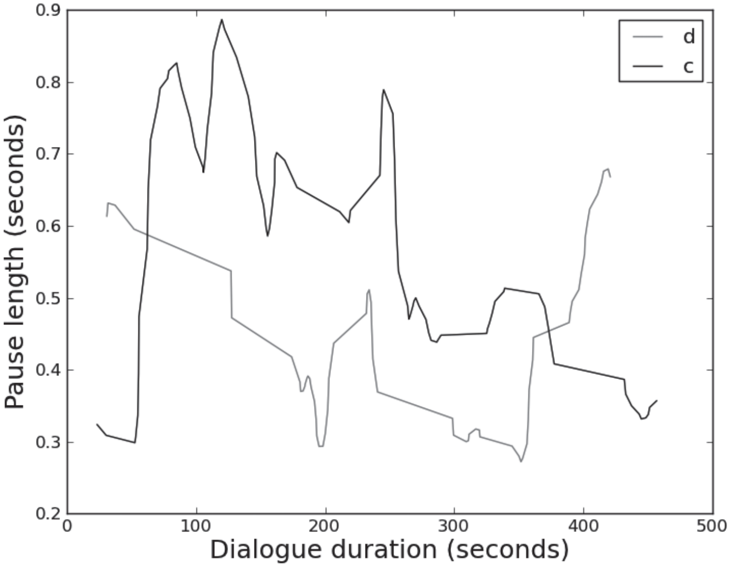

Figure 4.15: Pause length variations in dialogue 6

governing speech rate will become entrained, and that this entrainment facilitates turn changes and decreases the risk of for example two persons beginning to talk at the same time. Sacks et al. (1974) propose that speakers are able to project when another speaker will conclude her turn, and this ability to project TRPs may be connected to syntactic structure and pitch variation. Considering that we seem to have a sense of timing for the oscillations connected to syllable production rate, it is not unlikely that we also are attuned to the slower oscillations that mark breath groups and pause patterns. To discern patterns and regularities in data over time, we use time series analysis. It has been utilised in various fields, such as biology and economics, to recognise trends and to predict future events.

Merlo and Barbosa (2010) employed time series analysis when examining regularity in hesitation phenomena such as pauses, fillers, prolongations, repetitions, false starts and blocks. These

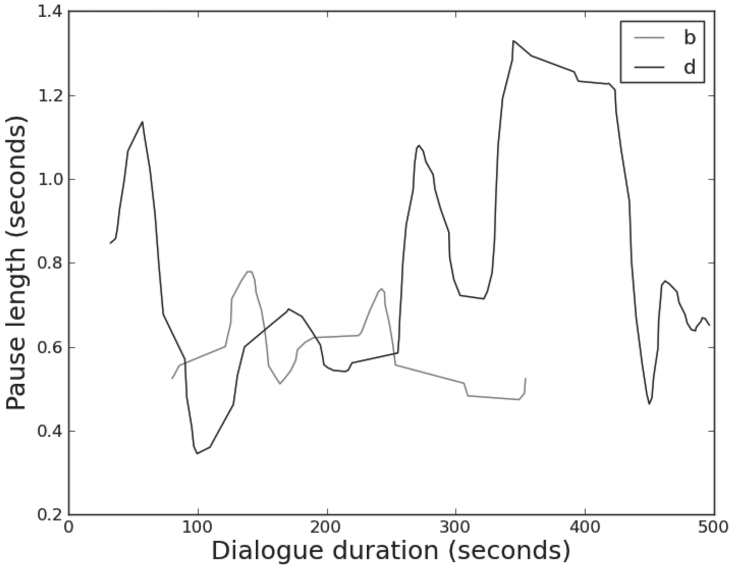

Figure 4.16: Pause length variations in dialogue 2

were found to oscillate periodically at different frequencies, and an average of three oscillations per speech sample was found, with cycles ranging from 1.9 to 77.6 seconds. Merlo and Barbosa did not differentiate between different types of hesitations. The presence of different oscillating frequencies suggests that there are several underlying processes involved in hesitation phenomena.

4.5.2 Aim

The aim of this study is to investigate whether pauses occur with regular intervals in dialogues, and in that case, which frequencies they oscillate at. Our hypothesis is that pauses do occur at somewhat regular intervals, since this will aid speakers in assessing for example when a turn is likely to end. If we find that pauses occur with regular intervals throughout dialogues, this is will be evidence that speakers' tend to prefer turns and parts of turns of

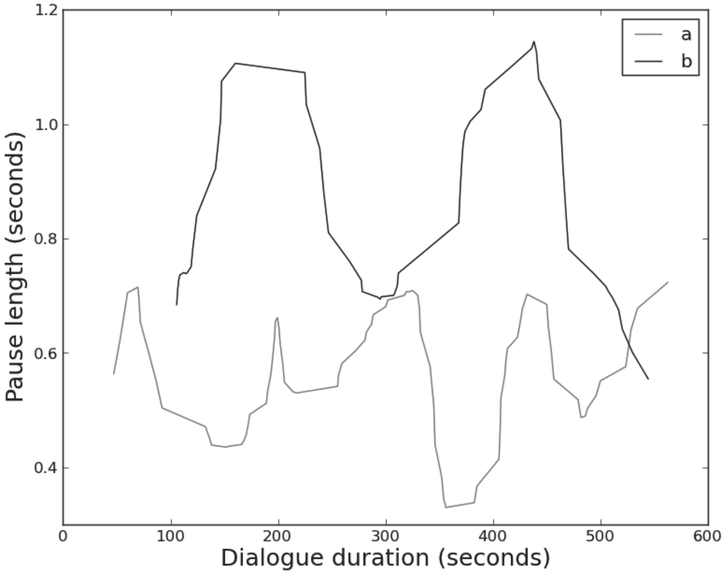

Figure 4.17: Pause length variations in dialogue 5

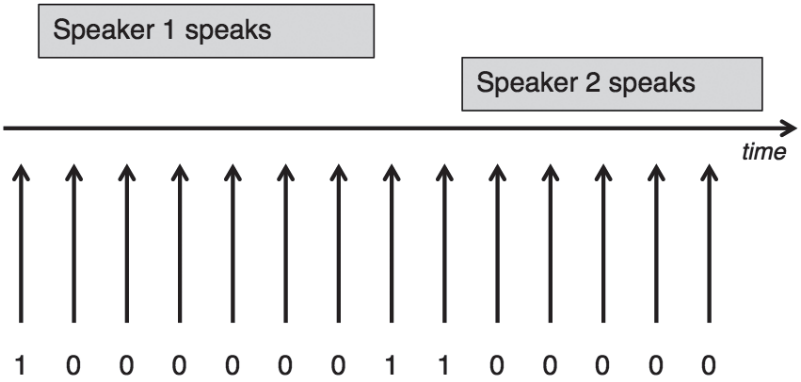

Figure 4.18: Sampling of dialogues

certain lengths, and we may be able to use that information when creating models for turn taking, both in humans and in dialogue systems.

4.5.3 Method

The six spontaneous dialogues described in section 3.3.1 were sampled at 100 Hz, which means that 100 samples were taken every second. At each sampling point, the script checked whether there was silence or speech at that time in the dialogue. A visual overview of the sampling process in presented in image 4.18.

No distinction was made between the two speakers in each dialogue, since the focus was on the overall structure of the dialogue, and not the individuals speakers. All pauses types were included in the material. Samples consisting of speech were labeled with 0′s, and samples consisting of pauses were labeled with 1′s. It might seem counterintuitive to label pauses as 1′s instead of 0′s, but we choose this since the pauses are the focus of the analysis. For each dialogue, a text file consisting of a string of 1′s and 0′s was generated, providing approximately 60000 samples per dialogue. Sampling of the data was done in Python.

Before performing a Fast Fourier Transform, the values were detrended, to remove linear trends. This means that slow linear increases or decreases in cycle lengths are excluded. Following this, the data was analysed using a Fast Fourier Transform. A Fourier transform decomposes the signal and expresses the outcome as a series of frequencies of waves that make up the signal, and the relative strength of these frequencies. The signal does not have to be periodic. The Fast Fourier Transform is calculated in a different and faster way than an original Fourier transform, but will yield the same results. Even though Fourier Transformation produces results that show more spectral energy at certain frequencies, it is important to ascertain that the variations are not simply due to noise in the data. For significance testing, a variation of the Fisher test developed by Bølviken was used (Bølviken, 1983; Jakobsson, 2013). This test is constructed to identify significant periodicities at unknown frequencies (Jakobsson, 2013).

We chose to focus on frequencies below 0.67 Hz, consequently excluding periodicities shorter than 1.5 seconds. This was done to make sure that the periodicities detected are not caused by pause length. If we had included higher frequencies, there is a risk that the data would not reflect the distance between pauses but rather the length of the pauses themselves.

4.5.4 Results

The results of the Fast Fourier Transform indicate that all five dialogues exhibit statistically significant periodic oscillations with regards to pauses. This substantiates the hypothesis that pauses occur at regular intervals throughout the dialogues.

In table 4.9, the strongest periodicities in each dialogue are presented, in descending strength. The four strongest periodicities for each dialogue, or all periodicities with a p-value < 0.0001 (according to the Bølviken test) are shown. Frequencies in bold text are multiples of the same cycle.

There are some apparent similarities between the dialogues: in three of the dialogues (dialogue 1–3), a periodicity at 0.06 Hz is present, which equals a cycle of 16.7 seconds. In the other three dialogues we do not find that exact frequency, but we do find a

Table 4.9: Pause periodicities in dialogues, in Hertz. (** = p < 0.0001)

D1	D2	D3	D4	D5	D6
0.61 (p=.0001)	0.29**	0.06**	0.06**	**0.14****	0.22**
0.14 (p=.0005)	0.08**	0.33**	0.33**	0.59**	0.06**
0.26 (p=.0012)	0.50 (p=.0011)	0.10**	0.10**	0.19**	0.37**
0.19 (p=.0116)	0.32 (p=.0035)	0.23**	0.23 (p=.0001)	0.44**	0.45**
		0.06**		**0.56****	0.58**
		0.17**			
		0.12**			

periodicity of 0.14 Hz (7.1 seconds) in dialogue 4, and 0.08 Hz (12.5 seconds) in dialogue 2 and 5. The cycle of 7.1 seconds found in dialogue 4 is roughly half of the cycle found in dialogues 1–3, and the 12.5 second cycle is in the neighbourhood of the 16.7 second cycle in dialogues 1–3. In dialogue 2 and dialogue 4 there are multiples of the same cycle.

4.5.5 Discussion

Even though the conversation topics and speakers vary, there are unifying features in the periodicities of pauses. The longest cycles in the dialogues vary between 7.1 and 16.7 seconds, and these cycles are unlikely to be connected to breath groups, as these tend to be around 3–6 seconds (McFarland, 2001). It seems more likely that these cycles are connected to turn lengths, which suggests that the pauses underlying these cycles are those associated with turn changes (TRP-pauses with turn-change and initial pauses). In dialogue 1, mean turn lengths are around 6 seconds, and the second strongest frequency in this dialogue is 0.14 Hz (a cycle of 7.1 seconds). Dialogue 2 has differed from the other dialogues in earlier analyses, for example in the lack of entrainment. Mean turn lengths in dialogue 2 differ markedly between speakers: 6.6 seconds for Beatrice and 11.3 seconds for Doris. The second strongest oscillation in dialogue 2 is 0.08 Hertz, which equals a cycle of 12.5 seconds. Since Doris speaks quite a lot more than Beatrice in dialogue 2, it is reasonable to believe that the 12.5 second cycle found in the dialogue is connected to her turn lengths. In dialogue 3, mean turn length for both speakers was 16.4 seconds, which suggests that the 0.06 Hz (16.7 seconds) oscillation found in the dialogue is related to turn length.

The mean turn lengths in dialogue 4 are 9.1 seconds (Emma) and 13.7 seconds (Doris) respectively. The 10 second cycle (0.10 Hz), which is the third strongest in the dialogue, could be connected to Emma's turn lenghts, while Doris' turn lengths are closest to the 0.06 Hz frequency. In dialogue 5 mean turn lengths are slightly less than 7 seconds, which fits with the strongest frequency in the dialogue (0.14 Hz). Finally, in dialogue 6, mean turn length for Doris is 5.8 seconds and 9.1 seconds for Cilla. The second strongest frequency in the dialogue is 0.06 Hz (a cycle of 16.7 seconds) which is roughly the double turn lengths for the speakers. To conclude, we propose that turn length in dialogues can be

Cycles associated with breathing should be in the interval 0.15–0.35 Hz, and there are several such candidates in the dialogues: 0.19 Hz and 0.26 Hz in dialogue 1, 0.29 Hz and 0.32 Hz in dialogue 2, 0.17 Hz and 0.22 Hz in dialogue 3, 0.23 Hz and 0.33 Hz in dialogue 4, 0.19 Hz in dialogue 5, and 0.22 Hz in dialogue 6.

Finally we have some higher frequencies present in some of the dialogues: 0.61 Hz in dialogue 1, 0.50 Hz in dialogue 2, 0.55 Hz, 0.59 Hz and 0.64 Hz in dialogue 3, 0.59 Hz and 0.56 Hz in dialogue 5, and 0.45 Hz and 0.58 Hz in dialogue 6 . These show that shorter cycles are present in the dialogues, ranging from 1.6 seconds to 2.2 seconds. In the dialogues, we find that the mean durations of speech stretches that are delimited by pauses range from 1.5 seconds to 2.1 seconds, making them suitable candidates for the underlying higher frequencies. The analysis of periodicities tells us more than if we would merely look at average length for different types of speech stretches in dialogues. When we look at for example mean length of turns, we do not know how often the turns actually are of that length; it could be that half the time they are twice the mean length, and half the time they are half of the mean length. With time series analysis, we can establish that a speech stretch of a certain length regularly occurs over the course of the dialogue.

In the current analysis all pauses were treated as one group. Analysing different types of pauses separately might shed more light on the periodicities in dialogues, and might make it possible to more confidently associate different frequencies with different underlying structures in speech.

4.6 Syntactic context of different pause types

4.6.1 Introduction

When we pause, we choose a place that is suitable with regards to the reason why we're pausing. Simply put: when we pause, but are not finished with our turn, we pause in the middle of something, to make it clear that we're continuing. When we pause because we are finished, or because we want to check if the other person wants to speak, we pause at a TRP to signal to the other speaker that they make the turn if they wish.

Pauses often occur at sentence boundaries, and also between clauses. Further, pauses are also common before content words (Strangert, 1993). Zellner (1994) reports that studies have shown that pauses are more frequent and longer between words that have lower cohesion, whereas they are less frequent and shorter between words that are strongly connected. Megyesi and Gustafson-Čapková (2002) found that pauses occur mainly at turn shifts (28 %), but that they also occur at phrase breaks: 16 % in front of noun phrases, 10 % in front of adverb phrases, 10 %

in front of conjunctions and 9 % in front of prepositional phrases. It is common that this type of pause occurs before a content word, or after a discourse marker such as "and" or "but" (van Donzel and Koopmans - van Beinum, 1996). Strangert (2003) found that pauses can be used to signal focus. However, the level of emphasis does not seem to correlate with pause length.

4.6.2 Aim

The aim of this study is to explore the relationship between parts of speech and pauses. To investigate the syntactic context of pauses, half of the spontaneous conversation material described in 3.3.1 was analysed. The specific questions we were focusing on were whether the parts of speech surrounding a pause differ between pause types, and in which way. The pauses analysed in this study are the pauses that are not adjacent to turn changes: TRP-pauses without turn changes, and planning pauses. It is important to note that this study will have been affected by pause annotations. Since pauses are categorised with the question "could the speaker have finished here?" (see section 2.2.1), we expect that the categorisation will depend partly on lexical and syntactical context. Therefore, the question is not "is the context different?" but "how is it different?". Our hypothesis is that the parts of speech before and after pauses will vary depending on the pause type, and if we hope that this information may be helpful in differentiating between pause types in models of turn taking.

4.6.3 Method

The material described in section 3.3.1 was automatically tagged using the hunpos open source Hidden Markov Models tagger (Halácsy et al., 2007). The hunpos tagger is a reimplementation of the Trigrams'n'Tags part-of-speech tagger developed by Thorsten Brandts (Brants, 2000). In the current study an existing Swedish model for the hunpos tagger was used, and that model is trained on the Stockholm-Umeå Corpus (Megyesi, 2009).

To avoid some problems that can arise when using a tagger trained on written Swedish on spoken Swedish, a few words were changed in the data from the speech variant to the written variant. This was done manually, since for example "å" can mean both "och" (and) and "att" (to). After having tagged the data automatically, the words immediately preceding and succeeding pauses were extracted, along with their tags. This data was then inspected and obvious errors corrected. The most common error was words being tagged as proper nouns, despite being other parts of speech. On average 20 % of words were tagged as proper nouns, whereas the correct amount is 0 %. Further, interrupted words such as "b-" or "en-" were removed, together with hesitation markers/fillers such as "eem" or "öö".

4.6.4 Results

Results are presented in table 4.11 and figures 4.19 and 4.20. The numbers displayed are percentages of all parts of speech in that position. The part-of-speech tags used in the material are the tags used in the Stockholm-Umeå Corpus, and these are presented in table 4.10. The table is reproduced from the Stockholm-Umeå Corpus manual (Gustafson-Capková and Hartmann, 2006).

What stands out the most in the data presented in table 4.11 is conjunctions after pauses. Conjunctions make up 40.6 % of the parts of speech that follow TRP-pauses not at turn changes, whereas after planning pauses, they are not at all as common (they make up 8.6 % of the parts of speech following planning pauses). The difference in distribution of parts of speech after pauses is statistically significant (Wilcoxon signed rank test: z=–2.433, p=0.015). Before pauses, there is less variation in the distribution of parts of speech between the two types of pauses, and no part of speech stands out as much more common than the others. The difference in distribution of parts of speech before pauses is not statistically significant (Wilcoxon signed rank test: z=–0.335,

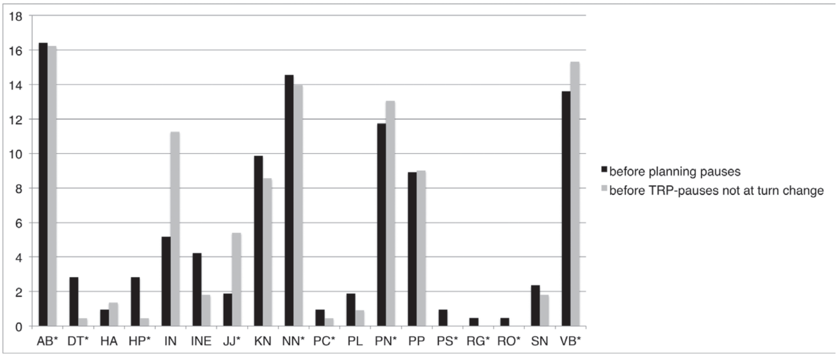

Figure 4.19: Parts of speech before pauses within speakers' turns

p=0.738).

The three most common parts of speech before and after pauses are presented below, with examples of utterances from the material. Utterances are displayed in Swedish with an accompanying English translation.

It is evident from the results in 4.12 that the same three parts of speech are most common both before planning pauses and TRP-pauses without turn change: adverbs, nouns and verbs. This is consistent with the fact that the distribution of parts of speech before pauses are not significantly different statistically.

We can see in table 4.13 that the most common part of speech after pauses that occur within turns at TRPs are conjunctions, followed by pronouns and adverbs. Conjunctions and adverbs are parts of speech that are often used as discourse markers, which can be used to introduce new segments in spoken conversation (Fraser, 1999). An example of this from the material is the utterance "men i alla fall om" ("but anyway if"), which follows a pause at a TRP (this example is not given in table 4.13. Often the conjunction is used to connect what is said after the pause to what was said before the pause (and thereby joining turn constructional units), as in the example in 4.13: "och hur hela situationen ska gå till" ("and how the whole situation should be").

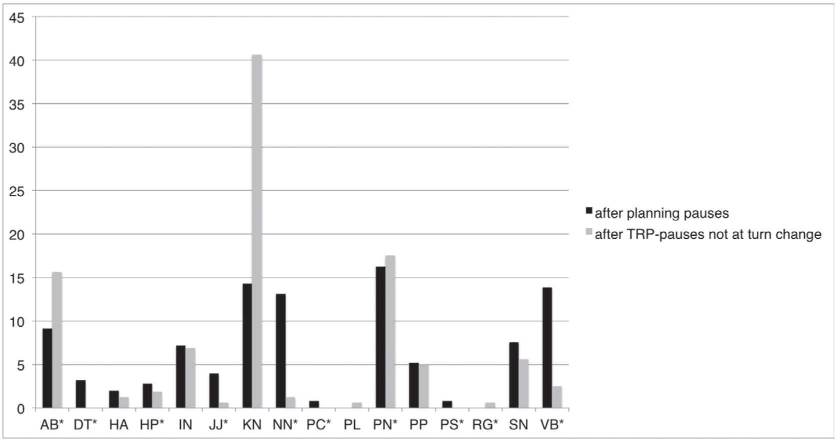

Figure 4.20: Parts of speech after two types of pauses within speakers' turns

4.6.5 Discussion

The parts-of-speech that occur before and after of pauses do not differ as much as one could expect, and we do not find evidence for our hypothesis that the contexts of different pause types would be clearly different. Based on what we know about how the pauses were categorised, this is somewhat surprising: pauses were categorised as occurring or not occurring at TRPs based on what transpires before the pause. The question was whether the speaker could have finished at the pause, or if it was evident that the speaker was going to continue speaking after the pause. If we look at the examples of utterances in table 4.12, the traces of categorisation seem evident: compare for example the utterances ending with nouns. Before the pause that is not at a TRP, we see the phrase "men om nu pappan" ("but if now the father") which seems incomplete and gives the impression that speaker will expand the utterance. We can compare this to the utterance "ursäkta att jag gäspar dig i ansiktet" ("sorry about yawning you in the face") which is found before the pause at a TRP; this utterance is complete in itself and does not demand a continuation. Evidently, examining only the last part of speech before the pause is not enough to capture this difference. The pauses in this material are not followed by content words to the extent found in previous research (van Donzel and Koopmans - van Beinum, 1996; Megyesi and Gustafson-Čapková, 2002).

4.7 Summary

In this chapter, we have presented two methodological studies and four studies on pause production. The methodological studies related to inter-rater reliability in pause categorisation, and automatic versus manual annotation of pauses. The results of the categorisation of pauses showed substantial agreement between raters, and we therefore conclude that our categorisation flowchart is adequate. The comparison of automatic and manual methods for pause annotation showed that automatic methods are not yet as good as manual annotations, and that to get reliable data on pauses, manual annotations are necessary. The four studies on pause production are all based on the PauDia corpus.

We found that speakers tend to be rather consistent with regards to the ratio of planning pauses and TRP-pauses without turn changes, and this is a preference that does not seem to change depending on the conversational partner. We had hypothesised that the mean duration of TRP-pauses without turn changes would be longer than planning pauses overall, but this hypothesis did not hold. Rather, we found that pause lengths vary quite a lot, which was further confirmed when we investigated entrainment of pause lengths in dialogues. Our analysis of pause length entrainment indicates that speakers vary their pause lengths to a large degree, but that these variations are synchronised between speakers in a dialogue.

In our third study we investigated the periodicity of pauses in dialogues, and found that pauses do occur regularly throughout dialogues, and that the regularities can be connected to different types of speech stretches, such as interpausal units, breath groups and turns. Finally, we examined the syntactic context of two different types of pauses, and found that the parts-of-speech that occur just before and after pauses are not reliable predictors of pause types.

Table 4.10: Stockholm-Umeå Corpus part-of-speech categories

Code	Swedish category	Example	English translation
AB	Adverb	inte	Adverb
DT	Determinerare	denna	Determiner
HA	Frågande/relativt adverb	när	Interrogative/Relative Adverb
HD	Frågande/relativ determinerare	vilken	Interrogative/Relative Determiner
HP	Frågande/relativt pronomen	som	Interrogative/Relative Pronoun
HS	Frågande/relativt possessivt pronomen	vars	Interrogative/Relative Possessive
IE	Infinitivmärke	att	Infinitive Marker
IN	Interjektion	ja	Interjection
JJ	Adjektiv	glad	Adjective
KN	Konjunktion	och	Conjunction
NN	Substantiv	pudding	Noun
PC	Particip	utsänd	Participle
PL	Partikel	ut	Particle
PM	Egennamn	Mats	Proper Noun
PN	Pronomen	hon	Pronoun
PP	Preposition	av	Preposition
PS	Possessivt pronomen	hennes	Possessive
RG	Grundtal	tre	Cardinal number
RO	Ordningstal	tredje	Ordinal number
SN	Subjunktion	att	Subjunction
UO	Utländskt ord	the	Foreign Word
VB	Verb	kasta	Verb

Table 4.11: Parts of speech before and after pauses

	Before planning pauses	Before TRP-pauses w/o turn change	After planning pauses	After TRP-pauses w/o turn change
Adverb	16.4 %	16.2 %	9.1 %	15.6 %
Determiner	2.8 %	0.45 %	3.2 %	0 %
Interrogative/relative adverb	0.9 %	1.4 %	1.9 %	2.8 %
Interrogative/relative pronoun	2.8 %	0.45 %	2.8 %	1.9 %
Infinitive marker	4.2 %	1.8 %	0 %	0 %
Interjection	5.2 %	11.3 %	7.1 %	6.9 %
Adjective	1.9 %	5.4 %	4.0 %	0.6 %
Conjunction	9.9 %	8.6 %	14.3 %	40.6 %
Noun	14.6 %	14,0 %	13.1 %	1.25 %
Participle	0.9 %	0.5 %	0.8 %	0 %
Particle	1.9 %	0.9 %	0 %	0.6 %
Pronoun	11.7 %	13.1 %	16.3 %	17.5 %
Preposition	8.9 %	9.0 %	5.2 %	5 %
Possessive pronoun	0.94 %	0 %	0.8 %	0 %
Cardinal number	0.47 %	0 %	0 %	0.7 %
Ordinal number	0.47 %	0 %	0 %	0 %
Subjunction	2.3 %	1.8 %	7.5 %	5.6 %
Verb	13.6 %	15.3 %	13.9 %	2.5 %

Table 4.12: Most common parts of speech before pauses

Part of speech (percentage)	Example from the material	English translation	Pause type
adverb (16.4 %)	"fast det kanske man inte"	"but maybe you do not"	planning pause
noun (14.6 %)	"men om nu pappan"	"but if now the father"	planning pause
verb (13.6 %)	"ja precis och att kunna ta"	"yes exactly and to be able to take"	planning pause
adverb (16.2 %)	"jag tänkte säja det men det var det inte"	"I was going to say that but that wasn't it"	TRP-pause w/o turn change
verb (15.3 %)	"ja jag kommer inte ihåg vad det hette"	"well I don't remember what it's called"	TRP-pause w/o turn change
noun (14.0 %)	"ursäkta att jag gäspar dig i ansiktet"	"sorry about yawning you in the face"	TRP-pause w/o turn change

Table 4.13: Most common parts of speech after pauses

Pause type	Part of speech	Example from the material	English translation
planning pause	pronoun (16.3 %)	“jag har aldrig haft det problemet med”	“I’ve never had that problem with”
planning pause	conjunction (14.3 %)	“och så blir man så”	“and then you become so”
planning pause	verb (13.9 %)	“har du ingen luft så”	“if you have no air then”
TRP-pause w/o turn change	conjunction (40.6 %)	“och hur hela situationen ska gå till”	“and how the whole situation should be”
TRP-pause w/o turn change	pronoun (17.5 %)	“vi trollar inte liksom”	“we don’t do magic you know”
TRP-pause w/o turn change	adverb (15.6 %)	“kanske inte fråga direkt så men”	“maybe not ask right out but”

Chapter 5

Perception of pauses

Silence is an answer too.

Perception and understanding of speech and communication is highly dependent on context. Pauses can be used to alter the context of spoken words, and as a result they may also modify the meaning of the words. In this chapter, we will investigate how pauses of varying lengths can affect understanding and processing of spoken language.

5.1 The effect of pause length on perception

The length and placement of pauses will affect the perception of what is being said. For example, the length of a pause will determine whether or not the pause is perceived as a sign of trouble in the conversation. Roberts and Francis (2013) investigated gaps (which we call TRP-pauses with turn change) ranging from 200 to 1200 ms, to conclude when the gaps start to become troublesome. Participants listened to short telephone conversations between friends, each containing a request, invitation or assessment. The request, invitation or assessment were followed by a positive reply by the conversation partner, but the gap before the reply differed in length between 200 and 1200 ms. The study participant were asked to rate their perception of how positive the reply was. No significant differences were found between the 200 ms gap and any of the gaps of a duration shorter than 600 ms. The difference in ratings between the 200 ms gap and the 600 ms gaps were significant, and when the gaps were longer than 600 ms, the ratings of willingness to comply with the request began to drop. Between 700 ms and 800 ms, there was a significant drop in ratings of willingness, but after 900 ms the ratings flattened out. This suggests that when gaps are as long as 600 ms and longer, they take on a social communicative meaning; the gap is seen as too long for typical production and utterance planning.

The effect of gap length may differ depending on the language spoken. Roberts et al. (2011) compares the impact of gaps of different lengths in English, Italian and Japanese. As in the study described above, participants listened to short telephone conversations between friends, each containing a request or an assessment and a positive reply. Three different gap lengths were analyzed: 0 s, 600 ms and 1200 ms. In all three languages, the person replying was considered less positive when the gap length was longer. However, language differences were evident in results: Japanese listeners rated the speakers as more agreeable over all. The responses to the 600 ms gap length reveal statistically significant differences in how agreeable the speaker was rated; the Japanese listeners were the most tolerant of the 600 ms gap, while the Italian listeners rated the speaker after the 600 ms gap as much less positive. Roberts et al. (2011) propose that while a longer silence before a reply is universally perceived as an indication of trouble in the conversation, the scale may differ slightly between different languages.

Pause length also has an effect on the perceived emotional state of the speaker (Tisljár Szabó and Pléh, 2014). Longer pauses will lead to the speaker being perceived as sadder and more scared, whereas shorter pauses will cause the speaker to be perceived as happier.

Pauses in speech have affect the speech processing of the listener. MacGregor et al. (2010) found that silent pauses will make listeners more ready to process something unexpected. Further, subjects are more likely to recall words preceded by a silent pause. Reich (1980) found that the placement of pauses will influence the ability to recall a sentence. In the study, Reich (1980) distinguished between "grammatical" ("The man the boat rocked PAUSE threw his spear at the whale.") and "non-grammatical" ("The man the boat rocked threw his PAUSE spear at the whale.") pauses, and discovered that sentences with non-grammatical pauses were significantly harder to recall.

5.2 Aim

The aim of this study is to investigate how pauses of typical and atypical lengths affect speech perception and memory. Three types of sentences were used: sentences with no pause, sentences with a 500 ms pause and sentences with a 4000 ms pause (the stimuli are described in more detail in section 5.5.2). The shorter pause was chosen to be of typical length for speech production, and short enough as to not be perceived as a sign of trouble in the conversation. The longer pause is chosen to be long enough to disturb the processing of the auditory stimuli.

Our hypothesis is that the shorter pause will not negatively influence perception or memory, but that the longer pause will disturb the processing of the complete sentence and the ability to accurately remember it.

5.3 Eye tracking and the visual world paradigm

To study the processing of sentences, we used eye tracking. Eye tracking has been used extensively to analyse the processing of language. The experiment setup was based on the visual world paradigm, which has been used since the mid 1970's as a research tool to study perception in real-time. Since the mid 1990's it is widely used in psycholinguistics. The most common version of the visual world paradigm consists of auditory stimuli and visual stimuli, often presented on a computer screen. While the stimuli is being presented, the subject's gaze and eye movements are being recorded with an eye tracker. The task may be to manipulate the visual stimuli in some way, or to simply listen to the auditory stimuli and look at the visual stimuli (Huettig et al., 2011). The method is based on the time-locking hypothesis, which states that what the subject is looking at reflects their cognitive and linguistic processing (Holmqvist et al., 2011). To conclude where a person is looking, fixations are identified. During a fixation the eye is relatively still in a position. Fixation are mixed with saccades when the eye is moving and the gaze is shifting to a new location, and during saccades no visual information is transmitted to the brain. Fixations are typically around 200–300 ms long, but may vary from tens of milliseconds to several seconds (Holmqvist et al., 2011).

5.4 Subjects

8 subjects took part in the study. They were recruited through flyers posted at Lund University. All subjects were female and native speakers of Swedish, and had unimpaired hearing. The subjects were aged between 18 and 61 (mean age 29.1 years). One subject had to be excluded from the analysis, as the data recorded in that session was incomplete.

5.5 Method

The eye-tracking study was carried out at Lund University Humanities Lab [1]. Each subject was given an introduction and in-

[1] http://www.humlab.lu.se/en/

structions, and they were also given an information text and consent form to read through before the test.

The test started out with a calibration of the eye tracker, which also gave the subjects some time to familiarise themselves with the test situation. The calibration was repeated between three and five times, depending on the outcome. The subjects then moved on to a short on-screen information text, before starting the actual test. During the test the test leader sat next to the subject, but stayed quiet and non-involved unless a subject needed help to, for example, remember which button to press to move to the next screen in the test.

The test was made up of 60 items, which consisted of auditory and visual stimuli. These are described in detail in section 5.5.2. The subjects' task was to listen to the auditory prompt and click on the image that best corresponded to what they heard. The stimuli were presented in a randomised order.

After the test, the subjects were given an explanation of the objectives of the study (this could not be done beforehand as that would most likely impact the results). They were then asked if they consented to having their data included in the study. If they accepted they were asked to sign the consent form they had previously read through. All subjects consented to have their data included in the study. Each subject received a movie ticket to compensate for their time (approximate value 10€).

5.5.1 Eye tracker and software

The eyetracker used was an SMI RED250, which is a remote eye tracker positioned below the computer screen. The sampling rate of the eye tracker is 250 Hz. The experiment was set up in Experiment suite 360, a software developed by SMI.

5.5.2 Auditory and visual stimuli

60 auditory stimuli were created for the test from recordings made by one speaker. All auditory stimuli were in Swedish. Each

spoken prompt consisted of three parts spliced together, for example: “click on”, “the dotty” and “chair”. The first part, “click on” was the same in all spoken prompts, but the adjectives and nouns varied.

The auditory prompts were of three different types. In group 1, the auditory prompt did not contain any pauses, whereas in group 2 and 3 a pause was inserted between the adjective and the noun in each prompt. In group 2, the pause was 500 ms long, and in group 3, the pause was 4000 ms long.

Each spoken prompt was accompanied by four images. All images where black and white drawings, depicting inanimate objects such as clothes and furniture. The images used were based on open clipart, with added patterns. Black and white drawings of inanimate objects were used to avoid images that might skew the attention of the test subjects. For example, using the color red might make the subjects look more at that image regardless of the auditory prompt, or a picture of a cute dog might distract the subjects.

The images consisted of one target image, which was the image described in the auditory prompt, one competitor image and two distractor images. An example of the visual stimuli can be seen in figure 5.2 . The competitor image shared either the adjective or the noun with the target image. For example, when the target image depicted a dotted chair, the competitor image could be either a dotted hat (as in image 5.2) or a black chair.

In eye tracking research, areas of interest (AOIs) are used as a tool to analyse eye tracking data (Holmqvist et al., 2011). Each region in an image that is of interest to the investigation is defined as an area of interest, and the parts of the images not defined as AOIs are “whitespace”. In this study, each drawing in the visual stimuli, for example the chair, the hat, the pen and the coat hanger in figure 5.2 , were defined as an AOI, resulting in four separate AOIs.

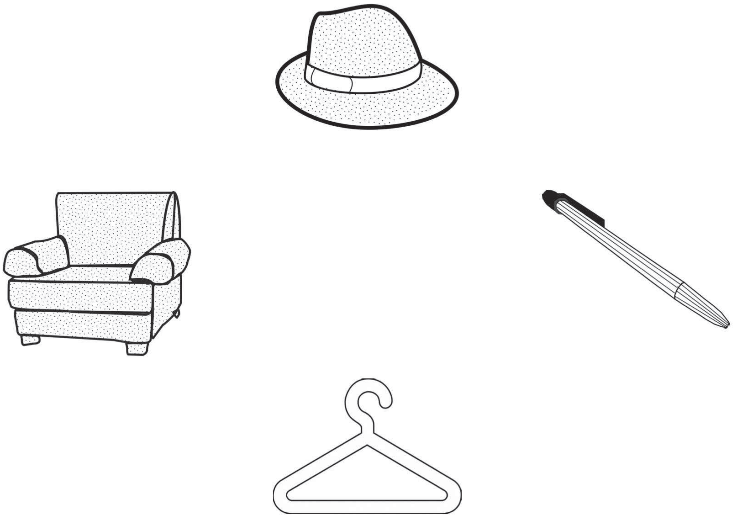

Figure 5.1: Example of visual stimuli

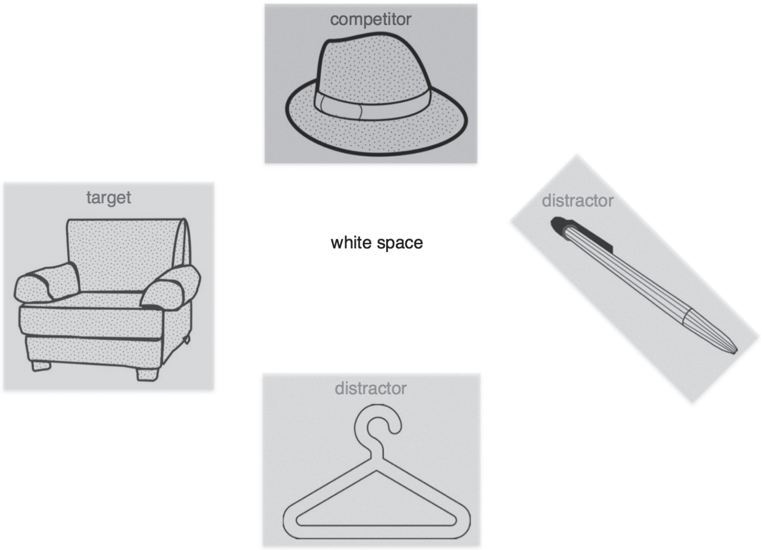

Figure 5.2: Areas Of Interest outlined in the visual stimuli

5.5.3 Memory test and questionnaire

After the main task, subjects were given a memory task, where they were asked to specify whether or not they had heard certain sentences during the task. The sentences were presented in written form. The memory task consisted of 36 sentences: 18 which had been included in the original task, and 18 that the subjects had not heard before. The sentences that had not been heard before followed the same form as the sentences that had been heard during the experiment. Subjects were also asked some background information regarding age, linguistic background etc.

5.6 Results

Analysis was based on eye movements, mouse clicks and answers to the memory task.

5.6.1 Gaze behaviour

In our experiment, we expect the presence or absence of a pause, and the length of the pause to have an effect of the processing of the spoken stimuli. To investigate this, we examined three aspects of fixation duration: the duration of the first fixation after the onset of the final word in the stimuli sentence, the duration of the first fixation on the target image after the onset of the final word in the spoken sentence, and overall fixation duration. The first fixation after the onset of a stimuli is directly connected to the very first processing of the stimuli, and it has a special status in eye tracking research. Longer fixation durations have generally been linked to "deeper and more effortful cognitive processing" (Holmqvist et al., 2011), and we expect longer fixation durations to occur when subjects are processing the sentences with the longer pauses, since these pauses are disrupting the sentences in an atypical way.

When investigating fixations connected to auditory stimuli, it is important to consider that perception of the stimuli and planning of eye movements take some time, and therefore we have looked at fixations starting 200 ms after the onset of the target noun. 200 ms was chosen since this is the approximate time it takes to plan an eye movement (Holmqvist et al., 2011).

First fixation after onset of stimuli

The first fixation after the onset of the target noun was examined.

Table 5.1: First fixation duration after onset of target noun

Sentence type	Mean duration fixation (Standard deviation of duration)
No pause	202 ms (202 ms)
500 ms pause	351 ms (328 ms)
4000 ms pause	433 ms (527 ms)

Means and standard deviations are presented in table 5.1. The analysis of variance revealed a significant difference (F = 5.77, p = 0.004). The mean duration of the first fixation after the onset of the target noun is shortest in the sentences with no pause, and longest in the sentences with the 4000 ms pause. However, since we are looking at the first fixation it is possible that the difference in fixation duration is a result of different images being fixated in the different conditions. Indeed, in table 5.2 we can see that the area of interest of the first fixation varies between the conditions.

Table 5.2: Area of interest of the first fixation

Sentence type	Target	Competitor	Distractor	White space
No pause	16	10	15	21
500 ms pause	33	17	9	6
4000 ms pause	31	7	2	5

In the sentences with no pause, the first fixation after the onset of the target noun is on the target image in 25 % of the cases.

When the target noun is preceded by a 500 ms pause, the first fixation after the onset of the target noun is on the target image in 51 % of the cases, and when the target noun is preceded by a 4000 ms pause, the first fixation is on the target in 69 % of the cases. This suggests that the shorter first fixations in the sentences with no pause could be influenced by the fact that fewer of the fixations are on the target image.

First fixation on target area of interest

In the previous section we investigated the fixation duration of the first fixation after the onset of the target noun. The fixation duration varied, but we could not rule out that the area of interest being fixated was influencing the fixation durations. Therefore, we also investigated the first fixation on the target image after the onset of the target noun, and the mean values and standard deviation values are presented in table 5.3.

Table 5.3: First fixation duration on target after onset of target noun

Sentence type	Mean duration fixation (Standard deviation)
No pause	345 ms (244 ms)
500 ms pause	430 ms (461 ms)
4000 ms pause	496 ms (589 ms)

The mean duration of the fixations are longer when preceded by a 4000 ms pause than by a 500 ms pause or no pause. However, these differences were found to be not significant (F= 1.34, p = 0.27).

Average fixation duration

After focusing on the first fixation after the onset of the target word, we also wanted to examine the mean duration of all fixations during each trial, since it is possible that the change in fixation duration that was seen in the first fixation after the onset of the noun will be present already during the pause.

Table 5.4: Mean fixation duration and standard deviation during entire trial

Sentence type	Mean duration fixation (Standard deviation)
No pause	391 ms (460 ms)
500 ms pause	524 ms (581 ms)
4000 ms pause	872 ms (750 ms)

The results (presented in table 5.4 and figure 5.3) show that when hearing the sentences with the longer pauses, the mean duration of fixation was 872 ms, which is longer than when hearing the two other types of sentences. The mean duration of fixations was 524 ms when hearing the sentences with the shorter pause, and 391 ms when hearing the sentences with no pause. The variance was found to statistically significant (F=7.8, p = 0.001), which means that when hearing the stimuli that included a longer pause, subjects' fixations were significantly longer than when listening to the sentences with shorter or no pauses. Post hoc tests showed that the difference in mean duration of fixation between the sentences with no pause and the sentences with a 500 ms pause was not significant. The longer fixation durations in the condition with longer pauses suggests that processing of these sentences was more difficult.

5.6.2 Time until click

When the subjects had decided which image best matches the audio prompt, their task was to use the computer mouse to click on that image. The clicks were recorded and the time it took for each subject to click was measured for each trial. The time until click is calculated from the end of the adjective (just before the beginning of the pause, where a pause had been inserted) .

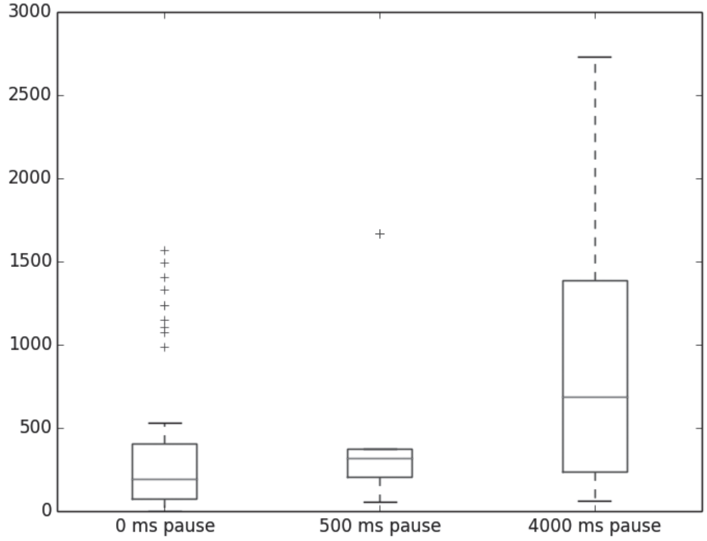

Figure 5.3: Average fixation durations in the different pause conditions

In table 5.5 the mean time it takes until the subjects click on the target image is presented. In this table any clicks that were recorded before the onset of the target noun were excluded.

Table 5.5: Time until clicking on target image

Sentence type	Mean time until click (Standard deviation)
No pause	1715 ms (455 ms)
500 ms pause	1569 ms (485 ms)
4000 ms pause	1379 ms (349 ms)

As can be seen in table 5.5, there is an observable difference in the time it takes until the participants click on their chosen answer image, and the difference is significant (F = 7.45, p = 0.0008). Post hoc tests confirm that there is a significant difference between the time until click after the 4000 ms pause and the two other types of pauses, whereas the difference between the 0 ms pause and the 500 ms pause was not significant (F = 1.75, p = 0.08). This suggests that while processing of the sentences was more effortful (as evidenced from average fixation durations), the comprehension of the sentences was not slower.

In the above analysis clicks that were recorded before the onset of the target noun were excluded, and these clicks were only found in the trials that included the longer pauses. Out of 56 clicks, 10 clicks (18 %) were recorded before the onset of the target noun, which means that they were recorded during the 4000 ms pause. At this point the subjects did not know what the target image was, as they had only been given the adjective, which could fit two different images (the target image and the competitor image). No premature clicks were recorded during the 500 ms pause, or in connection with the sentences that had no pause.

Data about the clicks that were recorded before the onset of the noun are presented in table 5.6.

The mean time for the early clicks, measured from the onset of the pause was 1808 ms (standard deviation 535 ms). This means

Table 5.6: Clicks recorded before onset of noun in the sentences with a 4000 ms pause

Number of early clicks	10
Mean time until click (measured from the beginning of the pause)	1808 ms
Standard deviation	535 ms
Minimum	1136 ms
Maximum	3104 ms

that the participants that clicked early, "gave up" on receiving more information after around two seconds, and chose to answer despite not having all of the necessary information.

5.6.3 Memory task

Results of the memory task showed that it is generally a difficult task to remember which sentences have been heard. We calculated the number of correct answers for the sentences where the target image could not be determined from the adjective (where the subjects had to wait for the noun to learn which image was the correct one). Recall was 58% for the sentences containing no pause, 88% for the sentences containing a 500 ms pause and 71% for the sentences containing a 4000 ms pause. The results suggest that adding a pause to the utterance will aid recollection of the sentence, and that a shorter pause is more effective than a longer pause. A chi square test shows that the difference between the conditions with no pauses and long pauses is not significant (chi-square = 5.1231, p = 0.077186), but that shorter pauses significantly improves the ability to remember a sentence (chi-square = 5.1692, p = 0.02299).

The effect of the pause does not occur when the target image can be determined from the adjective alone: results here show no difference in recognition. 63% for no pause, 63% for short pause and 58% for long pause.

It is important to note here that the noun in the sentences with 4000 ms pauses were sometimes skipped, when subjects chose an answer before hearing the noun. Having not heard the sentence in full will of course affect the ability to remember it.

5.7 Discussion and summary

The analysis of gaze behavior has shown that there seems to be a difference in fixation durations, depending on the presence and type of pause in the auditory stimuli. The most evident difference between conditions is the mean fixation duration, which is the longest when the subjects are hearing the sentences including a 4000 ms pause. It is not possible to draw definitive conclusions as to why fixations are longer in this condition, but generally, longer fixations are associated with more effortful cognitive processing. That would be consistent with what the subjects are experiencing in this condition, hearing an unusually long pause and trying to perceive what is being said. Still, to establish this, more research is needed to ascertain that the longer fixations are not due to other factors.

The time it takes for the subjects to click on the target image after the onset of the final word was analysed, and results show that after the 4000 ms pause subjects are significantly faster in clicking on the target image. This could be explained by the subjects having had time to familiarise themselves with the images, and therefore knowing where the target image is when they have heard the final word.

The recording of the mouse clicks of the subjects also showed that in 18% of the trials with the 4000 ms pause, the subjects clicked an image before getting all of the needed information. This suggests that the subjects may have believed that since the pause was so long, no more information was going to be presented. The long pause disrupts the perception of the auditory stimuli not only when it comes to processing of the entire sentence, but also by indicating to the subjects that the sentence has been completed, in spite of the lack of necessary information to complete the task. The mean time until the early clicks was 1808 ms, and this could be interpreted as pauses of this length being perceived as "too long" by the subjects. This indicates that the length of the pause becomes more important in the perception of the completeness of the utterance, than the syntactic information (which signals that more words should be coming).

The results of the memory task suggest that pauses, and specifically pause length, may have an effect on how well a person is able to remember auditory stimuli. Sentences that included a pause were easier to remember, compared to the sentences that did not include a pause. A shorter pause of 500 ms had a more positive effect than a longer, 4000 ms, pause. This is in line with previous research (MacGregor et al., 2010) which showed that pauses may help memorisation, but our results also show that the length of the pause has an effect on the memory process.

In the current study, the placement of the pauses was consistent, with the pause always occurring between the adjective and the noun. It future studies, it would be interesting to investigate if the placement of pauses have an effect on subjects' ability to recall sentences.

Chapter 6

An updated turn taking model

6.1 Introduction

In this chapter we are returning to the turn taking model suggested by Sacks et al. (1974) (the model was presented in section 2.2.1), and proposing several updates to it based on the results from our studies. A central question about pauses is whether an occurring pause is to be interpreted as a signal from the speaker that she is open to yielding the turn, and this puts pauses into the context of turn taking. The basis of our updated turn taking model is the notion of *turn change potential*, which is continuously updated throughout any conversation. Turn change potential (TCP) is a value that denotes the possibility of a turn change at any point during the dialogue, in contrast with TRPs (Transition Relevance Places) that occur only at specific points during a conversation, and mark the juncture between turn constructional units. TCP is not an objective measure, but something that is continuously estimated by speakers.

6.2 Background

Gravano and Vidal (2014) suggested a similar idea for dialogue systems, where they propose moving away from the binary switch/hold classification, and instead estimating the likelihood of a turn end based on extractable turn-yielding cues.

According to earlier theories, a turn may consist of one or several turn constructional units (TCUs). If a turn is made up of several TCUs, they are separated by transition relevance places (TRPs). Each TCU is seen as “possibly complete” (Clayman, 2013). TRPs are foreshadowed by different cues to the fact that the turn is winding down, which gives the listener ability to project the upcoming TRP (de Ruiter et al., 2006).

6.3 Turn Change Potential

Imagine the following situation: two persons are involved in a dialogue; let’s call them speaker 1 and speaker 2. Speaker 1 has the turn and is speaking, and speaker 2 is trying to deduce whether or not speaker 1 has reached a potential end of her turn. Commonly, speaker 2 can not be absolutely sure as to when speaker 1 has finished her turn, but she can make an estimate based on current information. The two main questions speaker 2 wants an answer to are “is speaker 1 reaching a point where she could end her turn” and “if I want to speak, and if speaker 1 is ending her turn, when should I speak in relation to the end of speaker 1’s turn”. A place where TCP is high will often be equal to a TRP.

So, the relevant TCP is not necessarily speaker 1’s actual intention, it is rather speaker 2’s perception of speaker 1’s intention. This value is dependent on a complex interplay between a number of underlying cues, such as pauses, syntax, pitch, voice quality, articulation rate, hand gestures, cognitive load level etc. We suggest that the definitions of turns and transition relevance places are too absolute, and that it would be advantageous to view them as more fluid and continuous. Furthermore, our model gives an explanation for what happens when turn taking cues are in conflict. Much research has focused on finding one definitive turn taking cue, and to find out if prosody or syntax are more important (for an overview, see de Ruiter et al. (2006)), but it likely that we use a range of cues when projecting possible turn change. For example, even though pitch movements at the end of a phrase can be categorised and correctly perceived as turn yielding or turn holding, they occur too late in the turn to be used as the sole signal of end of turn (de Ruiter et al., 2006).

A high turn change potential does not necessarily lead to a turn change, in the same way that a TRP does not demand a turn change. Whether turn change occurs will depend on the speaker 2, and whether she wants to take the turn or let speaker 1 continue. The idea of turn change potential suggests that there are no absolute right or wrong places for turn taking — just more or less likely and appropriate places for turn taking. It does not seem reasonable to aim to be able to predict every turn change in every conversation; rather, we aim to establish a model that is *good enough* at explaining and predicting possible turn change locations. Humans are not perfect at this either: for example, speakers may interrupt each other without meaning to.

There are several underlying processes at work during conversation, and in turn taking we need to be aware that the listener is working both at *what* she is going to say, and *when* she is going to say it. This is not necessarily a linear process, where the listener first waits to see when she is going to speak, and then plans what to say. Bögels et al. (2015) have shown that when answering a question, response planning begins within half a second of receiving the information needed to respond, regardless of whether the information is given early or late in the question.

Our aim is that this model should both be able to describe turn taking in humans, and be a basis for modelling turn taking behaviour in human-machine interaction (dialogue systems). We will first describe the model in general, and then explain the different components in more detail. After this, we will provide a few examples of the model at work and how it can be used to explain various turn taking patterns.

6.4 Overview of the model

There are at least three main factors that affect turn taking in our model: TCP, the threshold which the TCP must exceed to allow turn change, and the micro-timing of the turn transition. Whether turn taking does take place also depends on speaker 2's wish to take the turn, and we explore this in section 6.7. TCP is made up of several subcomponents which all keep track of turn holding and turn yielding cues, and these will be expanded upon in section 6.5. The TCP threshold indicates how high the turn change potential must be to trigger a potential turn change, and the threshold value is primarily dependent on the conversational style. We will return to this in section 6.6. TCP varies over time, and what happens when the TCP crosses the TCP threshold will depend on whether the other person wants to speak or not.

6.5 Components of turn change potential

Experimental work has shown that different turn taking cues seem to carry different weight (de Ruiter et al., 2006). In our turn taking model, turn taking cues can be weighted differently so that any specific cue can have a smaller or larger impact on the overall turn change potential, and this adds to the flexibility of the model. It also possible to completely eliminate a component. For example, non-verbal cues are not relevant when speaking to someone on the telephone, and in that case they will not affect the TCP.

Figure 6.1 displays TCP and its components: prosodic cues, non-verbal cues, duration of speech chunks and pauses, semantic cues and pragmatic cues. Each of these components is composed of features that have been shown to affect the perception of turn holding or turn yielding in conversation. Turn change

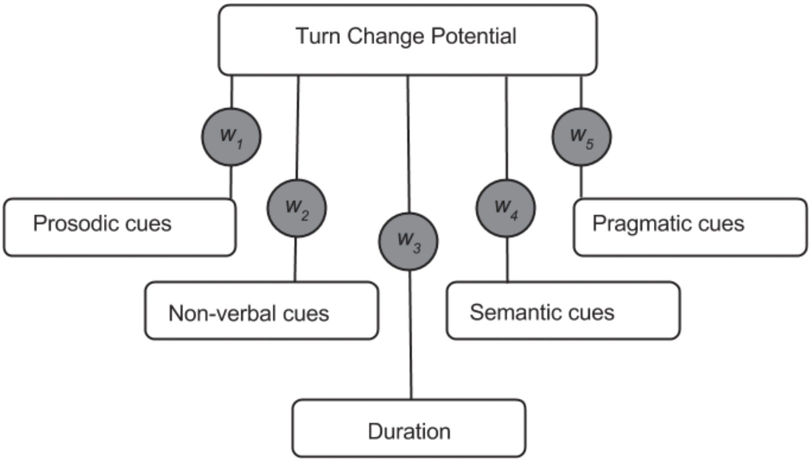

Figure 6.1: Turn change potential with components

potential is defined as the sum of turn taking cues expressed by the speaker, and it has been shown that turn taking cues have an additive effect, where more cues have a higher prognostic value (Hjalmarsson, 2011). The components of the TCP may have different weights.

It is also possible for the listener to affect TCP. If speaker 2 in our imagined conversation does not want the turn, and signals this through, for example, providing feedback, this will have an impact on TCP. We will return to this in section 6.7. Prior to that, we will examine the components that make up TCP, and explore the TCP threshold.

6.5.1 Prosodic cues

There are several prosodic features that are used as turn taking cues. Some of the features are associated with turn holding, and others with turn yielding. Prosodic turn holding cues include disfluencies (Hjalmarsson, 2011; Clark and Fox Tree, 2002), speech production phenomena such as smacks, flat intonation (Hjalmarsson, 2011), coarticulation with upcoming sounds (Local and

Kelly, 1986), and so-called rush-throughs where speech rate is increased close to a TRP (Clayman, 2013).

Turn yielding cues include increased phrase-final lenghtening (Hjalmarsson, 2011), decreased loudness and tempo deceleration, and lax and creaky fonation (Local and Kelly, 1986). The significance of pitch as a turn yielding cue is somewhat uncertain: Hjalmarsson (2011) notes that falling pitch seems to be associated with turn yielding, whereas Local and Kelly (1986) did not find pitch to be a robust cue of turn holding or yielding. Gravano and Vidal (2014) suggests that turn holding is marked by a pitch plateau, whereas a rising or falling pitch indicates turn yielding. Phrase-final lengthening does not have a significant effect on turn taking (Hjalmarsson, 2011).

These cues help us determine if a speaker is approaching the end of her turn, and once the speaker pauses we use the cues to decide whether or not it is possible for us to take the turn. Local and Kelly (1986) note that in some cases, pauses that are not at TRPs were preceded by a glottal closure, that was then held throughout the pause and released at the end of the pause. Turn holding and turn yielding cues have been documented in human-human dialogues, but humans also produce these cues when speaking to dialogue systems (Gravano and Vidal, 2014), and recognise the cues when they are produced by a synthetic voice (Hjalmarsson, 2011).

6.5.2 Semantic cues

The words spoken will of course have a large influence on perceived turn holding and turn yielding, and we call this part of the TCP semantic cues. In section 4.6 we analysed the context of different types of pauses, focusing on the word types preceding and following pauses. We found that the word types immediately surrounding pauses do not differ enough to be reliable cues to the different types of pauses (and thereby, to differentiate between the middle of a turn and the end of a turn). Instead, it is necessary to widen the scope and examine a larger context, to get enough information about whether a speaker is approaching a possible turn change. Semantic completeness is also described as arriving at a lexico-syntactic completion points, or a TRP (Transition Relevance Place). We will not define semantic completeness in terms of syntactic structure or similar, but rather we will for now suggest that it be defined as an affirmative answer to the question "is this a complete response to the previous turn" (Hjalmarsson, 2011). Semantic completeness will have a turn yielding effect, whereas semantic incompleteness will have a turn holding effect (Hjalmarsson, 2011).

Semantic cues carry much weight when a speaker needs to predict the end of a turn. de Ruiter et al. (2006) suggest that lexicosyntactic information is enough for humans to be able to reliably predict the end of a speaker′s turn, and Gravano and Vidal (2014) found that close to three quarters of speaker switches could be predicted from a written transcript of a dialogue. Bögels and Torreira (2015) addressed the claim by de Ruiter et al. (2006) and found that while lexico-syntactic information is important, intonational cues are also needed to reliably predict turn ends.

6.5.3 Duration of speech stretches and pauses

The duration of a current speech stretch or pause will also affect the perception of turn holding and turn yielding, and thereby the TCP. For example, when a speaker has just begun speaking, we will not expect the speaker to yield her turn. In section 4.5 we showed that pauses tend to occur regularly throughout conversations, and consequently speech stretches and turns are inclined to be of a certain duration. Therefore, we suggest that when speaker 1′s speech stretch or turn is approaching the typical duration of a speech stretch or turn in the current conversation, this will raise TCP.

When a speaker pauses, the pause will raise the turn taking potential, since ceasing to speak is a turn yielding cue. The longer the pause, the more it raises the turn taking potential. The length of the pause will be evaluated in comparison to the current perceived typical duration of the pause type. In section 4.4 we showed in a majority of dialogues, pause length is entrained by speakers, and therefore speakers will have a notion of how long a typical pause should be.

The turn change potential value at the beginning of a pause will aid the listener when deciding whether the pause is a planning pause or a pause where turn change is possible. Still, if a pause that has been categorised as a planning pause exceeds its typical length, the length of the pause itself may raise the turn change potential above the threshold, and thus turn change may occur regardless. Our study on the perception of pause lengths showed that when subjects hear sentences with a very long pause, the length of the pause will lead the listeners to believe that the sentence is complete, even when semantic information contradicts this (see Section 5.6.2.

The perception of a pause's duration will also depend on whether the persons are "just talking" or engaged in an activity at the same time as they are talking. If the persons are engaged in an activity long pauses may be attributed to the demands of the activity and will therefore not have the same effect on turn change potential.

6.5.4 Pragmatic cues and non-verbal cues

Pragmatic cues incorporate the types of speech acts being produced. For example, questions and the first part of adjacency pairs will have a turn yielding effect. Several non-verbal cues, such as gaze, gestures, and body positioning, are used in turn holding and turn yielding. For example, gaze directed towards the other speaker indicates an upcoming turn end (Novick et al., 1996). We will not explore the turn holding and turn yielding effects of pragmatic and non-verbal cues in more detail here.

6.6 Turn change potential threshold

In the previous section, we described TCP (Turn Change Potential) and the subcomponents, such as prosodic and semantic cues, that contribute to the TCP value. In this section we will expand on the TCP threshold. When the TCP (Turn Change Potential) value is higher than the TCP threshold, a turn change is possible. We can think of the TCP threshold as having a default value, defined by the conversational style. Tannen (2005) has explored conversational styles and suggests that speakers fall somewhere on a continuum from the high-involvement to the high-considerateness style. Turn taking in the high-involvement style is rapid, overlaps common and speakers give a lot of feedback, whereas turn taking in the high-considerateness style is characterised by longer pauses between turns and less interruptions (Norrby, 2004). The conversational style will be influenced by a speaker's background but also by the relationship between speakers and the context for the conversation.

We return to speaker 1 and speaker 2 and their conversation. Each speaker's estimation of the TCP threshold is set to a value based on the speaker's perception of the appropriate conversational style. However, the TCP threshold can be modified both by external and internal factors. If speaker 1 is talking, and speaker 2 suddenly spots a strange bird, and speaker 2 wants to get speaker 1 to see the bird before it flies away, this could external event could lower speaker 2's TCP threshold, allowing speaker 2 to speak even though speaker 1 may be in the middle of speaking, or pausing to plan what to say. Internal events that modify the TCP threshold include speaker 2's emotions: if suddenly speaker 2 becomes highly emotionally aroused, either by positive or negative emotions, this could also lower the TCP threshold to allow speaker 2 to speak even when speaker 1 is not clearly finished. After an event where the TCP threshold has been modified, it will reset to the default value for the conversation.

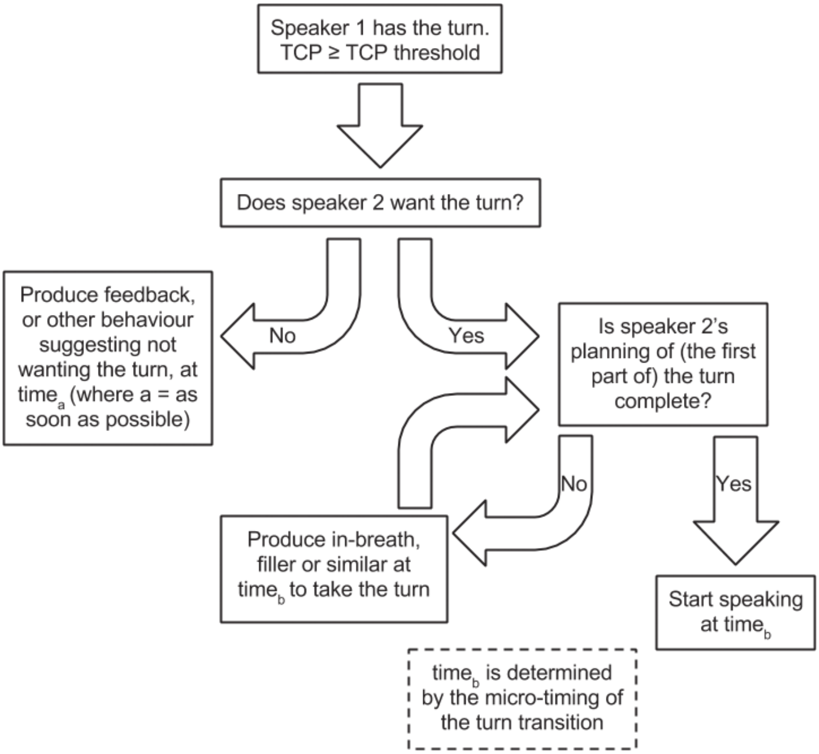
Speaker 1 has the turn.
TCP ≥ TCP threshold
Does speaker 2 want the turn?
No
Yes
Produce feedback, or other behaviour suggesting not wanting the turn, at time$_a$ (where a = as soon as possible)
Is speaker 2's planning of (the first part of) the turn complete?
No
Yes
Produce in-breath, filler or similar at time$_b$ to take the turn
Start speaking at time$_b$
time$_b$ is determined by the micro-timing of the turn transition

Figure 6.2: A schematic overview of what happens when TCP exceeds the TCP threshold

6.7 Speak or give feedback

Even when the TCP is higher than the TCP threshold, turn change does not always take place. Turn change is dependent on whether speaker 2 wants to take the turn or not. In figure 6.2 a schematic overview of what happens when TCP exceeds the TCP threshold is presented.

The actions of speaker 2 will depend on whether she wants the turn or not. If she does not want the turn, she will show this by producing feedback or other behaviour that is in keeping with not desiring to speak. If she does want the turn, she may already have planned, or be planning, what she is going to say, and her actions will depend on whether the planning of the first unit is complete. If speech planning is not complete, speaker 2 will take the turn by for example an audible inhalation while completing the production planning (Torreira et al., 2015). When speech planning is complete, speaker 2 will begin to speak at the first possible time, and this point in time is determined by the micro-timing of the transition, which will be explored in section 6.8.

In figure 6.2 we simplify the model by representing speaker 2's desire to speak as binary choice, but we believe that like TCP, a speaker's desire to take the turn would probably be more appropriately described as a continuous variable.

6.8 Micro-timing of turn transition

When TCP is higher than the TCP threshold, and speaker 2 wants to take the turn, evidence suggests that she does not start speaking the instant she is ready. Rather, the overall rhythm of the conversation will affect the micro-timing of turn transition. Wilson and Wilson (2005) proposed an oscillatory model of turn taking, where speakers' internal oscillators are entrained through syllable duration. The speakers' oscillators will be counter-phased, which means that the shortest pause between speaker 1's turn and speaker 2's turn will be equal in length to half the duration of a syllable. Sacks et al. (1974) originally suggested that the optimal turn change has no overlap and no pause, but studies have shown that a short pause is the norm (Heldner and Edlund, 2010; Stivers et al., 2009). In fact, in some cases a pause that is too short may be perceived as problematic (Roberts et al., 2011).

The micro-timing of turn transitions will be facilitated by entrainment of pause lengths, which we showed evidence for in our study presented in section 4.4.

6.9 A case study of turn change potential

We have presented the different parts of our turn taking model, and will now present an example from our corpus, to show how turn change potential can explain the production of feedback, when feedback does not occur at a pause.

6.9.1 Feedback

If speakers mainly react to pauses when deciding whether a turn is ending, feedback would only occur in connection with pauses. However, research shows that only 53% of feedback items overlap with pauses (Lundholm Fors, 2012). In figure 6.3, we see an excerpt from dialogue 5. The figure shows the pitch of the speaker (Beatrice), and below that we can see the transcription of what Beatrice is saying, and what Anna is saying. In translation, Beatrice is saying “the daycare staff (587 ms) who has asked him to come there- come to us with the boy” (the value within the parenthesis denotes the length of the pause between “staff” and “who”). Anna gives Beatrice feedback at the end of the outtake, overlapping with Beatrice saying “with the boy”.

If we were to think about the feedback given by Anna in relationship to pauses in Beatrice’s speech, it is hard to explain why it occurs where it does, and not closer to the pause at the beginning. We would expect Anna to give feedback when Beatrice is pausing, to let her know that she can continue talking. However, let us look at from a TCP standpoint instead. The pause will raise the TCP somewhat, but both pitch and semantic structure suggest that TCP should remain quite low: the slight pitch plateau at the end of “dagispersonalen” (“the daycare staff”) signals turn holding (Gravano and Vidal, 2014), and the lack of semantic completion also suggests turn holding (Hjalmarsson, 2011). The feedback provided by Anna starts approximately 140 ms after the completion of “oss” (“us”) by Beatrice. A speaker needs at least 200 ms to produce a simple utterance (Bögels and Torreira, 2015), and 200 ms before the start of the feedback utter-

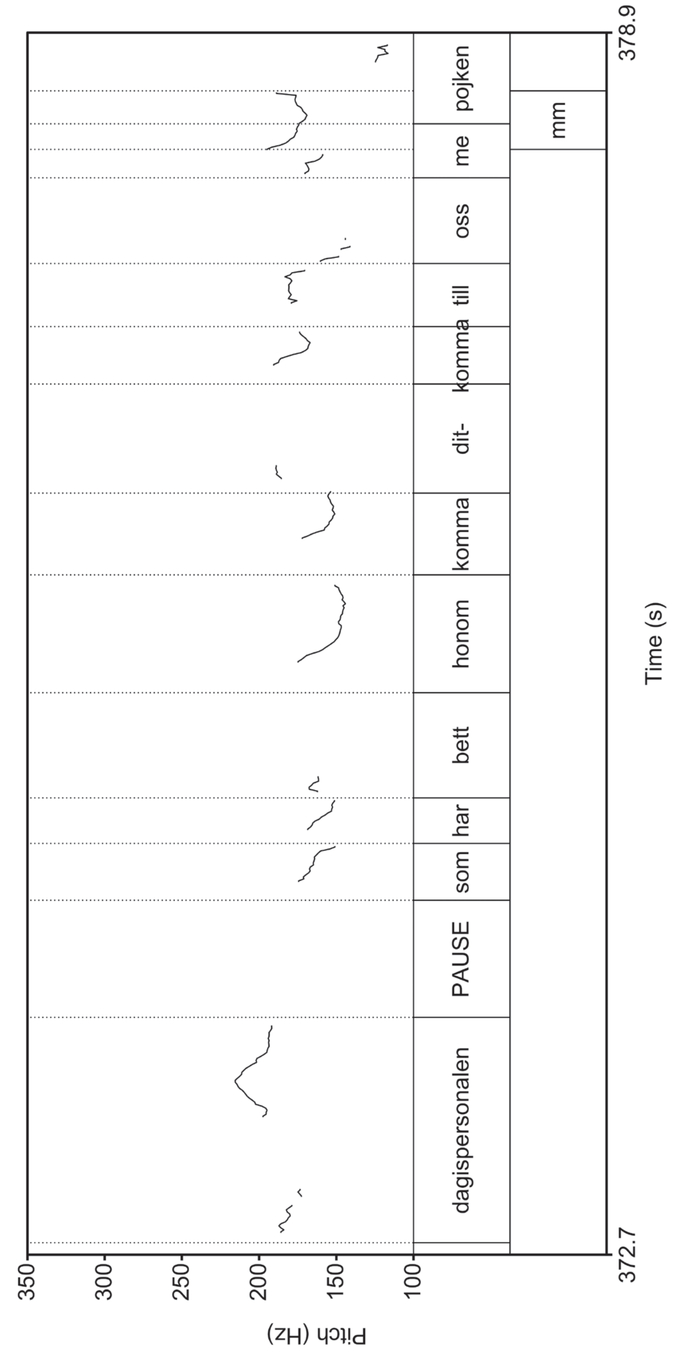

Figure 6.3: The interplay between TCP and feedback production

ance, Beatrice is in the middle of saying the [s] in “oss”. From that we can conclude that the reason Anna is providing feedback at that specific moment is that the TCP was high a few hundred milliseconds before the start of her feedback. If we look at the pitch of “till oss” (“to us”), we can see that it is falling, which is a sign of turn yielding. Further, at the end of “oss” Beatrice will reach a point of semantic completion, and these two factors contribute to pushing the TCP over the TCP threshold, triggering feedback from Anna. What Anna did not know when she was deciding to give feedback, was that Beatrice was actually not finished at the end of “oss”. She adds another couple of words to her utterance, and Anna’s feedback does not coincide with a pause, but rather with the end of Beatrice’s utterance.

6.10 Summary

In this chapter we have introduced an update to the turn taking model suggested by Sacks et al. (1974), and outlined the various components of this updated model. The results from our studies in sections 4 and 5 have been integrated into the model, and we propose that, for example, pause entrainment and the cyclical nature of speech are important factors in the perception of Turn Change Potential. We propose that TCP is continuously updated throughout dialogues, and that speakers use their estimate of TCP to predict when it is possible to take the turn.

Chapter 7

Summary and future work

7.1 Pause production, pause perception and turn taking

In this dissertation we have investigated different aspects of pause production, such as length, placement and how speakers' affect each other's pausing behaviour. Analysing these different aspects of pausing behaviour also meant developing and applying different methods that could capture the qualities we were interested in. We have found that pause lengths tend to be entrained by speakers involved in dialogues, and that pauses occur somewhat regularly throughout conversations. Pauses are part of the context of the surrounding conversation, where all different parts of language and communication will influence the final "product". This means that while it is interesting to look at pauses at their own, to really understand them we need to see the bigger picture. To highlight the role of pauses in conversation, we proposed an update to the turn taking model presented by Sacks et al. (1974), and suggest how the results from our studies can be integrated in this model.

7.2 Future work

We have identified several interesting areas of future work. Primarily, we would like to further formalise our turn taking model and explore how probabilistic reasoning can be used to model entrainment, production and interpretation of pausing behaviour. This would also include performing experiments to implement and test the model in dialogue systems. We would also like to investigate how feedback can be used to modulate Turn Change Potential.

Our work on pause length entrainment raised the idea that emotional arousal may have an effect on entrainment, and this is something that we wish to explore further. Entrainment in the micro-timing of turn transitions, described in Section 6.8 is also an interesting area in need of more research. Wilson and Zimmerman (1986) propose that pause lengths at turn changes should be multiples of half the syllable length in the current conversation, and we would like to test this experimentally. These are a few of the areas of potential work that we see in the future.

In research we aim to explain what we observe and to answer our questions, but, as always, we find that the answers lead to new questions, to be explored in future endeavours.

Bibliography

Agyekum, K. (2002). The communicative role of silence in Akan. *Pragmatics*, 12(1):31–51.

Boersma, P. and Weenink, D. (2014). Praat: doing phonetics by computer. Computer program.

Bögels, S., Magyari, L., and Levinson, S. C. (2015). Neural signatures of response planning occur midway through an incoming question in conversation. *Scientific Reports*, 5(12881).

Bögels, S. and Torreira, F. (2015). Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, 52:46–57.

Bølviken, E. (1983). New tests of significance in periodogram analysis. *Scandinavian Journal of Statistics*, 10(1):1–9.

Brants, T. (2000). TnT: a statistical part-of-speech tagger. *Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP-2000)*, pages 224–231.

Brennan, S. E. and Brook, S. (1996). Lexical entrainment in spontaneous dialog. *1996 International Symposium on Spoken Dialogue, ISSD-96*, pages 41–44.

Brown, P. and Levinson, S. C. (1978). Universals in language usage: Politeness phenomena. In *Questions and politeness: Strategies in social interaction*, pages 56–311. Cambridge University Press.

Butterworth, B. (1979). Hesitation and the production of verbal paraphasias and neologisms in jargon aphasia. *Brain and language*, 8(2):133–61.

Campione, E. and Véronis, J. (2002). A large-scale multilingual study of silent pause duration. In *Speech Prosody 2002, International Conference*. Citeseer.

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

Clayman, S. E. (2013). Turn-constructional units and the transition-relevance place. In Sidnell, J. and Stivers, T., editors, *The handbook of conversation analysis*, pages 151–166. Wiley Online Library.

de Ruiter, J. P., Mitterer, H., and Enfield, N. J. (2006). Projecting the end of a Speaker's Turn: A Cognitive Cornerstone of Conversation. *Language*, 82(3):515–535.

Duez, D. (1993). Acoustic correlates of subjective pauses. *Journal of Psycholinguistic Research*, 22(1):21–39.

Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., and House, D. (2010). Spontal: A swedish spontaneous dialogue corpus of audio, video and motion capture. In *LREC*, pages 2992–2995.

Edlund, J., Heldner, M., and Hirschberg, J. (2009). Pause and gap length in face-to-face interaction. In *Proceedings of Interspeech 2009*, pages 2779–2782.

Fraser, B. (1999). What are discourse markers? *Journal of pragmatics*, 31(November 1996):931–952.

Fujio, M. (2004). Silence during intercultural communication: a case study. *Corporate Communications: An International Journal*, 9(4):331–339.

Goldman-Eisler, F. (1958). Speech production and the predictability of words in context. *The Quarterly Journal of Experimental Psychology*, 10(2):96–106.

Goldman-Eisler, F. (1961). The distribution of pause durations in speech. *Language and Speech*, 4:232–237.

Gravano, A. and Vidal, C. A. J. (2014). A Study of Turn-Yielding Cues in Human-Computer Dialogue. In *15th Argentine Symposium on Artifical Intelligence (ASAI 2014)*, pages 9–17.

Gustafson-Capková, S. and Hartmann, B. (2006). *Manual of the stockholm umeå corpus version 2.0*. University of Gothenburg.

Halácsy, P., Kornai, A., and Oravecz, C. (2007). HunPos: an open source trigram tagger. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, number June, pages 209–212.

Hansson, P. (1998). Pausering i spontantal. Bachelor's thesis in phonetics, Department of linguistics, Lund University.

Heldner, M. (2011). Detection thresholds for gaps, overlaps, and no-gap-no-overlaps. *The Journal of the Acoustical Society of America*, 130(1):508–13.

Heldner, M. and Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

Hjalmarsson, A. (2011). The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication*, 53(1):23–35.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., and van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.

Horne, M., Strangert, E., and Heldner, M. (1995). Prosodic boundary strength in Swedish: Final lengthening and silent interval duration. *Proceedings ICPhS*, (3).

Howell, P. and Sackin, S. (2001). Function word repetitions emerge when speakers are operantly conditioned to reduce frequency of silent pauses. *Journal of psycholinguistic research*, 30(5):457–74.

Huettig, F., Rommers, J., and Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta psychologica*, 137(2):151–71.

Jakobsson, A. (2013). *An introduction to time series modeling*. Studentlitteratur.

Jefferson, G. (1989). Preliminary notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation. In Roger, D. and Bull, P., editors, *Conversation: An interdisciplinary perspective*, Intercommunication Series. Multilingual Matters.

Kendall, T. (2009). *Speech Rate, Pause, and Linguistic Variation: An Examination Through the Sociolinguistic Archive and Analysis Project*. PhD thesis, Duke University.

Kircher, T. T., Brammer, M. J., Levelt, W., Bartels, M., and McGuire, P. K. (2004). Pausing for thought: engagement of left temporal cortex during pauses in speech. *NeuroImage*, 21(1):84–90.

Kousidis, S. and Dorran, D. (2009). Monitoring convergence of temporal features in spontaneous dialogue speech. In *Conference papers*.

Kousidis, S., Schlangen, D., and Skopeteas, S. (2013). A crosslinguistic study on turn-taking and temporal alignment in verbal interaction. In *Proceedings of Interspeech 2013*, pages 803–807.

Krahmer, E., Swerts, M., Theune, M., and Weegels, M. (2002). The dual of denial: Two uses of disconfirmations in dialogue and their prosodic correlates. *Speech communication*, 36(1-2):133–145.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

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

Levitan, R. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Twelfth Annual Conference of the International Speech Communication Association*, pages 3–6.

Lindblad, P., Karlsson, S., and Heller, E. (1991). Mandibular movements in speech phrases - A syllabic quasi-regular continous oscillation. *Scandinavian Journal of Logopedics and Phoniatrics*, pages 36–42.

Linell, P. (2004). *The written language bias in linguistics: Its nature, origins and transformations*. Routledge.

Lisker, L. (1957). Closure duration and the intervocalic voiced-voiceless distinction in English. *Language*, 33:42–49.

Local, J. and Kelly, J. (1986). Projection and 'silences': Notes on phonetic and conversational structure. *Human Studies*, 9:185–204.

Lundholm, K. (2000). Pausering i spontana dialoger: En undersökning av olika paustypers längd. Bachelor's thesis, Lund University.

Lundholm Fors, K. (2012). The temporal relationship between feedback and pauses: a pilot study. In *Interdisciplinary Workshop on Feedback Behaviors in Dialog (an Interspeech 2012 satellite event)*, pages 43–45.

MacGregor, L., Corley, M., and Donaldson, D. (2010). Listening to the sound of silence: Investigating the consequences of disfluent silent pauses in speech for listeners. *Neuropsychologia*, 48(14):3892–3892.

McFarland, D. H. (2001). Respiratory markers of conversational interaction. *Journal of speech, language, and hearing research : JSLHR*, 44(February):128–143.

Megyesi, B. (2009). The open source tagger hunpos for swedish. In *Proceedings of the 17th Nordic conference on computational linguistics (NODALIDA)*.

Megyesi, B. and Gustafson-Čapková, S. (2002). Production and perception of pauses and their linguistic context in read and spontaenous speech in Swedish. In *Interspeech*.

Menenti, L., Pickering, M. J., and Garrod, S. C. (2012). Toward a neural basis of interactive alignment in conversation. *Frontiers in human neuroscience*, 6(June):185.

Merlo, S. and Barbosa, P. A. (2010). Hesitation phenomena: a dynamical perspective. *Cognitive processing*, 11(3):251–61.

Mushin, I. and Gardner, R. (2009). Silence is talk: Conversational silence in australian aboriginal talk-in-interaction. *Journal of Pragmatics*, 41:2033–2052.

Nakane, I. (2006). Silence and politeness in intercultural communication in university seminars. *Journal of Pragmatics*, 38(11):1811–1835.

Newman, H. (1982). The sounds of silence in communicative encounters. *Communication Quarterly*, 30(2):142–149.

Norrby, C. (2004). *Samtalsanalys: så gör vi när vi pratar med varandra*. Studentlitteratur.

Norwine, A. and Murphy, O. (1938). Characteristic time intervals in telephonic conversation. *The Bell System Technical Journal*, 17:281–291.

Novick, D., Hansen, B., and Ward, K. (1996). Coordinating turntaking with gaze. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 3.

Oehmen, R., Kirsner, K., and Fay, N. (2010). Reliability of the manual segmentation of pauses in natural speech. In *Advances in Natural Language Processing: 7th International Conference on NLP, IceTAL 2010, Reykjavik, Iceland, August 16-18, 2010, Proceedings*, volume 6233, pages 263–268. Springer-Verlag New York Inc.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, 119(4):2382.

Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *The Behavioral and brain sciences*, 27(2):169–90; discussion 190–226.

Reich, S. S. (1980). Significance of pauses for speech perception. *Journal of Psycholinguistic Research*, 9(4):379–389.

Roberts, F., Francis, A., and Morgan, M. (2006). The interaction of inter-turn silence with prosodic cues in listener perceptions of "trouble" in conversation. *Speech Communication*, 48(9):1079–1093.

Roberts, F. and Francis, A. L. (2013). Identifying a temporal threshold of tolerance for silent gaps after requests. *The Journal of the Acoustical Society of America*, 133(6):EL471–7.

Roberts, F., Margutti, P., and Takano, S. (2011). Judgments Concerning the Valence of Inter-Turn Silence Across Speakers of American English, Italian, and Japanese. *Discourse Processes*, 48(5):331–354.

Rochet-Capellan, A., Bailly, G., and Fuchs, S. (2014). Is breathing sensitive to the communication partner? In *7th Speech Prosody Conference, Dublin : Ireland (2014)*.

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

Sajavaara, K. and Lehtonen, J. (2011). The silent Finn revisited. In Jaworski, A., editor, *Silence: Interdisciplinary Perspectives*, pages 263–283. Walter de Gruyter.

Scollon, R., Scollon, S., and Scollon, R. (1981). *Narrative, literacy and face in interethnic communication*. Ablex Norwood, NJ.

Sifianou, M. (2011). Silence and politeness. In Jaworski, A., editor, *Silence: Interdisciplinary Perspectives*. Walter de Gruyter.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J. P., Yoon, K.-E., and Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences of the United States of America*, 106(26):10587–92.

Strangert, E. (1993). Speaking style and pausing. *Phonum (Dept Phon., UmeåUniv.)*, (2):121–137.

Strangert, E. (2003). Emphasis by pausing. *Proc. 15th ICPhS, Barcelona*, pages 0–3.

Szczepek Reed, B. (2010). *Analysing Conversation: An Introduction to Prosody*. Palgrave Macmillan.

Tannen, D. (2005). *Conversational style: Analyzing talk among friends*. Oxford University Press.

ten Have, P. (2007). *Doing Conversation Analysis: A Practical Guide*. Introducing Qualitative Methods series. SAGE Publications.

Tisljár Szabó, E. and Pléh, C. (2014). Ascribing emotions depending on pause length in native and foreign language speech. *Speech Communication*, 56:35–48.

Torreira, F., Bögels, S., and Levinson, S. C. (2015). Breathing for answering: the time course of response planning in conversation. *Frontiers in Psychology*, 6(March):1–11.

van Donzel, M. and Koopmans - van Beinum, F. J. (1996). Pausing strategies in discourse in Dutch. In *Fourth International Conference on Spoken Language Processing*, pages 1029–1032, Philadelphia, PA, USA.

Wilson, M. and Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin and Review*, 12(6):957–968.

Wilson, T. and Zimmerman, D. (1986). The Structure of Silence between Turns in Two-party Conversation. *Discourse Processes*, 9:375–390.

Wlodarczak, M. and Wagner, P. (2013). Effects of talk-spurt silence boundary thresholds on distribution of gaps and overlaps. In *Interspeech 2013*, pages 1434–1437.

Yanushevskaya, I., Kane, J., de Looze, C., and Chasaide, A. N. (2014). The distribution of pitch patterns and communicative types in speech-chunks preceding pauses and gaps. In *Proceedings of the 7th International Conference on Speech Prosody (SP2014)*, pages 959–963.

Zellner, B. (1994). Pauses and the temporal structure of speech. In Keller, E., editor, *Fundamentals of speech synthesis and speech recognition*, pages 41–62. John Wiley, Chichester.

Production and Perception of Pauses in Speech

Kristina Lundholm Fors

Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg

Dissertation edition, September 2015