Bachelor Thesis

A Corpus-Based Approach to the Classification and Correction of Disfluencies in Spontaneous Speech

Jana Besser
Matriculation Number: 2507562

November 7, 2006

Supervised by:
Prof. Dr. Wolfgang Wahlster,
Dr. Jan Alexandersson

Saarland University,
Faculty of Humanities,
Department of General Linguistics,
Chair of Computational Linguistics

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides Statt, dass ich diese Arbeit selbstständig und nur unter Zuhilfenahme der ausgewiesenen Hilfsmittel angefertigt habe.

Declaration of Academic Honesty

I hereby declare to have written this thesis on my own, having used only the listed resources and tools.

Location, Date

Signature

Abstract

This thesis presents a classification scheme for disfluencies occurring in spontaneous spoken language. Disfluencies are irregular speech phenomena that range from rather simple entities such as sounds of hesitation and slips of the tongue to more complex structures as self-corrections. The presented classification scheme covers phenomena, which affect the structural, non-semantic layer of an utterance.

The designed scheme is an extension of previous work on the topic and was conducted in a data-driven approach. It is based on an analysis of data from the AMI Meeting Corpus (McCowan et al., 2005). The corpus contains human-human dialogues on a prescribed topic. The examination of the data led to the identification of 1205 disfluency instances within a total of 792 dialogue acts, showing that disfluencies are prevalent phenomena in spontaneous speech. Therefore their examination is of significant relevance for natural language applications.

The scheme is designed to serve as a basis for the development of a computational tool for automatic disfluency corrections. However, it may be used for other purposes as well, e.g. for speaker-related data collection, such as a person's characteristic speech behaviour. For this reason, the classification contains several layers, which correspond to different degrees of subdivision of the disfluency types. In this way an adequate layer can be chosen depending on the application, the scheme is used for.

The present work includes also an annotation scheme that was developed according to the classification. Its reliability was tested by comparing annotations of several annotators, who used the scheme but were not familiar with the topic of disfluencies. The annotations were compared according to a number of different metrics and the inter-annotator agreement was calculated using two different statistics: the κ- and the AC1-statistics (Gwet, 2002). The results show that the agreement on the annotations is very high (about 0.95), once the boundaries of a DF were identified similarly.

Contents

Acknowledgements

I am sincerely grateful to all the people, who have supported and encouraged me whilst the process of accomplishing this work.

First I want to thank Wolfgang Wahlster, who permitted me to work on the topic at the Intelligent User Interface department of the DFKI GmbH (German Research Center for Artificial Intelligence).

I am especially indebted to my supervisor Jan Alexandersson. He gave me invaluable support throughout the process, always having a critical but encouraging look at the research I was doing. I appreciate the cooperation and correspondence with him and would also like to thank him for advice and discussions that go beyond the scope of this work.

A special note of thanks goes to Thomas Kleinbauer. His advice and constructive contributions were strongly conducive to the progress of my research. His practical tips brought solutions to a number of challenges (both large and small) I encountered during the process.

Thanks to Tilman Becker (and Jan Alexandersson) for providing me with the opportunity to write this thesis within the AMI project, allowing me to use the AMI meeting corpus for my investigations and “lending” me two of the project’s research assistants for disfluency annotations. I also want to express my gratitude to the annotators themselves: Beata Biehl and Alexandria Kimsey. Special thanks to Alexandria for taking her time for language related corrections of the annotation manual.

Sincere thanks also go to the annotators, who did the work in their free time: Lisa Fiedler and Steven B. Poggel.

Other people to whom I am extremely grateful: Stephan Lesch, who contributed with technical support in different areas; Benjamin Lang and Jochen Frey, who made the evaluation of the annotation scheme possible on a technical level; Dan Flickinger for support on the LKB system; Kilem Gwet for support on the generalisation of the AC1-formula for multiple raters; all the people, who gave me valuable recommendations for essential background literature on the topic; all the people that helped with spell-checking and other corrections of the work.

Finally, I would like to thank my mother, my boyfriend, and my friends, who contributed with a lot of encouragement and care.

Chapter 1

Introduction

Speech differs highly from written language. When people talk, they produce a lot of linguistic irregularities, so called *disfluencies* (henceforth, DFs). Linguistic irregularities in the present work mean syntactic and grammatical errors according to standard syntax and grammar. DFs are often named *self-corrections* or *self-repairs* but the present work avoids these terms as a superordinate concept for all phenomena. Not all DFs include a speaker's self-correction. They can stay uncorrected or serve to add new information to the context.

The causes for DFs are various. Some cases can be explained by the assumption that speakers sometimes start an utterance before they have formulated it. Then the DF is used as a provisional start, which is revised later if it did not express what the speaker wanted to say. This even holds for mid-sentence DFs, since the process of formulation continues while speaking.

The present work examines the regularities of the occurring DF phenomena and the possibility to divide them into classes according to their structure. Disfluency phenomena have a quite regular structure. This has been done in other research before, but these approaches did not cover the existent phenomena to a satisfying degree. The present work is more fine-grained than the previous approaches and covers a larger set of DF types, see chapter 2.

In the current approach DFs are divided into three subgroups: *uncorrected* phenomena, *deletable* phenomena, and *revisions*, see chapter 3. Only *revisions* have the structure that is typically associated with DFs and that is described in almost all work on the topic. It consists of an erroneous part (*reparandum*), an optional medial region (*interregnum*) that prepares the correction, and finally the correction (*reparans*), see chapter 3.1. *Uncorrected* and *deletable* DFs do not include a self-correction by the speaker. Therefore they consist only of a *reparandum*.

DFs can be left uncorrected for different reasons: Either the speaker did not notice her error or she noticed it but decided not to correct it since e.g.

a small error could gain too much focus if it gets corrected. Furthermore the communicative intention may not be disturbed by the error. Human listeners are quite good in ignoring errors and interpreting slightly imperfect speech, which makes certain corrections superfluous.

The present work considers only phenomena that actually lead to the interruption of the syntactic or grammatical fluency of an utterance. This excludes meta comments and certain stylistic devices from the classification. The approach is only concerned with the structural correctness of an utterance and thus no analysis of the semantic or pragmatic impacts of DFs were considered. The underlying psychological processes were neither examined.

1.1 Motivation

The number of natural language applications that focus on spontaneous speech input rises.Those systems have to be able to handle all the challenges that arise in conjunction with speech, e.g. DFs. Many current computational natural language parsers were though developed to handle written language and thus DFs can mean extensive trouble for them. This makes the topic interesting from a computational linguistic point of view. A tool for the automatic detection and correction of speech disfluencies, that is incorporated in a natural language application as a pre-processing module for the parser, could make it possible to keep an existent parser and allow for the processing of DFs anyway. This would make the system more robust and user-friendly, since since it would allow for less restricted spontaneous speech input.

The classification scheme developed and reported in this work is meant to contribute to the preparation of automatic detection and correction of speech disfluencies. Since such a correction tool would have to give a cleaned version of the original utterance as output, where all DFs have been cleared, the DF annotation scheme presented here generates an appropriate correction for each DF.

The DF classification scheme can also be used for examinations of other data in various fields. For example, an investigation could be made how disfluencies correspond to a person's social status, age, education, certainty, or other agent- or context related factors.

1.2 Aims

The aim of this work was to develop a fine-grained broad coverage classification scheme for DFs occurring in spontaneous spoken language. The classification has been made in a data-driven approach based on the analysis of data from the AMI meeting corpus (McCowan et al., 2005). The work is supposed to combine and extend previous research on the topic to produce a scheme that covers all kinds of DF phenomena and does not only focus on certain types of DF. The disfluency classification scheme can hopefully be used for a large set of approaches, which have to handle disfluencies in one way or another, irrespective of the application's domain and purpose of the application.

Another goal was to create an annotation scheme from the DF classification. The annotation scheme can be used for marking disfluency types and their erroneous vs. correcting parts or for inserting corrections for uncorrected DFs.

Finally, the classification scheme is meant to be used for enabling the automatic detection and correction of DFs in speech input to natural language applications. In this scenario the scheme could be used for training a correction tool, in the way described in the previous section.

1.3 AMI Project

The disfluency classification scheme was developed as part of the AMI project (http://www.amiproject.org). The project is operated by a multidisciplinary 15-member consortium, which consists of both academic partners, industrial partners, and research institutes. The name AMI stands for Augmented Multi-party Interaction. The project's goal is to develop technology to support and enrich communications between individuals and groups of people. According to (McCowan et al., 2005) some research topics of the project are 1) *Definition and analysis of meeting scenarios*, 2) *Infrastructure design, data collection and annotation*, 3) *Processing and analysis of raw multi-modal data*, 4) *Processing and analysis of derived data*, and 5) *Multimedia presentation.* For instance, one activity within the project is the work on automated meeting summarisations, which falls into category 4). Disfluency detection and correction is a nearly compulsive matter for reaching this goal.

1.4 Corpus

For the research within the AMI project a corpus was created, which contains records of business meetings that hold about 100 hours of meeting time. The meetings were recorded in instrumented rooms equipped with a variety of instruments, yielding high-quality multi-modal recordings, see (McCowan et al., 2005) for further information. Both real meetings and scenario-driven meetings are contained in the corpus. In the scenario-driven meetings the participants are talking about a prescribed topic.

The speech recordings have been transcribed into text for research purposes. A part of the transcriptions was functioning as the basis for the development of the classification scheme presented here.

The corpus contains unrestricted and uncontrolled human-human discussions, which means that a lot of aspects of DFs produced in natural speech should occur in the dialogues. Furthermore, the meetings were held in English, but a large proportion of the participants were non-native English speakers, resulting in a high variability of speech patterns, which is not covered by many other corpora (McCowan et al., 2005). On the other hand the meeting participants were sitting in the same room, having face-to-face conversations. This means that gestural cues and other factors, which can influence DF production could not be excluded. The suggestion is though, that the basic phenomena should not be affected by this.

For empirical approaches it is desirable to examine a large corpus. The larger the corpus, the more robust are the results gained by the investigation. Yet, due to time restrictions, only a limited amount of meetings could be analysed in the presented work. However, as chapters 2 and 3 show, it was still possible to identify an expanded range of DF phenomena. Chapter 4.4 reports that most occurring disfluency types are covered by the scheme.

1.5 Method

The classification is exclusively based on examinations of the meeting transcriptions. No acoustic material was used, neither while creating the classification scheme nor for the annotations. Time limitations made this restriction necessary in spite of the obvious advantages, which result from the inclusion of acoustic information, such as an abbreviation in taking decisions on class assignment in critical cases. However as stated in (Shriberg, 1994) and (Finkler, 1997), even if audio recordings are available, this does not guarantee that all ambiguities can be resolved. Finkler (1997) emphasises that in order to avoid interpretation by the annotator, it is often necessary to ask the speaker herself.

The transcriptions contain only plain text. This means that no non-linguistic material such as laughter, coughing or similar sounds were included in the investigations. Working on the transcriptions also meant to trust in the transcriber's analysis of the speech material. For example, a transcribed "I" was always treated as a full-word, a personal pronoun, while "i" was treated as a word fragment, even if it is not sure that the transcriber interpreted the sound correctly. The same counts for "uh" and "a", which can sound very similar, while the first is a hesitation and the second is an article or a word-fragment.

For the examinations, the corpus had to be segmented into discrete units for analysis. The most convenient way of segmenting the corpus was by dialogue acts (DAs), since it had already been split this way. Every DA was treated as one utterance and annotations were made on one segment a time, not including phenomena, which exceed segment boundaries. The annotators were not supposed to include information from the surrounding segments, from the whole discourse or their world knowledge in order to complete their class assignments. This also meant that no semantic analysis of the content and no semantic corrections were to be made.

The segments were processed with the parser of the LKB system (http://www.delph-in.net/lkb) before annotation. Only segments that could not be parsed were annotated with DF classes.

For the comparison of the annotations' similarity, a number of metrics were created to identify the DF instances that were similarly defined by the annotators. With two different statistics also the inter-annotator agreement was calculated, see chapter 4.4.

1.6 Chapter Summary

The goal of this thesis is to provide a classification scheme for speech disfluencies of all types. This is motivated by the fact that DFs in spoken language can cause serious trouble for computational listeners and the ability to handle DFs is therefore of importance for any computer system that deals with spontaneous speech.

The underlying research was carried out within the AMI project. In a data-driven approach, DF phenomena in human-human conversations from the AMI meeting corpus were analysed in order to gain an appropriate classification scheme. The DF class definitions reflect the surface structure of the encountered phenomena and do not consider any underlying processes.

The work also involves DF annotations by several annotators. The annotations were made according to a scheme developed on the basis of the created classification scheme. They were then compared with respect to similarity through measuring the inter-annotator agreement.

Chapter 2

Theoretical Background

Several researchers have investigated the topic of speech disfluencies before, by various motivations. Their work serves as an important basis for this thesis, since the present work extends the list of the previously identified DF phenomena with additional classes.

This chapter is supposed to give an overview on the existing research on the topic. The outline is by far not exhaustive but should give a good all-around impression of previous approaches. Two approaches that provide a classification of a great extent of DF phenomena will be presented in detail. The first one is (Shriberg, 1994), chapter 2.1. Shriberg's thesis is an absolute foundation in the research on DFs. She elaborated regularities in the production of DFs and created a detailed classification scheme of DF phenomena. Also Finkler (1997) provides a valuable and elaborate DF classification, which is presented in chapter 2.2. The DF classes identified by Finkler and Shriberg are also presented in tables 2.1 and 2.2 where they are compared to the corresponding classes of the current scheme. Chapter 2.3 presents relevant findings and classifications from miscellaneous other research. It is mainly concerned with differences in the naming of DF phenomena.

2.1 Shriberg - Regularities in DF Production

Shriberg (1994) examined disfluencies in spontaneous speech of adult normal speakers of American English. Her aim was to find and illuminate regularities in disfluency (DF) phenomena. According to Shriberg, these regularities have consequences for models of human language production, which she was concerned about as a psycholinguist. Furthermore, the observation of systematic disfluency patterns can help to improve the performance of speech applications.

Shriberg used a strongly data-driven approach in her investigations. Her work was based on the analysis of the three following corpora: 1) ATIS, which contains task-oriented human-computer dialogues on air travel planning, 2) AMEX, which contains task-oriented human-human dialogues, also on air travel planning, and 3) the SWITCHBOARD corpus, which consists of informal human-human dialogues on a prescribed topic. Shriberg's work includes an analysis of more than 5000 DF instances, for which she used both transcriptions and audio recordings. The data were analysed with respect to several different features. Shriberg defines features as observable characteristics in the data. For further information about the features examined see (Shriberg, 1994).

Shriberg included only same-turn and same-sentence DFs in her investigation. Thus no phenomena, which exceeded turn or sentence boundaries were considered. Moreover, she only regarded cases "in which a contiguous stretch of linguistic material must be deleted to arrive at the sequence the speaker intended[...]" (Shriberg, 1994).

This means that she did not consider any disfluencies where material has to be added or changed in order to gain the sequence the speaker (presumably) intended. Thus phenomena, which are classified as *Omission* or *Order* in the present work are not covered by Shriberg's classification.

As most other researchers Shriberg defines a three-phase structure of DFs. The terms she uses are adapted from Levelt (1983), with some modifications: The *reparandum* (RM) is the whole stretch of speech to be deleted. This goes in accordance with the present work but is against Levelt and Finkler (1997) where only the altered or corrected element(s) are considered as *reparandum.* The RM is followed by the *interruption point* (IP) and then the *interregnum* (IM), which is named *editing phase* in many other works. The third and last part is the *repair* (RR), which is defined as the stretch of speech that corresponds to and "corrects" the material in the RM. The RM and the IM are the regions which are to be deleted in order to arrive at the intended utterance.

Shriberg's classification contains the following classes (see also tables 2.1 and 2.2):

ART: This class contains DFs, which arise in connection with speech errors (= SOTs in the present work).

HYB: The Class HYB is used for phenomena, which include the presence of at least two substituted, inserted or deleted words.

SUB: One word from the RM is substituted by another one in the RR.

INS: The RR contains a word, which did not occur in the RM (insertion).

DEL: None of the material from the RM is repeated in the RR.

REP: Material from the RM is repeated in the RR.

CON: Denotes coordinating conjunctions, e.g. "and" between two sentences, connecting them to one sentence.

FP: Filled pauses (only "um" and "uh" were attended).

Shriberg's class DEL (deletion) covers the same phenomena as the class *restart* in the present work. The only difference is that for DEL only the RM is annotated whereas for *restart* both RM and *reparans* (= RR in Shriberg's work) are annotated.

Discourse markers and *explicit editing terms* (EETs) are accounted for in Shriberg's work but not considered as DFs. Contrary to the ideas presented in this paper, Shriberg regards them as expressions, which are not only distinct in their position in the utterance (whether or not they are within an IM) but also in the terms they use. However, also here EETs are directly associated with a DF and can only occur within the IM or directly after the repair.

As table 2.2 shows, the classes CON and HYB have no direct counterpart in the present work. HYB is covered by other classes, while a CON phenomenon does not have to lead to disfluency and therefore was not considered as relevant in the current approach. CON denotes cases, where two proper sentences are connected with "and", which does not cause syntactic irregularity. Shriberg (1994) also states that these phenomena are extremely rare.

As in the present work, serial and complex DFs are treated as consecutive basic DFs or as a hierarchically organized complex of basic DFs in Shriberg's approach. DFs are denoted as "serial", where one DF's repair (RR) is immediately followed by the *reparandum* (RM) of the next one. In a complex DF, two or more IPs bind both preceding and following material. The correction of a complex DF proceeds outward from the inmost to the outmost DF. Shriberg (1994, chap. 4.3.4.8, p. 68) gives the following example for a complex DF:

(1) he she she went

The disfluent sequence includes the first three words. It contains two disfluencies. The first "she" is the repair region (RR) of the first disfluency. It replaces the "he". The second "she" is the RR of the second DF. It repeats the first "she". The sequence is an example for a case, where one word is both repair of one disfluency and reparandum of another one.

One clear difference between Shriberg's approach and the present work is that the class REP (repetition) also includes cases where the first element of the repetition (= the RM) is a word fragment or a mispronunciation. The present work classifies only real repetitions as *repetition*, since the risk of false interpretation is otherwise too high. The following example from (Shriberg, 1994, chap. 4.3.4.5.2, p. 65) illustrates the difference:

(2) show me grand trouns- ground transportation

In Shriberg's approach, “grand” is analysed as a mispronunciation of “ground” and thus “ground” is seen as a repetition of “grand”. Furthermore “transportation” is marked as repeating the word “trouns”, which is both a misarticulation and a word fragment. In the approach presented here, none of these phenomena would be classified as *repetitions*:

2.2 Finkler - Automatic Generation of Self-Repairs

With PERFECTION Finkler (1997) created a system for the incremental syntactic generation of natural language. This means that the system starts generating speech output before the complete input information from the (human) conversational partner is given. According to Finkler (1997) this is supported by psycholinguistic models of speech production. Due to non-monotonic input specifications, the incremental processing sometimes makes later modifications of the already produced output necessary, which leads to a self-correction scenario.

Finkler's aim was to improve the speech production of a natural language system by making it more flexible and adaptable to the situational context and the conversational partner. The system replicates typical performance phenomena of human self-corrections. It is supposed to produce human-like speech even in situations where errors occur. By this, the user's acceptance of the system is meant to be increased and error situations should be resolved as good as in human-human communication.

In order to produce human-like self-corrections, Finkler analysed a corpus, the results of which he took as the empirical basis for his work. The corpus he used contained task-oriented human-human dialogues, recorded within the VERBMOBIL project (http://verbmobil.dfki.de/). The dialogue participants were supposed to arrange a meeting. Finkler had access both to audio recordings and transliterations of the dialogues. This allowed him to consult the acoustic material when the transliterated data did not give sufficient information for the classification of a phenomenon.

Finkler (1997) examined 336 dialogues, which contained a total of 4590 turns and correspond to 8 hours of meeting time. In the material he identified 1251 self-corrections. The corrections were evaluated for 20 criteria, e.g. where the *interruption point* typically is located, which of the possible corrections is chosen and how much of the uttered material is usually repeated at the beginning of the correction. The choice of the criteria shows that Finkler was interested in generating self-corrections, not in identifying them and correcting the errors which they respond to.

In his corpus analysis, Finkler identified ten classes of disfluencies, which he divided into the following four groups: (The original names of the classes are given in parentheses.)

1. Replacement, Deletion, Insertion
   (Austauschkorrektur, Löschung, Späteingabe)
2. Meta Comment, Syntactic Performance Problem (SPP), Repetition
   (Metakommentar, syntaktisches Performanzproblem, Wiederholung)
3. Stuttering, Slip Of The Tongue
   (Stotterer, Versprecher)
4. Uncorrected Errors, Rest
   (Unkorrigierte Fehler, Rest)

Representative for the fundamental structure of DFs, Finkler (1997) describes a DF of the type *replacement* in the following way: A *replacement* consists of the *original utterance*, an *editing phase* and the *continuation*:

The Original utterance is the part of the utterance, which reaches from its beginning to the point where the speaker interrupts herself. It contains the disfluency's *reparandum.*

The Editing Phase comes directly after the *interruption point.* It contains unfilled pauses, filled pauses, hesitations and/or correction introducers.

The Continuation of the actual utterance takes place after the editing phase. It contains the self-correction, the disfluency's *reparans.* Depending on the error, this last phase can look quite different.

In (Finkler, 1997) the *reparandum* (RM) is seen as the part of the *original utterance*, which is revised in the self-correction later on. There may be several other words between the *reparandum* and the *interruption point.* The *reparans* (RS) is seen in an equivalent way. It denotes the part of the *continuation*, which corrects the *reparandum.* Again there may be several words between the *editing phase* and the RS. Thus the terms do not say anything about the position of these parts in the complete utterance. This is a clear difference to the present work, where the RM reaches from the beginning of the erroneous part to the IP, which correspond to the start of the *interregnum* (= *editing phase* in Finkler's work), or the *reparans.*

Finkler defines the disfluency classes he detected in the following way: (Again, the original names of the classes are given in parentheses.)

Replacement (Austauschkorrektur): The RM is replaced by one or several substituting elements and may not be contained in the RS, except in an explicit negation.

Deletion (Löschung): One element from the *original utterance* is missing in the *continuation* and the whole utterance cannot be seen as an ellipsis.

Insertion (Späteingabe): Some element from the *original utterance* is repeated in the *continuation* and also additional information to this element is given in the *continuation.*

Meta Comment (Metakommentar): Those are sentences or sentence-like expressions, which are used for adding characterisations, constraints or an attitude to the uttered material, often in form of an ellipsis. This also includes cases where one of two well-formed sentences modifies or negates the other. Often there is no interruption in the original sentence.

SPP (Syntaktisches Performanzproblem): Speech material has been inserted between the interrupted phrase and the *editing phase* or there is an agreement error and the utterance does not fulfill the conditions for any of the other classes.

Repetition (Wiederholung): The last part of the *original utterance* was repeated in the *continuation* and a possibly existent abrupted word from the *original utterance* was fully articulated in the *continuation.* (A phenomenon is not classified as *repetition*, if the repetition is either caused by stuttering, or is a stylistic device or has any syntactic function.)

Stuttering (Stotterer): The utterance was interrupted within a syllable and there are no correction introducers, hesitations or long pauses in the *editing phase.* The correction has to start with a repetition of the interrupted word fragment.

Slip of the tongue (Versprecher): Those are parts of utterances which result from interferences in the normal speech production.

Uncorrected (Unkorrigiert): Those are utterances that contain uncorrected errors, which do not even fulfill the less stringent rules of speech grammar, or utterances, which contain obvious content errors.

Some differences between Finkler's work and the classification presented in this work may arise due to the fact that Finkler's aim was to produce human-like self-corrections, while the present work's aim is to correct disfluencies occurring in spontaneous speech. For example, the present work does not differentiate between grammar for spoken and written language, since the idea is to enable the use of written language parsers for spoken language.

Finkler also distinguishes between 1) different types of *insertions*, depending on the position, on which the new material was placed, relative to the repeated element, and 2) different types of *repetitions*, according to the position of the replacing material in the continuation. These factors might be interesting for a generation approach of self-corrections but are not as relevant in a correction approach.

The class *Meta Comment*, as it was defined by Finkler, strongly references an utterance's semantic content, which is outside the range of this work. Such phenomena are not considered as long as a sentence is syntactically well-formed. Finkler's class SPP has no direct counterpart in the present work but the examples for this class, given in (Finkler, 1997), are covered by other classes, see chapter 3.

2.3 Other Related Work

Generally, disfluency classification schemes fall into two groups: Schemes using a fine-grained classification and schemes using rather rough classifications. Rough classification schemes usually identify four different groups of disfluencies, which can be described in the following way:

Modifications: The RS is a modification of the RM. It often has a strong correspondence to the RM.

Repetitions: The RS repeats the RM.

Fresh Starts: The current utterance is abandoned and a new one is started, (which often does not correspond to the abandoned utterance).

Fillers: Phenomena which do not contribute to the meaning of the utterance, examples are *filled pauses*, *editing terms* etc.

The names used for these phenomena are various. Also the actual number of the defined classes differs. Some research considers less different categories while other studies include additional DF types, e.g. involving word fragments (Bear et al., 1993). This depends mostly on the aim of the work. Some terms used for *modifications*, are *revision* (Liu et al., 2003), *repair* (Shriberg, 2001) and *modification repair* (Heeman & Allen, 1999). *Fresh Starts* are also called *restart* (Liu et al., 2003), *deletion* (Shriberg, 1994, 1996, 2001), and *false start* (Shriberg, 1999). Other terms for *fillers* are *filled pause* (Shriberg, 1999, 2001), *cue words* (Bear et al., 1993), and *abridged repairs* (Heeman & Allen, 1999).

The *filler* class can have variable extent. Sometimes it only includes phenomena, which in the present work are referred to as *hesitation*, e.g. in (Shriberg, 1994, 1999, 2001), sometimes also *discourse markers* and *explicit editing terms* are included, e.g. in (Strassel, 2004; Heeman & Allen, 1999).

Besides the DF classification schemes given in chapters 2.1 and 2.2, another example for a more fine-grained classification scheme is the one defined in (Strassel, 2004). In this work, DFs are divided into two subgroups: *fillers* and *edit disfluencies.* The *filler* group includes *filled pauses*, *discourse markers*, *explicit editing terms*, and *asides and parentheticals.* Thus it corresponds to the range of phenomena that are gathered under the term *deletable* in the current work. Phenomena of the category *asides and parentheticals* are not considered in the present work, since they are quite hard to define exactly. *Asides* and *parentheticals* are both defined as short comments in (Strassel, 2004) that do not contribute to the content of the utterance they are enclosed in. The difference between them is that asides concern a new topic (e.g. addressing a new person entering the room), while parentheticals are breaking the flow of the original utterance with a remark on the same topic. Actually, no phenomena of either of the two types were found in the material that was used for DF type identification in the current approach.

Phenomena of the group *edit disfluencies* are the same as those considered as *revisions* in the present work. *Edit disfluencies* include *repetitions*, *revisions* (denotes the same phenomena as those that above were considered with the term "modification") and *restarts* (the same as the *restart* in the current work). Strassel (2004) analysed the structure of these DFs in a similar way to the one presented in this work in chapter 3.1.

Finally, it can be stated that all research on the topic accounts for more or less the same phenomena. They are considered from different perspectives though, which leads to differences in the classification schemes.

2.4 Comparison of Schemes

The classification schemes developed by Shriberg (1994) and Finkler (1997) were discussed in detail in chapters 2.1 and 2.2. The DF classes identified by them and the classes of the scheme presented in this work are gathered in tables 2.1 and 2.2 for an easier comparison of the schemes. Also the classification that was made by Strassel (2004) is included there. Table 2.1 lists the classes of the current scheme in the first column and their correspondences in the other schemes in the remaining columns. The table shows that predominantly *uncorrected* phenomena miss counterparts in the other approaches.

The first column in table 2.2 gives the classes of the other schemes that did not fit into the first table. This means, they have no direct correspondent in the present work, but are for instance covered by several classes in the current scheme.

There are three phenomena in the other research that are not covered by any class in the current scheme. Those phenomena were mostly excluded for the reason that they do not necessarily cause syntactic irregularity or that they require semantic analysis, which does not fall into the scope of this work. This counts e.g. for the class *Meta Comment* in Finkler's scheme.

However, if it turns out that those phenomena are frequent, it could be worth to consider to introduce a new class for them.

Present Work	**Finkler** (Finkler, 1997)	**Shriberg** (Shriberg, 1994)	**Simple MDE** (Strassel, 2004)
Deletion	Deletion		
Disruption			Incomplete SU
DM		DM	DM
EET		EET	EET
Hesitation		FP	Filled Pauses
Insertion	Insertion	INS	
Mistake			
Omission			
Order			
Other	Rest		
Repetition	Repetition	REP	Repetition
Replacement	Replacement	SUB	
Restart	DEL		Restart
SOT	SOT	ART	
Stuttering	Stuttering		

Table 2.1: The DF classes defined in the present work (column 1) and their correspondents in other work discussed in this thesis

	DF-Type	**Type in present work**
Finkler (Finkler, 1997)	Meta-Comment	(Not relevant for current approach)
	Problem of Syntactic Performance (quite various phenomena)	SOT, Restart, Replacement (to decide as the case arises)
	Uncorrected	Mistake, Omission, Order
Shriberg (Shriberg, 1994)	HYB	Combination of Deletion, Repetition, Replacement and Restart
	CON	(not relevant for current approach)
Simple MDE (Strassel, 2004)	Revision	Deletion, Replacement, Insertion
	Asides and parentheticals	(relevant only if it causes syntactic irregularity)

Table 2.2: The table shows the DF classes from other work that have no direct correspondent in the present work.

2.5 Chapter Summary

Some basic types of DFs are covered by all examined previous classification schemes. Those are fillers, repetitions, fresh starts and modifications. Only some of the schemes provide a more detailed classification. Differences in the classifications can partially be explained by researchers' different underlying motivations for their work on DFs. However, none of the presented schemes is exhaustive. Additional DF classes have to be defined in order to gain a satisfactory coverage of the existing DF phenomena, especially for errors that were not corrected by the speaker herself.

Chapter 3

Classification Scheme

This chapter presents the results of my investigations on existing DF phenomena. It both reports the general structure of DFs (3.1) and elaborates each of the identified DF classes in detail (3.2). The chapter does not only give the class definitions but discusses also critical cases of class assignment, implications for future research, and planned changes to the current classification scheme.

3.1 Disfluency Structure

Most disfluencies have the same surface structure. They consist of three parts. The first part contains the "erroneous", disfluent material, that will be replaced by the speaker. It is called *reparandum* (RM) in the present work. *Reparandum* is Latin and means "to be repaired". The RM is followed by the *interregnum* (IM), a term which is adapted from Shriberg (1994). This part is called "editing phase" in many other researcher's work. The term *interregnum* was chosen for two reasons: The first reason is that in this way the DF-parts all have Latin names, the second and more important reason is given in (Shriberg, 1994, chap. 2.3.1, p. 8):

> "Interregnum" is a more neutral term than "editing phase" (Levelt, 1983); it can be used to specify the temporal region from the end of the reparandum to the onset of the repair even if this region contains no editing term, and it does not imply an editing function for the speaker(...)

This can be the case when the IM contains only *hesitations*. *Hesitations* can be seen to have an editing function in some cases but not in all. The term is also reasonable with regard to the planned inclusion of acoustic information. By means of acoustic information also phenomena such as long unfilled pauses can be scanned.

The third part of a DF is the *reparans* (RS). *Reparans* is also Latin and means "repairing". This indicates this section's function; it corrects the disfluency of the RM.

Opposed to (Finkler, 1997) and (Levelt, 1983), but similar to (Shriberg, 1994), the *reparandum* denotes the whole stretch of material from the beginning of the DF first part to the beginning of the IM, not only the words that are replaced or corrected in the reparans. This is due to the fact that replacing the RM with the RS has to result in a meaningful, grammatically correct sentence, which would not always be the case, if only the modified parts were denoted as RM. The following example illustrates this. The italic text in (3) shows a RM according to the present work, while it shows a RM according to Levelt's and Finkler's definition in (4). The RS is written in bold face. As can be seen, replacing the RM in sentence (3) by the RS would result in a correct sentence, while it would not in sentence (4).

(3) *for two days ago I met* — **yesterday I met** my mother

(4) *for two days ago* I met — **yesterday I met** my mother

Between the RS and the IM, the *interruption point* (IP) is located. This is the point at which the utterance is interrupted and the correction is initiated. The IP does not say anything about the point at which the speaker has noticed her error, but only about the point at which she is going to do something about it. In this approach, the IP is not marked in the DF annotations, since RM, IM, and RS are marked and the IP comes directly after the RM, before the onset of the IM or RS. This means an implicit marking of the IP

Both the IM and the RS can be omitted in a disfluency. The IM can consist of *hesitations* or *explicit editing terms*, but a disfluency does not have to include such expressions. The error can be followed directly by the correction, which means that the IM does not apply in these cases. (The present work does not account for unfilled pauses.) The RS is omitted in all *uncorrected* phenomena, e.g. *order*, *omission* and *mistake*, see 3.2 for details.

3.2 Disfluency Class Definitions

The DF phenomena encountered during the corpus analysis were grouped into a set of classes. The grouping was done based on the disfluencies' surface similarity. No presumptions concerning the underlying causes for the DF were included in the classification.

Disfluencies can be divided into two subgroups: *independent phenomena* and *application dependent phenomena*. With *application dependent phenomena* cases are denoted, which can have the same effect as DFs in computer applications but which are not DFs in actual fact. Therefore they must not be included in a general classification scheme of DFs. Some examples for application dependent phenomena are named entities (NEs), slang expressions and contracted word forms (e.g. you're, couldn't etc.). It depends on the application if those things cause DF effects, e.g. on the existence of a reliable NE recognition. Therefore these phenomena would have to be defined and classified according to the approach.

The classification scheme presented here does only specify independent phenomena. Those are entities which objectively can be seen as disrupting the fluency of speech. The classes identified in the current approach are shown in figure 3.1. In the following, detailed descriptions for all classes will be given. The examples given in the descriptions are taken from the corpus.

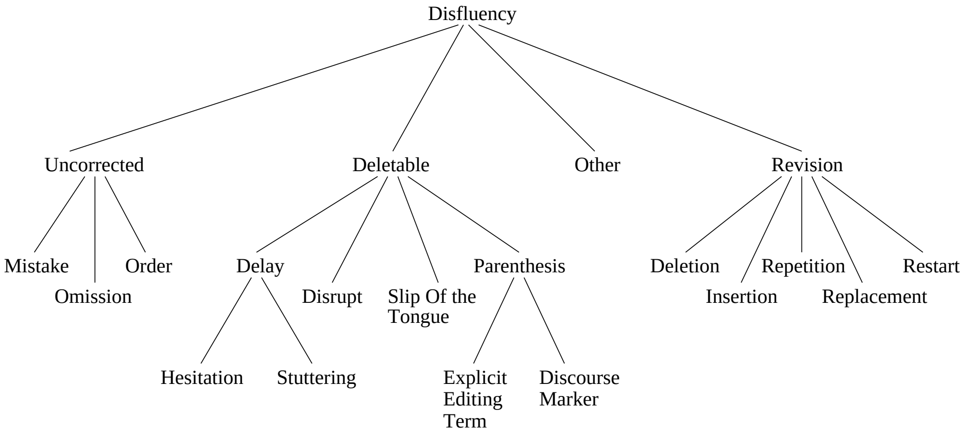


Figure 3.1: The figure displays the hierarchy of the disfluencies.

Some comments on notation: For clarification purposes the *reparandum* is enclosed by <RM> and </RM> and the *reparans* is enclosed by <RS> and </RS>. This is in the style of the XML-notation, which was used for the annotations. Furthermore, in most cases DFs are enclosed in tags representing the name of the DF's class, as they were used in the annotations.

The examples for the uncorrected phenomena are preceded by 'O' or 'C'. 'O' stands for "original", indicating the string of words originally uttered by the speaker. 'C' stands for "correction", giving the original string with an appropriate RS added to it. This only applies to the uncorrected DFs, since no RS has to be added in the other cases. There the original utterance contains the repair.

3.2.1 DFs of Type Uncorrected

Two conditions have to be fulfilled by a DF to be classified as uncorrected:

1. The speaker's original utterance may only contain a *reparandum* (RM). The *reparans* (RS) must be missing (and thus also the *interregnum* (IM)). This means, the speaker herself did not give a correction for the DF. The correction has to be created subsequently (e.g. by an annotator or a correction system).

2. The content of the RM is relevant for the sentence and may not just be left out. Therefore the DF cannot be corrected by only deleting the disfluent material. Instead, the correction must include the insertion of a suitable RS. The RM's propositional content has to be preserved in this RS.

3.2.1.1 Mistake

A *mistake* is an uncorrected speech error, which leads to a grammatically incorrect sentence. Examples for this class are agreement errors and other grammatical errors.

Examples for Mistakes:

(5) O: If it `<RM>`were`</RM>` flat on the bottom
C: If it `<mistake><RM>`were`</RM> <RS>`was`</RS></mistake>` flat on the bottom

(6) O: You know which way you're gonna `<RM>`pointing`</RM>` it
C: You know which way you're gonna `<mistake><RM>`pointing`</RM> <RS>`point`</RS></mistake>` it

Distinction from other classes:

Omission: In a *mistake* all relevant speech material was given by the speaker. The class is used for the correction of existent speech material by substituting an error with a correct form. In an *omission* some relevant speech material was not given by the speaker at all (empty RM).

Order: The *order* class is used for cases, where words occur in the wrong order. The involved word forms do not necessarily need to be changed in an *order* DF. A *mistake* always implies changes of one or several words.

Slip Of the Tongue and Stuttering: *Mistakes* are meaningful words that were used in the wrong form by the speaker and have to be corrected. In contrast, *SOTs* and *stutterings* are meaningless speech material. In order to gain a meaningful and correct sentence it has to be deleted, not corrected.

3.2.1.2 Omission

The class *omission* is for cases, in which the speaker omitted a word, which would be necessary for the segment in order to be grammatically correct. *Omissions* are quite special cases of (uncorrected) DFs, since they seemingly do not contain a RM. This means, that that part of the DF, which actually makes a phenomenon a DF, seems to be missing. In the presented approach *omissions* are treated as containing an empty RM. The RM is existent in the way, that it is the gap, which causes the utterance to be syntactically incorrect.

Typical examples for *omissions* are left out function words, e.g. articles and prepositions or personal pronouns etc. *Omissions* do not denote cases, where word stretches have to be added to adjust an utterance's semantic content or similar purposes. This would include too much interpretation, which should be avoided as far as possible. Of course, even with prepositions and articles it can be hard to decide, which one would be the most appropriate or likely addition. For example, in many cases it is almost impossible to judge whether a definite or an indefinite article should be used for example. However, the choice of the article should influence the utterance's content to an acceptable degree only.

Examples for Omissions:

(7) O: And project manager will design a better meeting
C: And `<omiss>the</omiss>` project manager will design a better meeting

Distinction from other classes:

DELETION: The difference between *omission* and *deletion* is that a *deletion* implies a self-correction, where the correcting part (the RS) leaves out some information from the original part (the RM), whereas an *omission* means that the speaker "forgot" some relevant speech material in the original utterance and did not correct herself. In a *deletion* both RM and RS are given by the speaker. In an *omission* the speaker utters neither RM nor RS, which results in an empty RM.

MISTAKE: *Mistakes* do always imply a change of the existing speech material and their RM may never be empty.

3.2.1.3 Order

Disfluencies of type *order* denote cases, in which a segment's word order has to be changed in order to make the utterance grammatically correct.

Examples for Order DFs:

(8) O: I don't know <RM>what's the idea</RM> for.
C: I don't know <RM>what's the idea</RM> <RS>what the idea is</RS> for.

Distinction from other classes:

MISTAKE: The *mistake* class implies that changes have to be made to one or several words of the existent material. This is not the case when a wrong word order was used.

3.2.2 Deletable Phenomena

The following preconditions have to be fulfilled by a DF to be classified as a deletable phenomenon:

1. The DF's content does not contribute to the meaning of the utterance. It can be discarded without impact on the utterance's statement.
2. The DF does only contain a RM and no correction, which is quite naturally following from 1, since non-contentional expressions can hardly be corrected.

Since deletable DFs are corrected by deleting them from the utterance, they are the only phenomena, which do not need a RS.

3.2.2.1 Hesitation

Hesitations are rather sounds than words. It can be assumed that they are used by the speaker in order to gain time and thus that they are expressions of the speaker's cogitation. However, independently from the underlying causes for these sounds, typical *hesitations* are: uh, uhm, eh, em, mm etc.

Examples for Hesitations:

(9) <hesit>Um</hesit> one thing I thought of

(10) And then marketing will look and see <hesit>uh</hesit> what people want.

(11) And uh then and then we are going to make

```
<repeat>
        <RM>
            And <hesit>uh</hesit> then
        </RM>
        <RS> and then </RS>
</repeat>
we are going to make
```

As example (11) shows, *hesitations* can also occur within the RM or RS of another DF, without influencing that DF's classification. Note that this only holds for cases in which the *hesitation* is surrounded by other material of the RM or RS on both sides. A *hesitation* may never stand at the beginning or end of a RM or RS. Otherwise it is either classified as part of the IM or as a discrete DF.

Distinction from other classes:

STUTTERING: Opposed to *stutterings*, *hesitations* are not related to the subsequent word. No phonetic similarity has to be given.

SLIP OF THE TONGUE: *SOTs* can be seen as an erroneous try by the speaker to phrase the next word, whereas hesitation sounds rather refer to a phase of cogitation before the speech is continuated. They are not attempts to formulate a real word.

DMS AND EETS: The difference between *DMs* and *hesitations* is that *DMs* always consist of real words, whereas *hesitations* are non-lexical speech sounds.

3.2.2.2 Stuttering

Syllables and speech sounds such as single consonants, which are similar to the beginning of the next fully articulated word, are classified as *stutterings*. The material may not form a whole meaningful word. This means, they are non-lexical word fragments. They may neither be similar to the whole next word. Such cases would be classified as *repetitions*.

The deletion of a *stuttering* may neither make the sentence ungrammatical or change its meaning. Thus words, which happen to be phonologically similar to the beginning of the next word are *not* considered as *stutterings*.

Examples for Stutterings:

(12) `<stutter>`D`</stutter>` do you have

(13) They all work on the same `<stutter>`prin`</stutter>` principle.

(14) `<stutter>`N n`</stutter>` no, I don't think so.

As (14) shows, sequences of stuttering sounds are seen as one single *stuttering* and are not treated separately.

Distinction from other classes:

SLIP OF THE TONGUE: *Stutterings* are always related to the subsequent word and have to be phonetically equal to the beginning of the following word. These conditions do not count for *SOTs*.

Hesitation: Neither are *hesitations* related to the speech material around them.

Mistake: *Mistakes* are fully articulated words that have been used by the speaker in a form that is grammatically incorrect in the current context. In contrast, a *stuttering* is never a complete word. *Stutterings* do not contribute to the segment's content either.

Repetition: Though a sound can be repeated several times in *stutterings*, it is not annotated as *repetition*. *Repetitions* may only contain whole (meaningful) words.

Discussion of the current definition:

The current annotation manual gives no example for a case, in which a *stuttering* stands between the material from RM and RS of another DF. It can be discussed, whether the *stuttering* should be treated as the end of the RM, as part of the IM or as the beginning of the RS. It could be seen as part of the RM, since it constitutes erroneous material, which is to be deleted. If such a *stuttering* is seen as having an editing function, it could be treated as part of the IM. Last it can also be seen as belonging to the RS, since it is strongly bound to the next fully articulated word and should not be separated from this. For the reason just given, the last interpretation is the preferred one in the current approach, even though it contradicts the idea, that the *reparans* should only contain correcting material. This is however not given in cases where a *stuttering* occurs in the middle of the RS either. Following is an example for one of the described cases:

(15) remote c rem remote...

```
<replace>
        <RM>remote <sot>c</sot></RM>
        <RS> <stutter>rem</stutter> remote...</RS>
</replace>
...
```

In the same way *stutterings* that stand before the onset of an erroneous region would have to be included in the RM:

(16) thi this is this would

```
<replace>
        <RM> <stutter>thi</stutter> this is</RM>
        <RS>this would</RS>
</replace>
```

Another case, which the current annotation manual does not account for, is the following: If a *hesitation* stands between the fragmentary material and the next word, it is still considered as *stuttering* (see (17)). This does not count for the case that other material, e.g. a *slip of the tongue* is situated there. This could wrongly be inferred from the declaration that the fragment has to be similar to the onset of the next *fully* articulated word. (18) [1]. is an example for such a case. Here the "u" is similar to the beginning of the next fully articulated word ("user") but it is not considered as *stuttering* anyway. Thus the definition of *stutterings* would have to be extended by the addition that no other speech material except for *hesitations* may stand between the *stuttering* and the word it belongs to.

(17) the `<stutter>`u`</stutter>` `<hesit>`uh`</hesit>` user interface

(18) the `<sot>`u`</sot>` `<sot>`fin`</sot>` user interface

3.2.2.3 Disruption

Segments are classified as *disruptions* if they do not form a meaningful statement and are so fragmentary that no meaning can be established by adding information either. (Such a case would be classified as *omission.*) Both whole segments and partial segments can be classified as *disruptions*. The latter applies, when the segment starts with a meaningful clause but its last part does not make sense (see example 20). If the fragmentary material occurs at the beginning of a segment, the phenomenon is classified as *restart* or *replacement* instead. Generally, the removal of material classified as disruption may not cause any loss of relevant information.

Examples for Disruptions:

(19) `<disrupt>`or like a`</disrupt>`

(20) Of course we're not only a electronics company `<disrupt>`but a `</disrupt>`

Distinction from other classes:

OMISSION: The *omission* class is only used for cases, where the meaning of the segment is apparent but some element is missing for the segment to be grammatically correct. With *omissions* it should also be clear, which word is missing. These things do not hold for *disruptions.*

3.2.2.4 Slip Of the Tongue (SOT)

SOTs are single speech sounds, syllable fragments or one or several syllables, which do not form a correct (existing) word and which cannot be classified

[1] Note that examples (17) and (18) are constructed, not taken from the corpus, since no such cases appeared in the examined material

as *stuttering*. This means, they may not be similar to the beginning of the next word.

Examples for SOTs:

(21) it may not be `<sot>`th`</sot>` as functional

(22) looking at the `<sot>`tex`</sot>` technical functions

(23) I'll be `<sot>`pro`</sot>` mostly dealing with properties.

Note that cases as example 23 also are classified as *SOTs*. In some other approaches it would have been treated as an *insertion*, since "pro" can be interpreted as the beginning of the later uttered "properties", which was interrupted for the insertion of additional material before saying "properties". However, since such considerations are too vague and dependent on interpretation, phenomena of this style are considered as *SOTs* in this work.

Segment (23) is also an example for a case, in which the DF-speech material could function as an own word. Such cases are classified as *SOTs* if it is unlikely, that the speaker wanted to use this word at this place. However, a certain vagueness has to be accepted in these cases.

Distinction from other classes:

Stuttering: *Stutterings* are always related to the subsequent word and have to be phonetically equal to the beginning of the following word. *SOTs* are not related to the following word.

Mistake: *Mistakes* are meaningful words that were used in the wrong form by the speaker and which have to be corrected. In contrast, *SOTs* form meaningless speech material, which is not to be corrected but to be deleted in order to gain a meaningful and correct sentence.

3.2.2.5 Discourse Marker (DM)

All of the following expressions may be classified as *discourse markers* (DMs). DMs do not contribute to the content of an utterance, but have rather a discourse related function. Their usage gives the speaker time to think of what to say next and to hold the turn. The following examples illustrate possible DMs. This list is not exhaustive.

actually
anyway
and yeah
basically
I mean
let's see
like
now
okay
see
so
well
yeah
you know
you see

It is difficult to decide whether an expression is a DM or not in many cases. This often counts for “kind of” and “sort of”. Those are used not only as DMs but also for weakening the following proposition. If “sort of” and “kind of” modify an adverb, it should be legitimate to classify them as DMs, see (25). However, a phenomenon should only be classified as DM if it is sure that it functions as a DM and does not contribute to the meaning of the utterance. This means, deleting it would not totally change the sentence’s meaning. In case of uncertainty the material should rather be left untouched than classified as a *DM*.

Examples for DMs:

(24) But, `<dm>`you know`</dm>`, they all sort of have the same functions

(25) All remotes are `<dm>`sort of`</dm>` quite similar

When “yeah” occurs as an own utterance, e.g. as a response to a statement made by another person, it is not classified as a DM.

Several consecutive DMs are considered as independent phenomena and classified one by one. As (27) shows, consecutive DMs of the same type are not interpreted as *repetitions*.

(26) Just `<dm>`sort of`</dm>` `<dm>`you know`</dm>`, your buttons

(27) I think, `<dm>`yeah`</dm>` `<dm>`yeah`</dm>` a universal remote

Distinction from other classes:

Explicit Editing Term: The above mentioned expressions are annotated as EETs only if they stand within a complex disfluency. All other cases are marked as DMs.

Hesitation: DMs consist always of real words, whereas *hesitations* are nonlexical speech sounds.

3.2.2.6 Explicit Editing Term (EET)

EETs are filler phrases the speaker uses to mark that she has made an error and is about to correct it. They can also be seen as words that are uttered in order to temporise when the speaker is planning to make a correction. Therefore an EET always stands in the IM of a revision. EETs are roughly the same expressions as DMs but can also consist of other expressions. As stated in (Strassel, 2004), one difference between EET and DM could be that DMs usually should not be terms as “sorry”, “oops” or similar expressions.

Examples for EETs:

(28) How would we go about making you know getting rid of our weak points?

```
How would we go about
                <restart>
                          <RM>making</RM>
                          <eet>you know</eet>
                          <RS>getting</RS>
                </restart>
rid of our weak points?
```

(29) The design of or the point of putting two sensors on each side

```
<replace>
         <RM>The design of</RM>
         <eet>or</eet>
         <RS>the point of</RS>
</replace>
putting two sensors on each side
```

Distinction from other classes:

Discourse Marker: DMs do not occur in IMs of revisions. DM expressions occurring in such a region are always classified as EETs.

Hesitation: EETs always consist of real words (and stand in the IM of a *revision*). *Hesitations* are non-lexical speech sounds and can occur anywhere in an utterance.

3.2.3 DFs of Type Revision

Revisions are phenomena, where both RM and RS are given by the speaker. They could also be named "self-corrections" or "self-repairs".

Revision phenomena cover the three most general forms of editing:

1. deletion of material

2. insertion of material

3. change of material (which could also be expressed as a combination of deletion and insertion)

The case "change of material" can be divided into two subgroups: partial change of material (*replacement* in this work) and total change of material (*restart* in this work).

In all *revision* classes, except for *restarts*, the revised material does not have to consist of a single stretch of words. It may contain several stretches, which are intervened by portions of non-revised, preserved material.

3.2.3.1 Deletion

The RS of a *deletion* repeats some parts of its RM, while omitting some other material. The deleted material has to be from the central region of the RM. Otherwise the DF is a *repetition* or a *replacement*.

Examples for Deletions:

(30) But it's really not it's not functional.

```
But
   <delete>
           <RM>it's really not</RM>
           <RS>it's not</RS>
   </delete>
functional.
```

Distinction from other classes:

Omission: A deletion implies a self-correction (RM and RS both uttered by the speaker). In case of an *omission* the speaker's utterance only contains a RM. Here the speaker "forgot" some relevant speech material.

Replacement: The RS of a *deletion* simply leaves out information of the RM, whereas the RS of a *replacement* replaces some of the original material with new material.

Restart: The RS of a *restart* consists of completely new material compared to the RM. The RS of a *deletion* does not contain new material at all but leaves out some of the information of the RM.

3.2.3.2 Insertion

The RS of an *insertion* repeats the RM with supplementary information added at some point. The last element(s) of RM and RS have to be similar. This means, the added information may not be the last material in the RS. Otherwise it would be classified as a *repetition*.

Examples for Insertions:

(31) What else it what else do we want it to do?

```
<insert>
        <RM>What else it</RM>
        <RS>what else do we want it<RS>
</insert>
to do?
```

A DF is not classified as *insertion*, if this classification would be based on a word, which seems to occur in both RM and RS but was not fully articulated in the RM. This would require too much interpretation of the speaker's intention. The following example clarifies the case:

(32) I'll be pro mostly dealing with properties

a. **Correct:** I'll be `<sot>`pro`</sot>` mostly dealing with properties.

b. **False:**

```
I'll be
   <insert>
           <RM>pro</RM>
           <RS>mostly dealing with properties</RS>
   </insert>
```

Distinction from other classes:

Replacement: The new material in the RS of a *replacement* is replacing some of the material of its RM. Thus some information of the RM is missing in a *replacement's* RS. The RS of an *insertion* preserves all the material of the RM and additionally contains some new information.

Restart: The RS of an *insertion* adds information to the utterance without replacing or deleting any material. The RS of a *restart* replaces all of the material of the RM with new material.

3.2.3.3 Repetition

Expressions that occur several times consecutively are classified as *repetitions.* This denotes both single words and whole phrases, but no word fragments. The class implies that RM and RS contain exactly the same material. Finkler (1997) expresses this in the following way:

> "A repetition usually is an unchanged reproduction of a segment
> of arbitrary length."

Therefore stretches as "infor information" are not classified as *repetitions.* Furthermore, nothing except for *hesitations* and EETs may stand between the repeated material.

Examples for Repetitions:

(33) Maybe we could draw it up on the on the board.

```
Maybe we could draw it up
            <repeat>
                    <RM>on the</RM>
                    <RS>on the</RS>
            </repeat>
board.
```

(34) After a while you have to point it towards the uh towards the equipment

```
After a while you have to point it
          <repeat>
                  <RM>towards the</RM>
                  <hesit>uh</hesit>
                  <RS>towards the</RS>
          </repeat>
equipment
```

Cases where an abbreviated form and a fully articulated form of an expression occur after one another are still classified as *repetitions*:

(35) You're you are the industrial designer.

```
<repeat>
        <RM>You're<RM>
        <RS>you are</RS>
</repeat>
the industrial designer.
```

Similar to Finkler (1997), this work does not classify phenomena as *repetitions*, if the repetition is a stylistic device. Consecutive *stutterings*, DMs, SOTs or EETs are not classified as *repetitions* either. *Repetitions* need to consist of proper words.

Discussion of the current definition:

Sometimes the speaker repeats an expression several times. This means, the *repetition* contains more than two instances of the repeated stretch. (36) is an example for such a case.

(36) We will look at the the the ball later.

In the present approach *repetitions* have been analysed in the way, that all instances of the repeated expression are included in the RM except for the last one, which is the RS. (37) shows an annotation example for this analysis.

(37) We will look at the the the ball later.

```
We will look at
        <repeat>
                <RM>the the</RM>
                <RS>the</RS>
        </repeat>
ball later.
```

I decided to change the analysis of such cases in the way that they are analysed as nested two-instance *repetitions*, see (38). On one hand the former analysis corresponded to the intuitional estimation that the repeated material forms one *repetition*. On the other hand the new analysis treats such cases as all other complex DFs (see chapter 4.2), which means an improvement in terms of consistency of the scheme. Furthermore, the proposal that RM and RS have to contain exactly the same material is not fulfilled in the old analysis.

(38) We will look at the the the ball later.

```
We will look at
        <repeat>
                <repeat>
                        <RM>
                                <RM>the</RM>
                                <RS>the</RS>
                        </RM>
                        <RS>the</RS>
                </repeat>
        </repeat>
ball later.
```

Another thing, which should be reviewed, is the following: The present definition of *repetitions* says that only *hesitations* and EETs may stand between the repeated material. This includes also cases, in which the RS is preceded by a *stuttering*. *Stutterings* appear to be strongly associated with the following word. In this way it would be more natural to classify such cases as *repetitions* anyway. (39a) and (39b) are examples for two different possibilities of analysing *repetitions* in combination with *stutterings*. (39a) is the preferred analysis for the reason given above: *stutterings* are strongly associated with the following word. It does not make sense to separate them from the word by treating the *stuttering* as part of the IM (see also 3.2.2.2), even if this would fit better to the claim that RM and RS of a *repetition* should be similar.

(39) to zap t to zap between channels

a.
```
<repeat>
        <RM>to zap</RM>
        <RS> <stutter>t</stutter> to zap</RS>
</repeat>
between channels
```

b.
```
<repeat>
        <RM>to zap</RM>
        <stutter>t</stutter>
        <RS>to zap</RS>
</repeat>
between channels
```

This way of treating *stutterings* in *repetitions* also makes sense with respect to the following circumstances: If such cases are not considered as *repetitions*, it is hard to decide whether *deletion* or *insertion* would be the appropriate classification. This depends on, whether the *stuttering* is considered as part of the RM (see (40)) or as part of the RS, see (41).

(40)
```
<delete>
        <RM>to zap <stutter>t</stutter> </RM>
        <RS>to zap</RS>
</delete>
between channels
```

(41)
```
<insert>
        <RM>to zap</RM>
        <RS> <stutter>t</stutter> to zap</RS>
</insert>
between channels
```

To treat such a case as *deletion* is not intuitive, since the deleted material should be in some way meaningful content. Furthermore, the *deletion* would only concern material, which is already marked for deletion by the <stutter>-tag. To treat such a case as *insertion* is not reasonable either. This would mean that the only material, the speaker wants to add, is material, which is to be deleted again as it is a *stuttering*.

3.2.3.4 Replacement

The RS of a *replacement* repeats some material of the RM but substitutes the remaining information with new material. The information may not just be left out. That would be marked as a *deletion*. It has to be replaced with other information.

Examples for Replacements:

(42) Otherwise the design of or the point of putting two sensors on both sides

```
Otherwise
        <replace>
                 <RM>the design of</RM>
                 <eet>or</eet>
                 <RS>the point of</RS>
        </replace>
putting two sensors on both sides
```

(43) Even if you designed it in some in a way that you know

```
Even if you designed it
          <replace>
                   <RM>in some</RM>
                   <RS>in a<RS>
          </replace>
way that you know
```

(44) So if there's a g a way of finding it quite easily

```
So if there's
     <replace>
              <RM>a <sot>g</sot></RM>
              <RS>a way</RS>
     </replace>
of finding it quite easily...
```

Example (44) shows that the class *replacement* also covers cases, where the replaced element is not a proper word but a word fragment.

Distinction from other classes:

Restart: The RS of a *restart* replaces all material of the RM, while the RS of a *replacement* only replaces some of the RM's content. The rest of a *replacement's* RM is preserved and repeated in its RS.

Deletion: The RS of a *deletion* only deletes material of the RM, without inserting any new information. The RS of a *replacement* instead substitutes the deleted stretch with new material.

3.2.3.5 Restart

The material in RM and RS of a *restart* is totally different. None of the RM's content is repeated in the RS. Thus the RS replaces all the information given in the RM. It restarts the segment of the sentence, which was started by the RM. In this way, the term “restart” refers to the function of the RS. Other terms, which have been used for this phenomenon as “false start” or “sentence correction”, are referring to the RM. The term “restart” was chosen because the names of the other revisions also correspond to the RS.

The classification is irrespective of the *restart's* position in the sentence. The DF does not need to imply a new start of the sentence in order to be classified as a *restart*. This is conforming with the class *deletion* in (Shriberg, 1994) (see 2.1). In her work she criticises other researchers, who only classify cases as *restart*, which concern the beginning (thus “new start”) of a sentence.

Examples for Restarts:

(45) How would we go about making getting rid of our weak points?

```
How would we go about
             <restart>
                      <RM>making</RM>
                      <RS>getting</RS>
             </restart>
rid of our weak points?
```

(46) So there are always the some restrictions

```
So there are always
           <restart>
                    <RM>the</RM>
                    <RS>some</RS>
           </restart>
restrictions
```

Distinction from other classes:

Replacement: The RS of a *restart* replaces all material of the RM, while the RS of a *replacement* only replaces some of the RM's content. The rest of a *replacement's* RM is preserved and repeated in its RS.

Deletion: The RS of a *restart* consists of completely new material compared to the RM. The RS of a *deletion* contains no new material at all but leaves out some of the information of the RM.

3.2.4 Other

DF structures that do not match any of the other classes can be classified as *other*. The class was introduced for revealing gaps in the classification scheme, which should be covered in a future extension of the scheme.

3.3 Chapter Summary

Disfluencies show a regular surface structure, which consists of a *reparandum* (RM), the part to be corrected, a *reparans* (RS), the correcting part, and an *interregnum* (IM), situated between RM and RS. It is optional and can be used for marking a planned modification of the just uttered material.

The current classification scheme includes DF phenomena from three groups: *uncorrected*, *deletable*, and *revisions*. *Uncorrected* phenomena are DFs where only the RM is given in the speaker's original utterance. The RS has to be added later by a "corrector". An *uncorrected* DF's content is relevant for the proposition of the utterance and may not be omitted in order to achieve a meaningful sentence. This does not hold for *deletable* phenomena. DFs of this type do not contain relevant material and may just be deleted from the utterance. *Revisions* are forms of a speaker's self-correction. Here both RM and RS are given by the speaker herself. The three groups of DFs were divided into a set of classes, which cover all of the encountered phenomena.

The correct and unambiguous definition of a DF class and the assignment of a DF to the appropriate class are not always trivial. Therefore adjustments of the class definitions will steadily have to be done as the research on the topic progresses.

Chapter 4

Annotation

An annotation manual was developed from the classification scheme. It contains DF class definitions, annotation instructions and a large number of annotation examples. In order to test the reliability and clearness of the DF class definitions, annotations according to the manual were done by several annotators on four meetings of the corpus. The annotations were then compared and evaluated statistically. This chapter presents the results of the evaluation (4.4) as well as general annotation issues (4.3) and the annotation standard (4.1).

4.1 Annotation Standard

The annotations were done in XML notation. Every DF class has a class tag assigned to it. Thus every encountered DF is enclosed in the tag corresponding to its class. For example, a *hesitation* is enclosed in the tag `<hesit>`, see (47), and a *repetition* is enclosed in the tag `<repeat>`, see (48):

(47) Can you draw `<hesit>`uh`</hesit>` `<hesit>`um`</hesit>` a rabbit?

(48) That's a that's a fish?

```
<repeat>
        <RM>That's a</RM>
        <RS>that's a</RS>
</repeat>
fish?
```

The following is a list of all DF classes and their corresponding tags:

Deletion	`<delete>`
Discourse Marker	`<dm>`
Disruption	`<disrupt>`
Explicit editing term	`<eet>`

Insertion	`<insert>`
Hesitation	`<hesit>`
Mistake	`<mistake>`
Omission	`<omiss>`
Order	`<order>`
Other	`<other>`
Repetition	`<repeat>`
Replacement	`<replace>`
Restart	`<restart>`
Slip of the tongue	`<sot>`
Stuttering	`<stutter>`

As (48) shows, there are also tags for marking RM and RS of a disfluency. These tags are used both for the DFs where RM and RS are given by the speaker (*revisions*) and for those where the RS is added by the annotator (*mistake* and *order*). This might seem slightly confusing but in fact it gets clear whether or not the RS belongs to the original utterance by the surrounding class tag.

The other phenomena are only enclosed by their class tag. In those cases it is clear that the enclosed material is a RM that is to be deleted in order to correct the utterance. The only exception to this are *omissions*. In *omissions* the enclosed material is a RS created by the annotator. Since this class' RM is empty, it is not marked in the annotation.

An IM is not annotated as such but marked by the annotation of the phenomena standing between RM and RS (namely EETs or *hesitations*). Neither is the *interruption point* marked explicitly. It is always located directly after the RM.

Generally, annotations show one of the following patterns:

1. `<class tag>`some speech material`</class tag>`

2. `<class tag>`
   `<RM>`disfluent material`</RM>`
   `<RS>`correction`</RS>`
   `</class tag>`

4.2 Complex Disfluencies

Originally, the DFs that are listed under *revisions* (chapter 3.2.3) in this thesis were considered as *complex DFs* in the approach. Phenomena of type *deletable* (chapter 3.2.2) and *uncorrected* (chapter 3.2.1) were considered as *simple DFs*. These terms were chosen with regard to the fact that the structure of the original utterance of *revisions* is more complex than the other types' structure. In *revisions* both RM and RS are given by the speaker, in *deletable* and *uncorrected* DFs only the RM is given. However, it was decided to change the definition of simple and complex phenomena for two reasons: Firstly, the annotators did not seem to find this differentiation helpful. In fact, it turned out to be rather confusing, since the annotation of uncorrected phenomena can include adding a RS to the DF. This results in a DF structure, which consists of RM and RS, precisely as in *revisions*. The second reason is that this definition would go against most other researchers' use of the terms "simple" and "complex". Therefore the definition of a *complex DF* was changed to what will be described in the following.

DF structures are considered as complex DFs where at least two IPs bind material that belongs to more than one DF. An example for that was (1) ("he she she went"). There the first "she" is both RS to the first DF and RM to the second.

When a DF is completely contained in the RM or RS of another DF, it is called a nested DF. Nested DFs do not cause any troubles for annotation. The annotation is simply carried out starting from the inmost DF (or DFs, in case there are several DFs on the inmost layer) and then proceeding stepwise outwards. The following example illustrates this:

(49) But then to go back to the to th s something along those things.

```
But then to go back
                <replace>
                        <RM>to the</RM>
                        <RS>
                            to
                            <sot>th</sot>
                            <stutter>s</stutter>
                            something
                        <RS>
                </replace>
along those things.
```

The only case where DFs that are completely enclosed by another DF are not annotated, are word fragments at the end of a *disruption*, see (50). This is, because it is uncertain, which phenomenon they represent. The fragment could be a *stuttering*, a SOT or just an interruption. Since the material following the *disruption* is missing, the appropriate classification is not determinable.

(50) `<disrupt>`Just something maybe if you ha`</disrupt>`

A troublesome event are *complex partially chained DFs*[1], where not all of one DF's output is the input to another DF. Example (51)[2] gives an example for this.

[1] The term goes in accordance with Shriberg (1994).

[2] taken from (Shriberg, 1994, chap. 4.3.4.8., p. 71)

(51) show me the flight the delta flight delta fare

Here “the delta flight” substitutes “the flight” by an insertion and “delta fare” replaces “delta flight”. The problem is that the first DF’s output (and second DF’s input) is not “delta flight” but “the delta flight”. This means, that “delta fare” actually replaces “the delta flight”. Thus “the” is omitted, which results in the sentence “show me delta fare” after the correction.

The classification scheme presented in this work does not provide a solution for this problem. Thus in the case of a partially chained DF some loss of information must be accepted. This disadvantage emerges due to the structure, which the XML-notation imposes.

4.3 Annotation Issues

In this section examples on phenomena will be given that can either not be analysed at all with the current scheme or that put high demands on the annotator. This can be both due to the limitations that arise from the used XML-notation and ambiguities in the utterances, the existence of several possible ways of analysing a phenomenon, or an overall complicated structure of the DF.

One case are phenomena that with the currently used annotation standard cannot be handled without any loss of information. One example for this was given in the previous section (example (51)). There the output of one DF was the input to another one (*chained DF).* In such cases the RS of the second DF always substitutes the whole RS of the first one, since it also is the RM of the second DF. This can mean information loss, if the RS of the second DF does not provide corresponding material to all of the RM’s elements (partially chained structure).

Another case, where information gets lost are anaphora. Consider the following example:

(52) the designer will get he will take care of it

If it is assumed that the sentence was restarted at the word “he”, this means, that the information, that “he” refers to the designer will be omitted. This is the way such cases are currently treated in the approach.

There are also lots of cases, where it is hard to decide whether one or the other annotation should be chosen. This problem arises due to the fact that some cases can be analysed in different ways. This applies especially to cases where a word is missing (class *omission*). For instance, the following case could be corrected either via inserting an article before “remote” or adding a plural ending to it:

(53) I can’t think of anything other than a long rectangle for remote

It can also happen that ungrammaticality emerges from a correction. The following example shows this:

(54) it has to be a `<dm>`you know`</dm>` international product

As soon as the DM is removed in order to adjust the sentence, the article should be "an" instead of "a". This is not accounted for in the current approach. However, this should not cause any troubles for parsing, since most parsers do not pay attention to the article used. Especially in spoken language applications parsers have to ignore these irregularities, since they always have to account for smaller errors that arise due to inaccuracy in the speech recognition.

In some cases it cannot clearly be decided how a DF should be solved. In the following example "there's" can either be seen as a *restart* to "you", which would result in analysis (55a), or as a *restart* of the whole segment, starting at "if". This would lead to analysis (55b).

(55) if you you there's a button...

```
a. if
   <restart>
       <reparandum>
           <repeat>
               <reparandum>you</reparandum>
               <reparans>you</reparans>
           </repeat>
       </reparandum>
       <reparans>there's</reparans>
   </restart>
   a button...
b. <restart>
       <reparandum>
           if
           <repeat>
               <reparandum>you</reparandum>
               <reparans>you</reparans>
           </repeat>
       </reparandum>
       <reparans>there's</reparans>
   </restart>
   a button...
```

After the correction, analysis (55a) would deliver the sentence "if there's a button...", whereas (55b) yields "there's a button". The propositions made by these sentences are actually quite distinct. The first one suggests the possibility that there might be a button, while the second one states that there actually *is* a button. Sometimes the syntactical structure of the rest of the utterance may indicate, which analysis should be chosen, but in other cases it may not.

However, the outcome of the correction is not always influenced in this way by different annotations. In some situations it may rather be an internal structural problem, which analysis is the best one to choose. The result of the correction will still be the same. The following segment exemplifies this. There, the outcome of the correction will be “but then I had” in both cases.

```
(56)  but then had I I had

      a. but then
         <insertion>
              <reparandum>
                   had
              </reparandum>
              <reparans>
                   <repeat>
                       <reparandum>I</reparandum>
                       <reparans>I</reparans>
                   </repeat>
                   had
              </reparans>
         </insertion>

      b. but then
         <replace>
             <reparandum>had I</reparandum>
             <reparans>I had</reparans>
         </replace>
```

The first analysis, (56a), follows the presented scheme stricter than the second one, (56b). “I had” cannot really be seen as a *replacement* of “had I” according to the scheme, since no information from the RM is replaced in the RS. The material is just ordered in a different way in the RS. However, it is discussible if the double “I” can be seen as a *repetition* apart from the structural fact that the word is repeated. From a more intuitive perspective one probably rather would prefer the second analysis, presuming that the IP lies after the first ”I” and the speaker simply wanted to adjust the order of the uttered words.

A more general issue is the question, where the RS of a DF should be set to end. This is clear for some phenomena, e.g. for *repetitions*, but in other cases it is nearly impossible to set the boundary at a certain point. Consider example (55). There the boundary was set after “there’s”, but the RS could just as well have included one or several further words. The boundary of the RS however does not have any impact on the correction, since all material from the RS is preserved. Actually, there are approaches, which do not mark the RS at all but only the regions to be deleted (Strassel, 2004). The current approach tags the RS in order to distinguish *revisions* from the other DFs, where no correction was made by the speaker.

4.4 Evaluation

Four meetings of the corpus were annotated with disfluencies in order to test the manual's applicability and clearness. Each meeting was annotated by four annotators. The annotations were then evaluated by comparing the data according to a number of different metrics (see 4.4.2). The aim with this comparison was to find the degree of agreement between the annotators. The agreement was estimated using two different statistics: The κ-statistic by Cohen and the AC1-formula by Gwet as they are presented in (Gwet, 2002). They are described in detail in chapter 4.4.1. The results derived from the evaluations are presented in chapter 4.4.3.

4.4.1 Statistics

The inter-annotator agreement was rated using both the κ-statistic and the AC1-formula (Gwet, 2002). Both statistics account for the probability of inter-rater agreement by chance. For this, they both subtract the supposable chance agreement from the total agreement, but they do this in different ways.

In the κ-statistic, the overall probability for a phenomenon to occur is included in the computation of the chance agreement probability. This means, if the overall probability for a phenomenon to appear is very high compared to other phenomena, this will increase the probability of chance agreement and thus decrease the amount of non-chance agreement. This results in a low κ-value, even if the agreement between raters seems to be very good at first sight. In contrast, the AC1-statistic does not compute the factor of chance agreement in dependency on the phenomena's likelihood to appear. This means that a constant value of total agreement of the annotators will result in a virtually constant AC1-value, no matter, how the ratings are distributed.

The κ-statistic has been widely used for estimating the agreement between raters. However, the high decline of agreement according to this statistic when the phenomena are not equally distributed, does not seem appropriate all times. In these cases the AC1-statistic is more intuitive. In the following both formulas will be given, designed for the comparison of multiple category annotations by two annotators. The following is a small "lexicon" for the signs used in the formulas:

N stands for the total number of compared annotations.

M stands for the number of categories.

i is an integer = 1,..., M

AGR$_i$ is the number of agreements of the annotators on category i.

A$_i$ is the number of annotations into category i by annotator A.

B$_i$ is the number of annotations into category i by annotator B.

1. **κ-statistic**

   $$KAPPA = \frac{p - e(\kappa)}{1 - e(\kappa)}$$

   'p' stands for the total agreement of the annotators, whereas e(κ) computes their agreement by chance. 'p' is calculated in the following way:

   $$p = \frac{\sum_{i=1}{}^{M}(AGR_i)}{N}$$

   e(κ) is derived in the following way. It computes a value between 0 and 1:

   $$e(\kappa) = \sum_{i=1}{}^{M} \left(\frac{A_i}{N}\right)\left(\frac{B_i}{N}\right)$$

2. **AC1-statistic**

   $$AC1 = \frac{p - e(\gamma)}{1 - e(\gamma)}$$

   Again, 'p' stands for the total agreement of the annotators. It is calculated in the same way as in the κ-statistic. The difference lies in the computation of the chance agreement e(γ), which here computes a value between 0 and 0.5:

   $$e(\gamma) = \frac{\sum_{i=1}{}^{M} P_i(1 - P_i)}{M - 1}$$

   where

   $$P_i = \frac{(A_i + B_i)/2}{N}$$

4.4.2 Metrics

Four different metrics were used to compare the annotations with respect to similarity. This means that the number of agreements and disagreements as well as the 'N' used in the presented formulas were gained by applying the metrics to the gathered data. In the following a description on each of the metrics will be given:

Strict comparison: Two DF annotations are equal if both annotators have marked the same stretch of material with the same disfluency type. If the DF contains RM and RS (and IM), also those have to be absolutely equal (starting and end point have to be the same).

Strict comparison without DF type: This metrics contains the same conditions as the first one besides that the phenomenon may have been annotated with different DF types. If e.g. one annotator classified the phenomenon as a *replacement* whereas the other classified it as a *restart*, the annotations would count as equal anyway. This is motivated by the existence of some relatively similar DF classes (as *replacement* and *restart*), which can be hard to distinguish.

Result oriented comparison: In this metrics the regions, which were marked for deletion by the annotators, are compared. This includes RMs, *hesitations*, *stutterings*, DMs, EETs, SOTs and *disruptions*. If the same regions are marked with one of these tags, they are counted as equal.

In this way the metrics accounts for the fact that the annotations finally would be used for corrections of the incoming material. f the same regions of a segment are erased, then the final outcome of the correction is the same, no matter, which class assignments were made.

Liberal concerning IM: This metrics compares annotations in the same way as the first metrics (strict comparison) but EETs are treated in a special way: Two annotations are counted as equal

- if they both contain an EET, which is annotated exactly in the same way
- if the boundaries of the EET are the same but the EET stands within the RM in one of the annotations

The annotations are also considered equal if the same region was labelled as EET in one annotation but as DM in the other. This means, all of the following annotations would be counted as equal:

1.
```
<reparandum>
        we should
</reparandum>
<eet>you know</eet>
<reparans>
        we want
</reparans>
```

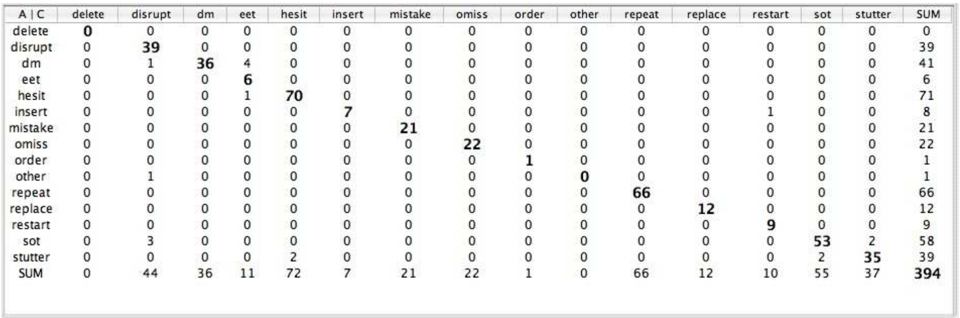
2.
   ```
   <reparandum>
           we should
           <eet>you know</eet>
   </reparandum>
   <reparans>
           we want
   </reparans>
   ```

3.
   ```
   <reparandum>
           we should
   </reparandum>
   <dm>you know</dm>
   <reparans>
           we want
   </reparans>
   ```

4.
   ```
   <reparandum>
           we should
           <dm>you know</dm>
   </reparandum>
   <reparans>
           we want
   </reparans>
   ```

4.4.3 Results

The results from the comparisons according to the different metrics were gathered in confusion matrices. Each matrix contains the comparison results of annotations by two annotators after a certain metrics. Figure 4.1 shows the confusion matrix derived by a strict comparison of the annotations by annotators A and C on meeting IS1003c.



A | C	delete	disrupt	dm	eet	hesit	insert	mistake	omiss	order	other	repeat	replace	restart	sot	stutter	SUM
delete	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
disrupt	0	**39**	0	0	0	0	0	0	0	0	0	0	0	0	0	39
dm	0	1	**36**	4	0	0	0	0	0	0	0	0	0	0	0	41
eet	0	0	0	**6**	0	0	0	0	0	0	0	0	0	0	0	6
hesit	0	0	0	1	**70**	0	0	0	0	0	0	0	0	0	0	71
insert	0	0	0	0	0	**7**	0	0	0	0	0	0	1	0	0	8
mistake	0	0	0	0	0	0	**21**	0	0	0	0	0	0	0	0	21
omiss	0	0	0	0	0	0	0	**22**	0	0	0	0	0	0	0	22
order	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	1
other	0	1	0	0	0	0	0	0	0	**0**	0	0	0	0	0	1
repeat	0	0	0	0	0	0	0	0	0	0	**66**	0	0	0	0	66
replace	0	0	0	0	0	0	0	0	0	0	0	**12**	0	0	0	12
restart	0	0	0	0	0	0	0	0	0	0	0	0	**9**	0	0	9
sot	0	3	0	0	0	0	0	0	0	0	0	0	0	**53**	2	58
stutter	0	0	0	0	2	0	0	0	0	0	0	0	0	2	**35**	39
SUM	0	44	36	11	72	7	21	22	1	0	66	12	10	55	37	**394**

Figure 4.1: Strict comparison of annotations by annotators A and C on meeting IS1003c

The numbers in the matrices show the number of agreements by the

DF type	**A**	**C**
Deletion	0	0
Disruption	51	70
DM	46	72
EET	6	13
Hesitation	72	86
Insertion	12	10
Mistake	37	35
Omission	37	28
Order	1	6
Other	5	10
Repetition	98	94
Replacement	17	48
Restart	16	28
SOT	62	69
Stuttering	45	44

Table 4.1: The table shows the total amount of annotations on a certain DF Type in meeting IS1003c by annotators A and C.

annotators in annotations on a certain disfluency type. The numbers at the line vs. column edges display the total number of phenomena, which were assigned to a certain class by one of the annotators and which have the same boundaries as a DF annotated by the other annotator according to the metrics used. Only those DF instances can be listed in the matrices, since all other phenomena are not comparable. This means, a number at the edge does not say anything about the total amount of DFs of this type annotated by the respective annotator in this meeting. To give an impression of the discrepancy between the total number of annotations on a certain DF type and the numbers given in the matrix, table 4.1 lists the total amounts of annotations on the different DF types by annotators A and C on meeting IS1003c.

For example, line 6 in the matrix (figure 4.1) displays an amount of 8 DFs of the type insertion annotated by annotator A, whereas table 4.1 shows a total amount of 12 insertions annotated by A in this meeting. The difference between those numbers rises from the fact, that the metrics are quite strict and count only DFs as equal, which have exactly the same boundaries. And only those are displayed in the matrix.

For each matrix the κ- and the AC1-value was calculated with the statistics described above (chapter 4.4.1). The total κ- and AC1-value for the annotations were then derived by calculating the average of all computed κ- and AC1-values of all meetings. By this, the results presented in table 4.2

	κ-value	AC1-value	Total agreement	Same DF type
Strict comparison	0.924	0.934	0.958	93.8 %
Liberal concerning IM	0.930	0.936	0.967	94 %

Table 4.2: The table shows the results of the calculation of the inter-annotator agreement according to both statistics for both the strict and the liberal comparison as well as the total agreement of the annotators and the percentage of DFs that were assigned to the same class.

were gained. Also the total agreement is given there (in the third column). It corresponds to the result of the p-formula used in both statistics. Column 4 shows the percentage of the DF instances that had equal boundaries and were also assigned the same DF type. There it becomes clear that only a very small percentage of phenomena with the same boundaries were not assigned to the same class.

Table 4.2 shows that once the boundaries of a DF were defined in the same way by the annotators, the agreement on the class assignment was very high. The more demanding task was rather to agree on the boundaries of a phenomenon. As mentioned before (in chapter 4.3), it can be quite hard to decide where the reparans of a DF ends. This could be one of the reasons for the difficulties in the identification of the appropriate boundaries of a DF. Another reason can be ambiguities: It is not always clear, which category a phenomenon belongs to, and the decision on the category assignment can also influence the definition of its boundaries.

To see how far the annotators agree in the selection of material that should be removed in a correction of the data and thus how equal their correction results would be, a comparison according to a result oriented metrics was made (see chapter 4.4.2). The evaluation of this metrics yielded that the annotators agreed to 77.5 % on the parts of the segments that should be deleted in order to correct them. Note that this evaluation only considered material that has to be removed for correction purposes. The classes of the category *uncorrected* were excluded from this evaluation, since their comparison can be quite hard to asses. For example, if annotator X marks that an “an” is missing, while annotator Y thinks that “the” is missing, their annotations would be judged as different, although they rightly could be seen as equal. Generally the analysis of the phenomena marked as *uncorrected* would probably need some semantic analysis of the annotations.

Altogether, the annotators identified 1205 DF instances on average in the four examined meetings. The data included a total of 792 segments. This means that the mean number of DFs per segment was 1.5, but of course there were segments that did not contain any DFs. This also means that on average there is at least one DF per dialogue act. Thus DFs are indeed quite prevalent in spoken language. The distribution of the different DF types is shown in figure 4.2. From there it can be seen that the classes are not equally distributed. There is a high discrepancy between the most common phenomenon (*hesitations*) and the scarcest one (*deletion*). The six most prevalent DF classes constitute 67 % of the encountered phenomena, whereas the five least common types correspond only to 5 % of the DF instances.

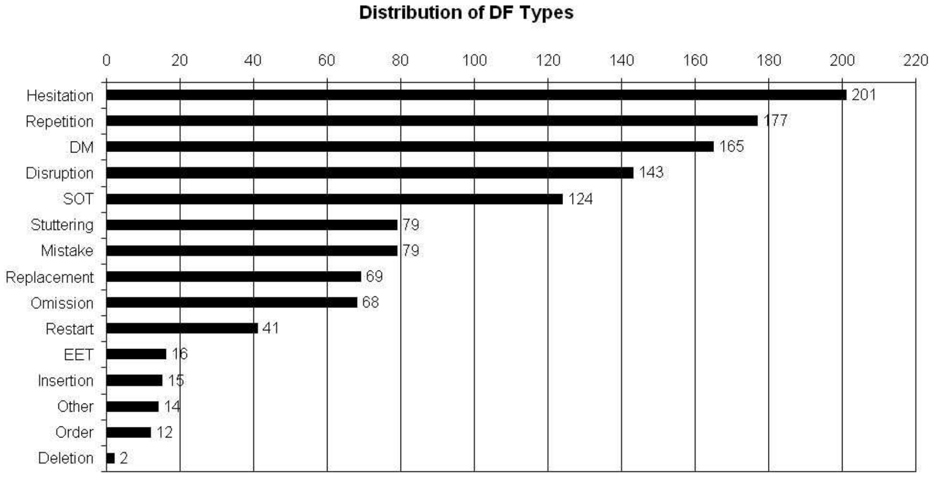


Figure 4.2: The figure displays the average distribution of the different DF types in the annotated data.

Also table 4.3 shows the number of occurrences of each DF type, together with the total and the proportional number of annotator agreement per DF type.

The annotator agreement on the classes *hesitation*, *stuttering*, SOT and *repetition* is especially high. For all these phenomena it applies that their structure is well identified, independent of the context they appear in. Even if they occur within complex multi-nested DF structures they can easily be recognised. The lowest agreement lies on the classes *disruption*, *other* and *order*. The assignment to these categories is to a high degree based on the annotator's estimation of the phenomenon. The structure of these phenomena cannot clearly be defined, since they look differently as the case arises. One also has to remember the fact, that only phenomena that were annotated with exactly the same boundaries were counted as equal. For the regarded classes it is particularly hard to say for sure where they end and start. For instance, a disruption can be seen to start at different points very often, depending on the annotators interpretation of the segment. However,

DF type	**Total Number**	**Total Agr.**	**Proportional Agr.**
Deletion	2	0	0
Disruption	143	16	0.112
DM	165	87	0.527
EET	16	7	0.438
Hesitation	202	171	0.847
Insertion	15	5	0.333
Mistake	79	27	0.342
Omission	68	24	0.353
Order	12	2	0.167
Other	14	1	0.071
Repetition	177	128	0.723
Replacement	69	27	0.391
Restart	41	10	0.244
SOT	124	97	0.782
Stuttering	79	65	0.823

Table 4.3: The second column of the table shows the number of occurrences of a certain DF type. The third column gives the total amount of average agreement on a specific DF type and the last column expresses the proportional agreement on the DF types.

such annotation differences do not necessarily have an impact on the meaning of the sentence that is gained after the correction. This means that some annotations that were counted as different in the current approach could actually be seen as equal.

This estimation would correspond to a less strict comparison of the annotations. In such a tolerant approach also phenomena could be counted as equal that overlap widely but do not have exactly the same boundaries. The presented work does not include a tolerant approach, since such a metrics is not easy to implement and some restrictions had to be made due to time limitation. The difficulties raised by a tolerant metrics arise due to the existence of complex disfluencies. They imply overlapping DFs do not always need too correspond to each other. The different layers of a complex DF do not give sufficient information on this issue either. The annotators can have defined a different number of layers in a complex DF. E.g. the inmost DF of one complex DF does not have to be the inmost DF of the other complex DF. Additionally, it could be the case that one of the annotators analysed the DF as being complex, whereas the other annotator did not.

4.5 Chapter Summary

Several annotators made disfluency annotations on four of the meetings from the corpus, according to an annotation manual that was developed on the basis of the classification scheme. The annotations follow an XML-notation with tags both for the different DF classes and for RM and RS.

The annotations were then compared according to a number of metrics and the agreement of the annotators on the DF annotations was estimated using two different kinds of statistics: the κ-statistic by Cohen and the AC1-formula by Gwet (2002). The results of these comparisons showed that the annotators agreed to a very high degree on DF class assignments of phenomena that were annotated with the same boundaries in the original text. The agreement was about 0.96 on those equally identified phenomena.

The identification of phenomena is complicated by the occurrence of complex DF structures and DFs that cannot definitely be assigned to a certain class or whose boundaries cannot be defined for sure.

Also the results were compared, which the annotations (this means corrections) would have delivered. The annotators agreed to a rate of 0.77 on the material that should have been removed in order to correct the existent irregularities.

It could be proven that DFs are quite common in spontaneous speech. In the 792 examined dialogue acts 1205 DF instances were identified, which means an average of about 1.5 DFs per dialogue act.

The DF types were not equally distributed, but some classes were predominant. The most frequent phenomena were *hesitations*. *Hesitations* together with the five subsequent most common classes correspond to 67 % of the DF instances. The five scarcest types correspond to only 5 % of the DFs.

The inter-annotator agreement on the different DF types was not equally distributed either. Annotators agreed significantly more on DF classes that have a structure, which can easily be recognised in any context.

Chapter 5

Summary and Conclusions

The aim of the present work was to develop a classification scheme for disfluencies (DFs) occurring in spontaneous speech. The term “disfluency” denotes all cases that lead to syntactical or grammatical irregularities. The scheme is supposed to serve as a theoretical basis for all applications that have to deal with such phenomena. It extends previous work that was done on the topic.

The identification of the existent phenomena was done in a data-driven approach via examinations of meeting transcriptions from the AMI meeting corpus (McCowan et al., 2005). The investigations led to an identification of 15 DF classes that were defined according to the disfluencies’ surface structure. They can be hierarchically organised and divided into three different subgroups of phenomena. Those are *uncorrected* DFs, *deletable* DFs, and *revisions*. *Uncorrected* DFs are phenomena that were not corrected by the speaker herself. This also counts for DFs of the type *deletable*, but in contrast to *uncorrected* DFs, those can simply be deleted in order to correct the utterance. For *uncorrected* DFs a correction has to be created to clear the irregularity. *Revisions* are DFs where the speaker corrected the error herself.

Also an annotation scheme for disfluencies was created from the observations of the corpus analysis. It gives detailed definitions for all of the identified classes. In order to evaluate the manual’s reliability, annotations according to the manual were done by several annotators on four meetings from the corpus. It turned out that the number of DFs identified by the annotators was quite high (1205 DFs in a total of 792 dialogue acts). This supports the suggestion that it would increase the performance of natural language applications to be able to deal with such phenomena.

The annotations were compared to a number of metrics. The metrics were quite strict and counted only phenomena as equal that were annotated with exactly the same boundaries by the annotators. On those DF instances the annotators’ agreement with respect to the DF type was very high (about

0.93). The inter-annotator agreement was measured by two different statistics: the κ-statistic and the AC1-formula (Gwet, 2002). However, they both yielded approximately the same value of agreement.

One of the metrics compared the correction results that the annotations would have delivered. It turned out that the annotators agreed to 77.5 % on the regions that should be deleted in order to correct the segments.

The evaluation showed that the occurring DF phenomena are not equally distributed. Some DF types are much more common than others. The most predominant type are *hesitations*. The most infrequent phenomenon are *deletions*. There was also a discrepancy in the accuracy of identifying the different DF types. The proportion of the similarly annotated DF instances of one type, compared to the total number of DFs of this type, varied strongly. Some types were identified much easier and more definite than others. This is assumed to depend rather on the distinct DFs' structures than on the clearness of the annotation manual, since the agreement was much higher on phenomena that have an easily recognised structure.

Generally spoken, the evaluation gave quite satisfying results, indicating that the definitions in the annotation manual were well elaborated and applicable. However, there was a number of phenomena that was not annotated with the same boundaries by the annotators. Future modifications to the manual will hopefully help to decrease this amount.

5.1 Future Work

The next step in my work on DF classification will be to revise the annotation manual according to the conclusions drawn from the evaluation of the annotations on one hand and my own observations on the other hand. This includes my considerations presented in chapter 3.2.

As stated in 1.5, annotations were only applied to segments, which could not be parsed by the parser of the LKB system (http://www.delph-in.net/lkb). In future annotation all segments of a meeting will be attended. Also parsable segments can contain DFs and the fact that a sentence is parsable does not mean that the found parse is correct. Applying DF annotation (and thus correction) to all segments might increase the percentage of correct parses. However, this is just a suggestion, which has to be proved.

Furthermore, the inclusion of acoustic information would be helpful for the investigations in several ways. It might e.g. help to decide on an appropriate DF class assignment in ambiguous cases. It probably also leads to the identification of new classes such as long unfilled pauses. They cannot be identified by the examination of transcribed speech as long as the transcription does not contain any indications on non-lexical events. Nevertheless, they can have impact on DF processing. For example, long pauses might function as an indicator for the speaker's detection of a DF. Thus they might occur especially often in the IM of a DF. If this can be evidenced, it would help e.g. a computational listener to identify a DF. However, these suggestions have to be investigated and proved before any propositions can be made on this field.

The final goal with this DF classification is to develop a computational tool for the automatic detection and correction of DFs occurring in spontaneous speech. This tool could then be integrated in a natural language application and function e.g. as a preprocessor for parsing.

References

Bear, J., Dowding, J., Shriberg, E., & Price, P. (1993). *A system for labeling self-repairs in speech* (Tech. Rep.). Stanford Research International. (Technical Note 522)

Finkler, W. (1997). *Automatische selbstkorrektur bei der inkrementellen generierung gesprochener sprache unter realzeitbedingungen.* Unpublished doctoral dissertation, Saarland University.

Gwet, K. (2002). *Kappa statistic is not satisfactory for assessing the extent of agreement between raters.* Series: Statistical Methods For Inter-Rater Reliability Assessment, No. 1.

Heeman, P. A., & Allen, J. F. (1999). Speech repairs, intonational phrases and discourse markers: Modeling speakers' utterances in spoken dialogue. *Computational Linguistics*, *25*(4), 527-571.

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

Liu, Y., Shriberg, E., & Stolcke, A. (2003). Automatic disfluency identification in conversational speech using multiple knowledge sources. In *Proceedings EUROSPEECH* (p. 957-960). Geneva.

McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., Post, W., Reidsma, D., & Wellner, P. (2005). The ami meeting corpus. In *Proceedings of Measuring Behaviour 2005 symposium on Annotating and Measuring Meeting Behavior.* Wageningen, The Netherlands.

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

Shriberg, E. (1996). Disfluencies in switchboard. In *Proceedings of the International Conference on Spoken Language Processing* (p. 11-14). Philadelphia, PA.

Shriberg, E. (1999). Phonetic consequences of speech disfluency. In *Proceedings of the International Congress of Phonetic Sciences* (p. 619-622). San Francisco.

Shriberg, E. (2001). To 'errrr' is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, *31*, 153-169.

Strassel, S. (2004). *Simple metadata annotation specification.* Linguistic Data Consortium.