A Comprehensive Disfluency Model for Multi-Party Interaction*

Jana Besser and **Jan Alexandersson**
DFKI GmbH
Stuhlsatzenhausweg 3
D-66123 Saarbrücken
GERMANY
{jbesser,janal}@dfki.de

Abstract

We present a disfluency model derived from analysing transcriptions of the AMI meeting corpus. Our model goes beyond previous work in that it discriminates several classes that are elsewhere regarded the same. Furthermore, we provide a formal account for naturally occurring phenomena that are rarely modeled in other schemes. Our annotations show significant occurrences of these classes. An evaluation of the annotations from four different annotators reveals a high agreement, $\kappa = 0.92 - 0.93$, $AC1 = 0.93$.

1 Introduction

Speech differs highly from written language. Spoken language contains a lot of linguistic irregularities, so called *disfluencies* (Henceforth DF), e.g., (Shriberg, 1994). In general, disfluencies can be classified on different levels, but in this work, we will solely treat syntactic and grammatical errors according to standard syntax and grammar. Hence, we present a classification scheme for speech DFs that defines DF classes according to their surface structure.

Previous approaches have failed to cover the existent phenomena to a satisfying degree. To our knowledge, the presented scheme is more fine-grained than previous schemes and covers a larger set of DF types. In fact, this scheme models almost 99% of the phenomena found in our corpus.

In a data-driven approach, we identified the existing phenomena via examinations of meeting transcriptions from the AMI[1] meeting corpus (McCowan et al., 2005). The corpus contains unrestricted and uncontrolled human-human discussions, recorded in business meetings. The meetings were held in English, but not all participants were native speakers.

We consider 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. Our 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.

The disfluency classification scheme was developed as part of the AMI project. The project's goal was to develop technology to support and enrich communications between individuals and groups of people. 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*, see also (McCowan et al., 2005). The project was, e.g., concerned with automated meeting summarizations. Disfluency detection and correction is a nearly mandatory matter for

* This research is funded by the EU 6th Framework Program under grants FP6-506811 (AMI) FP6-033502 (i2home) The responsibility lies with the authors.

[1] AMI = "Augmented Multi-party Interaction", see http://www.amiproject.org and its successor AMIDA = "Augmented Multi-party Interaction with Distance Access", see http://www.amidaproject.org.

Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, pages 182–189,
Antwerp, September 2007.

reaching this goal.

The paper is organised as follows: In the next section (2) the classification scheme is thoroughly described. Section 3 presents a scheme for DF annotations in XML format. In section 4 an evaluation of DF annotations according to some metrics is conducted. In 5 we present and discuss previous work. Finally, we conclude the paper with section 6.

2 A Classification Scheme

This section will give definitions for all DF classes that we have identified for the classification scheme. For some classes, XML-annotated examples will be presented. The annotations follow an annotation scheme for DFs that we have developed based on the DF classifications (see 3).

```
lorem ipsum
<DF>
    <RM>erroneous material</RM>
    <interregnum>editing material</interregnum>
    <RS>correction</RS>
</DF>
consectetuer adipiscing elit.
```

Figure 1: The general schema of a disfluency consists of the disfluency material—*reparandum* (RM)—followed by the *interregnum* (IM). The third part called *reparans* (RS) constitutes the actual repair.

Beforehand, we illustrate the general surface structure of a DF, see figure 1: DFs usually consist of three parts. The first part contains the "erroneous", disfluent material, that will be corrected later on, the *reparandum* (RM). The RM is followed by the *interregnum* (IM), a term which is adapted from (Shriberg, 1994). The third part of a DF is the repairing section, the *reparans* (RS). The RM denotes the whole stretch of material from the beginning of the DF's 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 DFs are grouped into three sets based on their surface similarity: *uncorrected* DFs, *deletable* phenomena, and *revision*s, see figure 2. Only *revision*s can optionally contain an IM whereas RS is omitted in all *uncorrected* phenomena. We divided the *deletable* DFs into two subgroups: *delay* and *parenthesis*. DFs of type *delay* are sounds, not words, that hold up the speech flow, e.g. for gaining time to plan the utterance. *Parenthesis* DFs are real words that do not contribute to the utterance's meaning.

In what follows, we provide definitions for all DFs and examples for some:

2.1 Uncorrected

The following two conditions have to be fulfilled by a DF to be classified as uncorrected:

1. The speaker's original utterance may only contain a RM. The RS (and thus the IM) is missing.
2. The content of the RM is relevant for the sentence and may not just be deleted. Therefore, the correction of the DF implies creating a suitable RS.

There are three types of uncorrected utterances:

Mistake: A *mistake* is an uncorrected speech error, which leads to a grammatically incorrect sentence. Examples are agreement errors and other grammatical errors.
Omission: The speaker omitted a word, which would be necessary for the segment in order to be grammatically correct.
Order: The segment's word order has to be changed in order to make the utterance grammatically correct.

2.2 Deletable

The following two conditions have to be fulfilled by a DF to be classified as uncorrected:

1. The DF's content can be discarded from the utterance without impact on the utterance's propositional content.
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.

There are six types of *deletables*. The types *Hesitation* and *stuttering* are grouped into *Delay*, and *EET* and *DM* are grouped into the class *Parenthesis*.

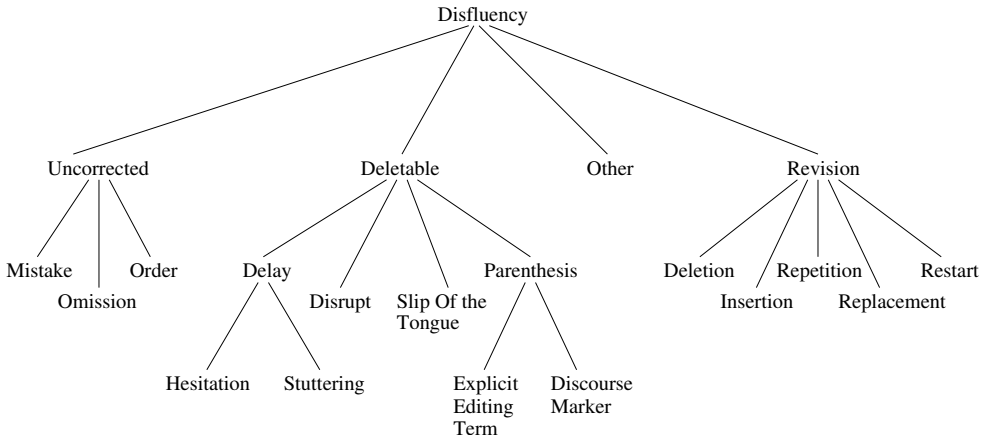

Figure 2: The hierarchy of our disfluencies where the classes are grouped into three main branches, *uncorrected*, *deletable* and *revision*s. The classes *stuttering* and *hesitation* are specializations of *delay* and *EET* and *DM* are specializations of *Parenthesis*.

Hesitation: Hesitations are rather sounds than words. They are usually used in order to gain time and are thus expressions of the speaker's cogitation. Typical *hesitation*s are: uh, uhm, eh, em, mm etc.

Stuttering: Stutterings are non-lexical word fragments, which are similar to the beginning of the next fully articulated word.

Example:

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

As the example shows, sequences of stuttering sounds are seen as one single *stuttering* and are not treated separately.

Disruption: Denotes whole or partial segments that do not form a meaningful statement and are so fragmentary that no meaning can be established by adding information. The fragmentary material may not occur at the beginning of a segment.

Slip Of the Tongue (SOT): *SOT*s are speech sounds, syllables or syllable fragments which do not form a correct (existing) word and cannot be classified as *stuttering*.

Example:

(2) looking at the <sot>tex</sot> technical functions...

Discourse Marker (DM): *DM*s do not contribute to the content of an utterance, but have a rather discourse related function. Their usage gives the speaker time to think of what to say next and to hold the turn. Examples are: I mean, so, well, you know, like etc.

Explicit Editing Term (EET): *EET*s are roughly the same expressions as *DM*s but they always stand in the IM of a *revision*.

Example:

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

2.3 Revisions

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

Deletion: The RS repeats some parts of its RM, while omitting some other material. The deleted material has to be from the central region of the RM.

Example:

(4) But
<delete>
<RM>it's really not</RM>
<RS>it's not</RS>
</delete> functional.

Insertion: The RS repeats the RM with supplementary information added at some point. The added information may not be the last material in the RS.
Example:

(5) <insert>
<RM>What else it</RM>
<RS>what else do we want it<RS>
</insert>
to do?

Repetition: Those are expressions that occur several times consecutively. This does not include word fragments. RM and RS have to contain exactly the same material.
Replacement: The RS repeats some material of the RM. The remaining information is substituted with new material.
Restart: The RS replaces all the information given in the RM. It restarts the region of the sentence, which was started by the RM. The *restart* does not have to occur at the beginning of the sentence.
Example:

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

Other: Those are DF structures that do not match any of the specified classes.

2.4 Complex Disfluencies

DFs are called *complex* if some of the contained material belongs to more than one DF. An example is shown in (7) where the first "she" is both RS to the first DF and RM to the second.

(7) he she she went

When a DF is completely contained in the RM or RS of another DF, it is called a *nested* DF. The annotation is simply carried out starting from the inmost DF and then proceeding stepwise outwards:

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

Troublesome events are *complex partially chained DFs* (Shriberg, 1994), where not all of one DF's output is the input to another DF, see (9)[2].

(9) 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 complication 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 resulting in the corrected sentence "show me delta fare".

This arises due to the fact that our annotations are made from left to right. Our annotation scheme does not yet provide a solution for this. Thus, in the case of a partially chained DF some loss of information must be accepted, see (Shriberg, 1994) for a discussion on this issue.

3 Annotation

In order to evaluate the reliability and clearness of the DF class definitions, we have annotated a subset of four meetings from the AMI meeting corpus (McCowan et al., 2005) based on an annotation manual we developed. The meetings contained a total of 2876 segments as identified during dialogue act (DA) annotation. These 2876 segments were parsed with the LKB parser (Copestake, 2002). The 792 segments (27.5%) that did not receive a parse were extracted and considered for manual annotation by four annotators. On average, 74% of these 792 segments received a DF annotation. In what follows, we call these segments "corpus A".

At the time of writing, the four meetings used in creating corpus A have been completely reannotated. Additionally, three more meetings have been annotated. For these annotations, the complete meetings were considered for annotation by the annotators. In total these meetings contain 4718 segments. 2095 segments, corresponding to 44% (ranging from 28% to 52%), were annotated with at least one disfluency.

3.1 Statistics and Metrics

We have applied two different statistics in order to rate the inter-annotator agreement: the κ-statistic

[2] taken from (Shriberg, 1994)

and the AC1-formula (Gwet, 2002). The reason for using AC1 is that it is insensitive to disproportionate distribution of class frequencies. Otherwise they share the same co-domain.

The formulae have been adapted to the comparison of multi-category annotations by two annotators. There, N stands for the total number of compared annotations, M is the number of categories, i is an integer (1,..., M), AGR_i is the number of agreements on category i, A_i is the number of annotations into category i by annotator A, and B_i is the number of annotations into category i by annotator B.

κ-statistic

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

p is the total agreement of the annotators, whereas e(κ) computes their agreement by chance (value between 0 and 1). p and $e(\kappa)$ are calculated in the following way:

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

AC1-statistic

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

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

$$e(\gamma) = \frac{\sum_{i=1}^{M} P_i(1 - P_i)}{M - 1}, P_i = \frac{(A_i + B_i)/2}{N}$$

The number of agreements and disagreements as well as the number of compared annotations (N) were gained by applying the following four metrics to the gathered data:

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.

Strict comparison without DF type: The conditions are the same as for the first metrics, but the annotated DF type may be different. If, e.g., annotator A classified the phenomenon as a *replacement* whereas B classified it as a *restart*, the annotations would count as equal anyway. This is motivated by the existence of some relatively similar DF classes, 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, *hesitation*s, *stutterings*, *DM*s, *EET*s, *SOT*s and *disruption*s. 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 if 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 *EET*s are treated in a special way: Two annotations containing an *EET* are also counted as equal, if the boundaries of the *EET*s are the same but the *EET* is annotated as part the RM in both or in one of the annotations . The annotations are also considered equal if the a region was labelled as *EET* in one annotation but as *DM* in the other.

It should be noted that *uncorrected* DFs were excluded from the result-oriented evaluation, since the comparison of their corrections can be quite hard to assess and would often some semantic analysis. For example, if annotator A adds "an" as missing determine (RS), and annotator B "the", their annotations are different from a shallow perspective, but they could be seen as equal regarding functional perspective.

4 Evaluation

The results from the comparisons according to the different metrics were gathered in confusion matrices. We then calculated the κ- and the AC1-value for each matrix with the statistics described above. The total agreement was derived by calculating the average of all computed κ- vs. AC1-values of all meetings. This gave the results presented in table 1. Column 4 shows the percentage of the DF instances that had equal boundaries and were also assigned the same DF type. It becomes clear that once the annotators identified the same boundaries for a

Table 1: Inter-annotator agreement according to both statistics for strict and liberal comparison, the total agreement, and the percentage of DFs that were assigned to the same class.

	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 %

DF, the agreement on the class assignment was very high. The demanding task was rather to agree on the boundaries of a phenomenon. It can, e.g., be quite hard to decide where the reparans of a DF ends. Also the decision on the class assignment of a phenomenon can influence the definition of its boundaries.

Additionally, we have computed the AC1- and κ-value for the three main classes in the DF hierarchy: *uncorrected*, *deletable* and *revision*. We received a κ-value of 0.998 and an AC1-value of 0.999 for the strict comparison. Thus, the annotators agreed invariably on the DF assignment to the main classes.

The evaluation of the result-oriented metrics yielded that the annotators agreed to 77.5 % on the material that would have to be removed for correction purposes.

Altogether, the annotators identified an average of 1206 DFs in the 792 segments. This means that the mean number of DFs per DA was 1.5.

Table 2 shows the number of occurrences of each DF type, along with the total and proportional annotator agreement for each class. The DF classes are not equally distributed and there is a high discrepancy between the most common phenomenon (*hesitation*s) and the scarcest one (*deletion*). The six most prevalent DF classes constitute 67 % of the encountered phenomena, whereas the five least common types correspond to only 5 % of the DF instances.

Classes rarely mentioned in previous schemes, e.g., *mistake* and *omission* are prevalent in our corpus. However, *order* only occurs in about 1% of the annotated segments. (Finkler, 1997) considers these three phenomena as one: "uncorrected". However, our findings support the division. Finally, *disruption*s are very common but seem to be hard to annotate reliably. A similar low reliability is found for *order*. This is probably due to their inhomogeneous structure. However, it is our hope that an annotator will improve the performance over time.

Table 2: The average number of annotations of a certain DF type in corpus A and corpus B. "**%**" depicts the proportion of a certain DF-type in the corpus and "**% Agr**" depicts the percentage of cases in which all four annotators agreed on the DF annotation.

Corpus	**A**			**B**	
DF	$\Sigma/4$	**%**	**% Agr**	Σ	**%**
Delete	2	0.0	0.0	2	0.0
Disrupt	143	11.9	11.2	509	11.9
DM	165	13.7	52.7	642	15.0
EET	16	1.3	43.8	43	1.0
Hesit	202	16.8	84.7	842	19.7
Insert	15	1.2	33.3	38	0.8
Mistake	79	6.6	34.2	259	6.0
Omiss	68	5.6	35.3	276	6.4
Order	12	1.0	16.7	32	0.7
Other	14	1.2	7.1	44	1.0
Repeat	177	14.7	72.3	641	15.0
Replace	69	5.7	39.1	165	3.8
Restart	41	3.4	24.4	190	4.4
SOT	124	10.3	78.2	366	8.5
Stutter	79	6.6	82.3	223	5.2
Σ	1206	100	—	4272	100

4.1 Discussion

The annotator agreement on the classes *hesitation*, *stuttering*, SOT and *repetition* is especially high. The structure of these phenomena is easy to identify, independent of their context. Even if they occur within complex multi-nested DF structures. 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. Moreover, the structure of these phenomena is inhomogeneous and cannot clearly be defined. Furthermore, we counted only phenomena as equal that were annotated with ex-

actly the same boundaries. For the regarded classes it is particularly hard to say for sure where they end and start. Annotation differences though do not necessarily have an impact on the meaning of the sentence after the correction has been applied, since different annotations can still result in the same correction.

Such facts could be accounted for via a less strict comparison of the annotations. Phenomena that overlap widely but do not have exactly the same boundaries could be counted as equal. The presented work does not include such an approach, since we could not implement a corresponding metrics due to time limitation. Such tolerant metrics is complicated by the existence of complex disfluencies. They imply that overlapping DFs do not always need to correspond to each other. They can even be assigned to different layers of a complex DF. The inmost DF of one complex DF does not have to be the inmost DF of another annotator's (complex) DF.

5 Related Work

Several researchers have investigated speech disfluencies before with different underlying motivations. There are four basic types of disfluencies that have been identified by most previous classification schemes, e.g. (Liu et al., 2003), (Shriberg, 1999), (Heeman and Allen, 1999), and (de Mareüil et al., 2005). Those are *filler*s (e.g. *filled pause*s, *discourse marker*s, and *editing term*s), *repetition*s, *fresh start*s and *modification*s. *Fresh start*s denote cases in which an utterance is abandoned and a new one is started. *Modification*s are self-corrections, in which the RS modifies the RM and has a strong correspondence to the RM.

Only some schemes go beyond this classification. One of them was developed in (Shriberg, 1994). Her thesis is an absolute foundation in this research field. She elaborated regularities in the production of DFs and created a detailed classification scheme of DF phenomena. The scheme has been adapted by several other approaches, for example by (Zechner, 2001) and (Strassel, 2004). Zechner has summarization in mind whereas the main motivation in (Strassel, 2004) is rich metadata annotation for the production of maximally readable transcripts. Another valuable and elaborate classification—also based on the findings in (Shriberg, 1994)—is presented in (Finkler, 1997). His main motivation is the incremental generation of natural language utterances.

Although some of these schemes are quite elaborated, they do not give a formal account for all disfluency phenomena occurring in our corpus. For example, in (Shriberg, 1994), no DFs were considered 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 our scheme are not covered by her classification. These phenomena have been mentioned in (Carbonell and Hayes, 1983), but are only informally described.

We also applied changes to some prevalent definitions of certain DF phenomena. An example for this is *repetition*. In Shriberg's approach these include also cases, where the first element of the repetition (the RM) is a word fragment or a mispronunciation. Our work is more rigid: a DF is only classified as *repetition* in case the RM consists of full words and RM and RS contain exactly the same material. Fragments are instead modelled in *stuttering*, *SOT* and *replacement*.

Moreover, our schema is more fine grained than the related work mentioned here. This concerns e.g. the *uncorrected* classes and the class *disruption*. Some schemata, do not differenciate between our *stuttering* and *slip-of-the-tongue* either.

6 Conclusions

Our aim has been to develop a classification scheme for disfluencies occurring in spontaneous speech. With the goal of serving as a theoretical basis for all applications that have to deal with such phenomena, our scheme extends previous work on this topic, e.g., (Shriberg, 1994; Finkler, 1997; Strassel, 2004; Heeman and Allen, 1999).

We identified the existent phenomena by examining transcriptions of business meetings from the AMI meeting corpus (McCowan et al., 2005). Our investigations led to an identification of 15 DF classes that we defined according to the disfluencies' surface structure. We developed a hierarchy of disfluencies and divided them into three subgroups. The subgroups are *uncorrected* DFs, *deletable* DFs,

and *revision*s. *Uncorrected* DFs are phenomena that were not corrected by the speaker. For these DFs, a correction has to be created to eliminate the irregularity. *Deletable* DFs are removed in order to correct the utterance. *Revision*s are DFs where the speaker made a self-correction.

We also developed an annotation manual for disfluencies. Four annotators annotated 792 segments from the AMI meeting corpus that could not be parsed by the LKB parser. It turned out that the number of DFs identified by the annotators was quite high (1206 DFs in a total). This supports the fact that disfluencies are very common in spontaneous speech. On the other hand, this might be due to the high number of non-native speakers in our corpus.

We defined four metrics for comparing the annotations. The metrics counted only phenomena as equal that were annotated with exactly the same boundaries. Annotations with the same boundaries showed a high agreement (0.93) with respect to the DF type. We also computed the agreement for the three main classes in the DF hierarchy. There we yielded a score of 0.999. The inter-annotator agreement was measured by the κ-statistic and the AC1-formula (Gwet, 2002). In this experiment, they both yielded approximately the same value. The result-oriented metrics, comparing the output of the annotations, gained 77.5% agreement.

Our evaluation showed that the DFs are not equally distributed ranging from 16.8% (*hesitation*) to approximately 0% (*deletion*). There is also a discrepancy in the accuracy of identifying the different DFs. The proportion of identically annotated DFs varied strongly. We attribute this to the DF structures rather than to the clearness of the annotation manual. This is motivated by the fact that the agreement was much higher for phenomena that have an easily recognised structure.

Future work will include more annotation of complete meetings and an evaluation thereof. The manual has already received some update, and we expect this to happen again. We plan to publish the annotations along with the complete AMI/AMIDA corpus.

References

Jaime G. Carbonell and Philip J. Hayes. 1983. Recovery strategies for parsing extragrammatical language. *Comput. Linguist.*, 9(3-4):123–146.

Ann Copestake. 2002. Implementing typed feature structure grammars. CSLI Publications, Stanford, CA.

Philippe Boula de Mareüil, Benoît Habert, Frédérique Bénard, Martine Adda-Decker, Claude Barras, Gilles Adda, and Patrick Paroubek. 2005. A quantitative study od disfluencies in french broadcast interviews. In *Proceedings od DiSS '05*, pages 27–32, Aix-en-Provence, France, September.

Wolfgang Finkler. 1997. *Automatische Selbstkorrektur bei der inkrementellen Generierung gesprochener Sprache unter Realzeitbedingungen*. Ph.D. thesis, Saarland University.

Kilem Gwet. 2002. Kappa Statistic is not Satisfactory for Assessing the Extent of Agreement Between Raters. Series: Statistical Methods For Inter-Rater Reliability Assessment, No. 1. `http://www.stataxis.com`.

Peter A. Heeman and James F. Allen. 1999. Speech Repairs, Intonational Phrases and Discourse Markers: Modeling Speakers' Utterances in Spoken Dialogue. *Computational Linguistics*, 25(4):527–571.

Yang Liu, Elizabeth Shriberg, and Andreas Stolcke. 2003. Automatic Disfluency Identification in Conversational Speech Using Multiple Knowledge Sources. In *Proceedings EUROSPEECH*, pages 957–960, Geneva.

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

Elizabeth Shriberg. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. thesis, University of Berkeley, California.

Elizabeth Shriberg. 1999. Phonetic Consequences of Speech Disfluency. In *Proceedings of the International Congress of Phonetic Sciences*, pages 619–622, San Francisco.

Stephanie Strassel. 2004. Simple Metadata Annotation Specification V6.2. Linguistic Data Consortium. http://www.ldc.upenn.edu/Projects/MDE.

Klaus Zechner. 2001. *Automatic Summarization of Spoken Dialogues in Unrestricted Domains*. Ph.D. thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, November.