Manual Annotation of Opinion Categories in Meetings

Swapna Somasundaran[1], Janyce Wiebe[1], Paul Hoffmann[2], Diane Litman[1]
[1]Department of Computer Science, University of Pittsburgh, Pittsburgh, PA 15260
[2]Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA 15260
{swapna,wiebe,hoffmanp,litman}@cs.pitt.edu

Abstract

This paper applies the categories from an opinion annotation scheme developed for monologue text to the genre of multiparty meetings. We describe modifications to the coding guidelines that were required to extend the categories to the new type of data, and present the results of an inter-annotator agreement study. As researchers have found with other types of annotations in speech data, inter-annotator agreement is higher when the annotators both read and listen to the data than when they only read the transcripts. Previous work exploited prosodic clues to perform automatic detection of speaker emotion (Liscombe et al. 2003). Our findings suggest that doing so to recognize opinion categories would be a promising line of work.

1 Introduction

Subjectivity refers to aspects of language that express opinions, beliefs, evaluations and speculations (Wiebe et al. 2005). Many natural language processing applications could benefit from being able to distinguish between facts and opinions of various types, including speech-oriented applications such as meeting browsers, meeting summarizers, and speech-oriented question answering (QA) systems. Meeting browsers could find instances in meetings where opinions about key topics are expressed. Summarizers could include strong arguments for and against issues, to make the final outcome of the meeting more understandable. A preliminary user survey (Lisowska 2003) showed that users would like to be able to query meeting records with subjective questions like "Show me the conflicts of opinions between X and Y" , "Who made the highest number of positive/negative comments" and "Give me all the contributions of participant X in favor of alternative A regarding the issue I." A QA system with a component to recognize opinions would be able to help find answers to such questions.

Consider the following example from a meeting about an investment firm choosing which car to buy[1]. (In the examples, the words and phrases describing or expressing the opinion are underlined):

(1)[2] OCK: *Revenues of less than a million and losses of like five million you know that's pathetic*

Here, the speaker, OCK, shows his strong negative evaluation by using the expression "That's pathetic."

(2) OCK: *No it might just be a piece of junk cheap piece of junk that's not a good investment*

In (2), the speaker uses the term "just a piece of junk" to express his negative evaluation and uses this to argue for his belief that it is "not a good investment."

(3) OCK: *Yeah I think that's the wrong image for an investment bank he wants stability and s safety and you don't want flashy like zip-*

[1] Throughout this paper we take examples from a meeting where a group of people are deciding on a new car for an investment bank. The management wants to attract younger investors with a sporty car.

[2] We have presented the examples the way they were uttered by the speaker. Hence they may show many false starts and repetitions. Capitalization was added to improve readability.

Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006, pages 54–61,
Sydney, July 2006.

ping around the corner kind of thing you know

The example above shows that the speaker has a negative judgment towards the suggestion of a sports car (that was made in the previous turn) which is indicated by the words "wrong image." The speaker then goes on to positively argue for what he wants. He further argues against the current suggestion by using more negative terms like "flashy" and "zipping around the corner." The speaker believes that "zipping around the corner" is bad as it would give a wrong impression of the bank to the customers. In the absence of such analyses, the decision making process and rationale behind the outcomes of meetings, which form an important part of the organization's memory, might remain unavailable.

In this paper, we perform annotation of a meeting corpus to lay the foundation for research on opinion detection in speech. We show how categories from an opinion (subjectivity) annotation scheme, which was developed for news articles, can be applied to the genre of multi-party meetings. The new genre poses challenges as it is significantly different from the text domain, where opinion analysis has traditionally been applied. Specifically, differences arise because:
1) There are many participants interacting with one another, each expressing his or her own opinion, and eliciting reactions in the process.
2) Social interactions may constrain how openly people express their opinions; i.e., they are often indirect in their negative evaluations.
We also explore the influence of speech on human perception of opinions.

Specifically, we annotated some meeting data with the opinion categories Sentiment and Arguing as defined in Wilson and Wiebe (2005). In our annotation we first distinguish whether a Sentiment or Arguing is being expressed. If one is, we then mark the polarity (i.e., positive or negative) and the intensity (i.e., how strong the opinion is). Annotating the individual opinion expressions is useful in this genre, because we see many utterances that have more than one type of opinion (e.g. (3) above). To investigate how opinions are expressed in speech, we divide our annotation into two tasks, one in which the annotator only reads the raw text, and the other in which the annotator reads the raw text and also listens to the speech. We measure inter-annotator agreement for both tasks.

We found that the opinion categories apply well to the multi-party meeting data, although there is some room for improvement: the Kappa values range from 0.32 to 0.69. As has been found for other types of annotations in speech, agreement is higher when the annotators both read and listen to the data than when they only read the transcripts. Interestingly, the advantages are more dramatic for some categories than others. And, in both conditions, agreement is higher for the positive than for the negative categories. We discuss possible reasons for these disparities.

Prosodic clues have been exploited to perform automatic detection of speaker emotion (Liscombe et al. 2003). Our findings suggest that doing so to recognize opinion categories is a promising line of work.

The rest of the paper is organized as follows: In Section 2 we discuss the data and the annotation scheme and present examples. We then present our inter-annotator agreement results in Section 3, and in Section 4 we discuss issues and observations. Related work is described in Section 5. Conclusions and Future Work are presented in Section 6.

2 Annotation

2.1 Data

The data is from the ISL meeting corpus (Burger et al. 2002). We chose task oriented meetings from the games/scenario and discussion genres, as we felt they would be closest to the applications for which the opinion analysis will be useful. The ISL speech is accompanied by rich transcriptions, which are tagged according to VERBMOBIL conventions. However, since real-time applications only have access to ASR output, we gave the annotators raw text, from which all VERBMOBIL tags, punctuation, and capitalizations were removed.

In order to see how annotations would be affected by the presence or absence of speech, we divided each raw text document into 2 segments. One part was annotated while reading the raw text only. For the annotation of the other part, speech as well as the raw text was provided.

2.2 Opinion Category Definitions

We base our annotation definitions on the scheme developed by Wiebe et al. (2005) for news articles. That scheme centers on the notion of subjectivity, the linguistic expression of private states. Private states are internal mental states that cannot be objectively observed or verified (Quirk et al. 1985) and include opinions, beliefs, judgments, evaluations, thoughts, and feelings. Amongst these many forms of subjec-

tivity, we focus on the Sentiment and Arguing categories proposed by Wilson and Wiebe (2005). The categories are broken down by polarity and defined as follows:

Positive Sentiments: positive emotions, evaluations, judgments and stances.

(4) *TBC: Well ca How about one of the the newer Cadillac the Lexus is good*

In (4), taken from the discussion of which car to buy, the speaker uses the term "good" to express his positive evaluation of the Lexus .

Negative Sentiments: negative emotions, evaluations, judgments and stances.

(5) *OCK: I think these are all really bad choices*

In (5), the speaker expresses his negative evaluation of the choices for the company car. Note that "really" makes the evaluation more intense.

Positive Arguing: arguing for something, arguing that something is true or is so, arguing that something did happen or will happen, etc.

(6) *ZDN: Yeah definitely moon roof*

In (6), the speaker is arguing that whatever car they get should have a moon roof.

Negative Arguing: arguing against something, arguing that something is not true or is not so, arguing that something did not happen or will not happen, etc.

(7) *OCK: Like a Lexus or perhaps a Stretch Lexus something like that but that might be too a little too luxurious*

In the above example, the speaker is using the term "a little too luxurious" to argue against a Lexus for the car choice.

In an initial tagging experiment, we applied the above definitions, without modification, to some sample meeting data. The definitions covered much of the arguing and sentiment we observed. However, we felt that some cases of Arguing that are more prevalent in meeting than in news data needed to be highlighted more, namely Arguing opinions that are implicit or that underlie what is explicitly said. Thus we add the following to the arguing definitions.

Positive Arguing: expressing support for or backing the acceptance of an object, viewpoint, idea or stance by providing reasoning, justifications, judgment, evaluations or beliefs. This support or backing may be explicit or implicit.

(8) *MHJ: That's That's why I wanna What about the the child safety locks I think I think that would be a good thing because if our customers happen to have children*

Example (8) is marked as both Positive Arguing and Positive Sentiment. The more explicit one is the Positive Sentiment that the locks are good. The underlying Argument is that the company car they choose should have child safety locks.

Negative Arguing: expressing lack of support for or attacking the acceptance of an object, viewpoint, idea or stance by providing reasoning, justifications, judgment, evaluations or beliefs. This may be explicit or implicit.

(9) *OCK: Town Car But it's a little a It's a little like your grandf Yeah your grandfather would drive that*

Example (9) is explicitly stating who would drive a Town Car, while implicitly arguing against choosing the Town Car (as they want younger investors).

2.3 Annotation Guidelines

Due to genre differences, we also needed to modify the annotation guidelines. For each Arguing or Sentiment the annotator perceives, he or she identifies the words or phrases used to express it (the *text span*), and then creates an annotation consisting of the following.

- Opinion Category and Polarity
- Opinion Intensity
- Annotator Certainty

Opinion Category and Polarity: These are defined in the previous sub-section. Note that the *target* of an opinion is what the opinion is about. For example, the target of "John loves baseball" is baseball. An opinion may or may not have a separate target. For example, "want stability" in "We want stability" denotes a Positive Sentiment, and there is no separate target. In contrast, "good" in "The Lexus is good" expresses a Positive Sentiment and there is a separate target, namely the Lexus.

In addition to Sentiments toward a topic of discussion, we also mark Sentiments toward other team members (e.g. "Man you guys are so limited"). We do not mark agreements or disagreements as Sentiments, as these are different dialog acts (though they sometimes co-occur with Sentiments and Arguing).

Intensity: We use a slightly modified version of Craggs and Wood's (2004) emotion intensity

annotation scheme. According to that scheme, there are 5 levels of intensity. Level "0" denotes a lack of the emotion (Sentiment or Arguing in our case), "1" denotes traces of emotion, "2" denotes a low level of emotion, "3" denotes a clear expression while "4" denotes a strong expression. Our intensity levels mean the same, but we do not mark intensity level 0 as this level implies the absence of opinion.

If a turn has multiple, separate expressions marked with the same opinion tag (category and polarity), and all expressions refer to the same target, then the annotators merge all the expressions into a larger text span, including the separating text in between the expressions. This resulting text span has the same opinion tag as its constituents, and it has an intensity that is greater than or equal to the highest intensity of the constituent expressions that were merged.

Annotator Certainty: The annotators use this tag if they are not sure that a given opinion is present, or if, given the context, there are multiple possible interpretations of the utterance and the annotator is not sure which interpretation is correct. This attribute is distinct from the Intensity attribute, because the Intensity attribute indicates the strength of the opinion, while the Annotator Certainty attribute indicates whether the annotator is sure about a given tag (whatever the intensity is).

2.4 Examples

We conclude this section with some examples of annotations from our corpus.

(10) *OCK: So Lexun had revenues of a hundred and fifty million last year and profits of like six million. That's pretty good*
Annotation: Text span=*That's pretty good* Category=Positive Sentiment Intensity=3 Annotator Certainty=Certain

The annotator marked the text span "That's pretty good" as Positive Sentiment because this this expression is used by OCK to show his favorable judgment towards the company revenues. The intensity is 3, as it is a clear expression of Sentiment.

(11) *OCK: No it might just be a piece of junk Cheap piece of junk that's not a good investment*
Annotation1: Text span=*it might just be a piece of junk Cheap piece of junk that's not a good investment* Category=Negative Sentiment Intensity=4 Annotator Certainty=Certain
Annotation2: Text span=*Cheap piece of junk that's not a good investment* Category=Negative Arguing Intensity=3 Annotator Certainty=Certain

In the above example, there are multiple expressions of opinions. In Annotation1, the expressions "it might just be a piece of junk", "cheap piece of junk" and "not a good investment" express negative evaluations towards the car choice (suggested by another participant in a previous turn). Each of these expressions is a clear case of Negative Sentiment (Intensity=3). As they are all of the same category and polarity and towards the same target, they have been merged by the annotator into one long expression of Intensity=4. In Annotation2, the sub-expression "cheap piece of junk that is not a good investment" is also used by the speaker OCK to argue against the car choice. Hence the annotator has marked this as Negative Arguing.

3 Guideline Development and Inter-Annotator Agreement

3.1 Annotator Training

Two annotators (both co-authors) underwent three rounds of tagging. After each round, discrepancies were discussed, and the guidelines were modified to reflect the resolved ambiguities. A total of 1266 utterances belonging to sections of four meetings (two of the discussion genre and two of the game genre) were used in this phase.

3.2 Agreement

The unit for which agreement was calculated was the turn. The ISL transcript provides demarcation of speaker turns along with the speaker ID. If an expression is marked in a turn, the turn is assigned the label of that expression. If there are multiple expressions marked within a turn with different category tags, the turn is assigned all those categories. This does not pose a problem for our evaluation, as we evaluate each category separately.

A previously unseen section of a meeting containing 639 utterances was selected and divided

into 2 segments. One part of 319 utterances was annotated using raw text as the only signal, and the remaining 320 utterances were annotated using text and speech. Cohen's Kappa (1960) was used to calculate inter-annotator agreement. We calculated inter-annotator agreement for both conditions: raw-text-only and raw-text+speech. This was done for each of the categories: Positive Sentiment, Positive Arguing, Negative Sentiment, and Negative Arguing. To evaluate a category, we did the following:

- For each turn, if both annotators tagged the turn with the given category, or both did not tag the turn with the category, then it is a match.
- Otherwise it is a mismatch

Table 1 shows the inter-annotator Kappa values on the test set.

Agreement (Kappa)	Raw Text only	Raw Text + Speech
Positive Arguing	0.54	0.60
Negative Arguing	0.32	0.65
Positive Sentiment	0.57	0.69
Negative Sentiment	0.41	0.61

Table 1 Inter-annotator agreement on different categories.

With raw-text-only annotation, the Kappa value is in the moderate range according to Landis and Koch (1977), except for Negative Arguing for which it is 0.32. Positive Arguing and Positive Sentiment were more reliably detected than Negative Arguing and Negative Sentiment. We believe this is because participants were more comfortable with directly expressing their positive sentiments in front of other participants. Given only the raw text data, inter-annotator reliability measures for Negative Arguing and Negative Sentiment are the lowest. We believe this might be due to the fact that participants in social interactions are not very forthright with their Negative Sentiments and Arguing. Negative Sentiments and Arguing towards something may be expressed by saying that something else is better. For example, consider the following response of one participant to another participant's suggestion of aluminum wheels for the company car

(12) *ZDN: Yeah see what kind of wheels you know they have to look dignified to go with the car*

The above example was marked as Negative Arguing by one annotator (i.e., they should not get aluminum wheels) while the other annotator did not mark it at all. The implied Negative Arguing toward getting aluminum wheels can be inferred from the statement that the wheels should look dignified. However the annotators were not sure, as the participant chose to focus on what is desirable (i.e., dignified wheels). This utterance is actually both a general statement of what is desirable, and an implication that aluminum wheels are not dignified. But this may be difficult to ascertain with the raw text signal only.

When the annotators had speech to guide their judgments, the Kappa values go up significantly for each category. All the agreement numbers for raw text+speech are in the substantial range according to Landis and Koch (1977). We observe that with speech, Kappa for Negative Arguing has *doubled* over the Kappa obtained without speech. The Kappa for Negative Sentiment (text+speech) shows a 1.5 times improvement over the one with only raw text. Both these observations indicate that speech is able to help the annotators tag negativity more reliably. It is quite likely that a seemingly neutral sentence could sound negative, depending on the way words are stressed or pauses are inserted. Comparing the agreement on Positive Sentiment, we get a 1.2 times improvement by using speech. Similarly, agreement improves by 1.1 times for Positive Arguing when speech is used. The improvement with speech for the Positive categories is not as high as compared to negative categories, which conforms to our belief that people are more forthcoming about their positive judgments, evaluations, and beliefs.

In order to test if the turns where annotators were uncertain were the places that caused mismatch, we calculated the Kappa with the annotator-uncertain cases removed. The corresponding Kappa values are shown in Table 2

Agreement (Kappa)	Raw Text only	Raw Text + Speech
Positive Arguing	0.52	0.63
Negative Arguing	0.36	0.63
Positive Sentiment	0.60	0.73
Negative Sentiment	0.50	0.61

Table-2 Inter-annotator agreement on different categories, Annotator Uncertain cases removed.

The trends observed in Table 1 are seen in Table 2 as well, namely annotation reliability improving with speech. Comparing Tables 1 and 2,

we see that for the raw text, the inter-annotator agreement goes up by 0.04 points for Negative Arguing and goes up by 0.09 points for Negative Sentiment. However, the agreement for Negative Arguing and Negative Sentiment on raw-text+ speech between Tables 1 and 2 remains almost the same. We believe this is because we had 20% fewer Annotator Uncertainty tags in the raw-text+speech annotation as compared to raw-text-only, thus indicating that some types of uncertainties seen in raw-text-only were resolved in the raw-text+speech due to the speech input. The remaining cases of Annotator Uncertainty could have been due to other factors, as discussed in the next section

Table 3 shows Kappa with the low intensity tags removed. The hypothesis was that low intensity might be borderline cases, and that removing these might increase inter-annotator reliability.

Agreement (Kappa)	Raw Text only	Raw Text + Speech
Positive Arguing	0.53	0.66
Negative Arguing	0.26	0.65
Positive Sentiment	0.65	0.74
Negative Sentiment	0.45	0.59

Table-3 Inter-annotator agreement on different categories, Intensity 1, 2 removed.

Comparing Tables 1 and 3 (the raw-text columns), we see that there is an improvement in the agreement on sentiment (both positive and negative) if the low intensity cases are removed. The agreement for Negative Sentiment (raw-text) goes up marginally by 0.04 points. Surprisingly, the agreement for Negative Arguing (raw-text) goes down by 0.06 points. Similarly in raw-text+speech results, removal of low intensity cases does not improve the agreement for Negative Arguing while hurting Negative Sentiment category (by 0.02 points). One possible explanation is that it may be equally difficult to detect Negative categories at both low and high intensities. Recall that in (12) it was difficult to detect if there is Negative Arguing at all. If the annotator decided that it is indeed a Negative Arguing, it is put at intensity level=3 (i.e., a clear case).

4 Discussion

There were a number of interesting subjectivity related phenomena in meetings that we observed during our annotation. These are issues that will need to be addressed for improving inter-annotator reliability.

Global and local context for arguing: In the context of a meeting, participants argue for (positively) or against (negatively) a topic. This may become ambiguous when the participant uses an explicit local Positive Arguing and an implicit global Negative Arguing. Consider the following speaker turn, at a point in the meeting when one participant has suggested that the company car should have a moon roof and another participant has opposed it, by saying that a moon roof would compromise the headroom.

(13) *OCK: We wanna make sure there's adequate headroom for all those six foot six investors*

In the above example, the speaker OCK, in the local context of the turn, is arguing positively that headroom is important. However, in the global context of the meeting, he is arguing against the idea of a moon roof that was suggested by a participant. Such cases occur when one object (or opinion) is endorsed which automatically precludes another, mutually exclusive object (or opinion).

Sarcasm/Humor: The meetings we analyzed had a large amount of sarcasm and humor. Issues arose with sarcasm due to our approach of marking opinions towards the content of the meeting (which forms the target of the opinion). Sarcasm is difficult to annotate because sarcasm can be

1) On topic: Here the target is the topic of discussion and hence sarcasm is used as a Negative Sentiment.

2) Off topic: Here the target is not a topic under discussion, and the aim is to purely elicit laughter.

3) Allied topic: In this case, the target is related to the topic in some way, and it's difficult to determine if the aim of the sarcasm/humor was to elicit laughter or to imply something negative towards the topic.

Multiple modalities: In addition to text and speech, gestures and visual diagrams play an important role in some types of meetings. In one meeting that we analyzed, participants were working together to figure out how to protect an egg when it is dropped from a long distance, given the materials they have. It was evident they were using some gestures to describe their ideas ("we can put tape like this") and that they drew diagrams to get points across. In the absence of visual input, annotators would need to guess

what was happening. This might further hurt the inter-annotator reliability.

5 Related Work

Our opinion categories are from the subjectivity schemes described in Wiebe et al. (2005) and Wilson and Wiebe (2005). Wiebe et al. (2005) perform expression level annotation of opinions and subjectivity in text. They define their annotations as an *experiencer* having some type of *attitude* (such as Sentiment or Arguing), of a certain intensity, towards a target. Wilson and Wiebe (2005) extend this basic annotation scheme to include different types of subjectivity, including Positive Sentiment, Negative Sentiment, Positive Arguing, and Negative Arguing.

Speech was found to improve inter-annotator agreement in discourse segmentation of monologs (Hirschberg and Nakatani 1996). Acoustic clues have been successfully employed for the reliable detection of the speaker's emotions, including frustration, annoyance, anger, happiness, sadness, and boredom (Liscombe et al. 2003). Devillers et al. (2003) performed perceptual tests with and without speech in detecting the speaker's fear, anger, satisfaction and embarrassment. Though related, our work is not concerned with the speaker's emotions, but rather opinions toward the issues and topics addressed in the meeting.

Most annotation work in multiparty conversation has focused on exchange structures and discourse functional units like common grounding (Nakatani and Traum, 1998). In common grounding research, the focus is on whether the participants of the discourse are able to understand each other, and not their opinions towards the content of the discourse. Other tagging schemes like the one proposed by Flammia and Zue (1997) focus on information seeking and question answering exchanges where one participant is purely seeking information, while the other is providing it. The SWBD DAMSL (Jurafsky et al., 1997) annotation scheme over the Switchboard telephonic conversation corpus labels shallow discourse structures. The SWBD-DAMSL had a label "sv" for opinions. However, due to poor inter-annotator agreement, the authors discarded these annotations. The ICSI MRDA annotation scheme (Rajdip et al., 2003) adopts the SWBD DAMSL scheme, but does not distinguish between the opinionated and objective statements. The ISL meeting corpus (Burger and Sloane, 2004) is annotated with dialog acts and discourse moves like initiation and response, which in turn consist of dialog tags such as query, align, and statement. Their statement dialog category would not only include Sentiment and Arguing tags discussed in this paper, but it would also include objective statements and other types of subjectivity.

"Hot spots" in meetings closely relate to our work because they find sections in the meeting where participants are involved in debates or high arousal activity (Wrede and Shriberg 2003). While that work distinguishes between high arousal and low arousal, it does not distinguish between opinion or non-opinion or the different types of opinion. However, Janin et al. (2004) suggest that there is a relationship between dialog acts and involvement, and that involved utterances contain significantly more evaluative and subjective statements as well as extremely positive or negative answers. Thus we believe it may be beneficial for such works to make these distinctions.

Another closely related work that finds participants' positions regarding issues is argument diagramming (Rienks et al. 2005). This approach, based on the IBIS system (Kunz and Rittel 1970), divides a discourse into issues, and finds lines of deliberated arguments. However they do not distinguish between subjective and objective contributions towards the meeting.

6 Conclusions and Future Work

In this paper we performed an annotation study of opinions in meetings, and investigated the effects of speech. We have shown that it is possible to reliably detect opinions within multiparty conversations. Our consistently better agreement results with text+speech input over text-only input suggest that speech is a reliable indicator of opinions. We have also found that Annotator Uncertainty decreased with speech input. Our results also show that speech is a more informative indicator for negative versus positive categories. We hypothesize that this is due to the fact the people express their positive attitudes more explicitly. The speech signal is thus even more important for discerning negative opinions. This experience has also helped us gain insights to the ambiguities that arise due to sarcasm and humor.

Our promising results open many new avenues for research. It will be interesting to see how our categories relate to other discourse structures, both at the shallow level (agreement/disagreement) as well as at the deeper level

(intentions/goals). It will also be interesting to investigate how other forms of subjectivity like speculation and intention are expressed in multiparty discourse. Finding prosodic correlates of speech as well as lexical clues that help in opinion detection would be useful in building subjectivity detection applications for multiparty meetings.

References

Susanne Burger and Zachary A Sloane. 2004. The ISL Meeting Corpus: Categorical Features of Communicative Group Interactions. *NIST Meeting Recognition Workshop 2004*, NIST 2004, Montreal, Canada, 2004-05-17

Susanne Burger, Victoria MacLaren and Hua Yu. 2002. The ISL Meeting Corpus: The Impact of Meeting Type on Speech Style. *ICSLP-2002*. Denver, CO: ISCA, 9 2002.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Meas.*, 20:37–46.

Richard Craggs and Mary McGee Wood. 2004. A categorical annotation scheme for emotion in the linguistic content of dialogue. *Affective Dialogue Systems*. 2004.

Laurence Devillers, Lori Lamel and Ioana Vasilescu. 2003. Emotion detection in task-oriented spoken dialogs. *IEEE International Conference on Multimedia and Expo (ICME)*.

Rajdip Dhillon, Sonali Bhagat, Hannah Carvey and Elizabeth Shriberg. 2003. "Meeting Recorder Project: Dialog Act Labeling Guide," *ICSI Technical Report TR-04-002*, Version 3, October 2003

Giovanni Flammia and Victor Zue. 1997. Learning The Structure of Mixed Initiative Dialogues Using A Corpus of Annotated Conversations. *Eurospeech 1997*, Rhodes, Greece 1997, p1871—1874

Julia Hirschberg and Christine Nakatani. 1996. A Prosodic Analysis of Discourse Segments in Direction-Giving Monologues Annual Meeting- *Association For Computational Linguistics 1996*, VOL 34, pages 286-293

Adam Janin, Jeremy Ang, Sonali Bhagat, Rajdip Dhillon, Jane Edwards, Javier Mac´ıas-Guarasa, Nelson Morgan, Barbara Peskin, Elizabeth Shriberg, Andreas Stolcke, Chuck Wooters and Britta Wrede. 2004. "The ICSI Meeting Project: Resources and Research," *ICASSP-2004 Meeting Recognition Workshop*. Montreal; Canada: NIST, 5 2004

Daniel Jurafsky, Elizabeth Shriberg and Debra Biasca, 1997. *Switchboard-DAMSL Labeling Project Coder's Manual*. http://stripe.colorado.edu/˜jurafsky/manual.august1

Werner Kunz and Horst W. J. Rittel. 1970. Issues as elements of information systems. *Working Paper WP-131*, Univ. Stuttgart, Inst. Fuer Grundlagen der Planung, 1970

Richard Landis and Gary Koch. 1977. The Measurement of Observer Agreement for Categorical Data *Biometrics*, Vol. 33, No. 1 (Mar., 1977) , pp. 159-174

Agnes Lisowska. 2003. Multimodal interface design for the multimodal meeting domain: Preliminary indications from a query analysis study. *Technical Report IM2. Technical report, ISSCO/TIM/ETI*. Universit de Genve, Switserland, November 2003.

Jackson Liscombe, Jennifer Venditti and Julia Hirschberg. 2003. Classifying Subject Ratings of Emotional Speech Using Acoustic Features. *Eurospeech* 2003.

Christine Nakatani and David Traum. 1998. *Draft: Discourse Structure Coding Manual version 2/27/98*

Randolph Quirk, Sidney Greenbaum, Geoffry Leech and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language. Longman*, New York.s

Rutger Rienks, Dirk Heylen and Erik van der Weijden. 2005. Argument diagramming of meeting conversations. In Vinciarelli, A. and Odobez, J., editors, *Multimodal Multiparty Meeting Processing, Workshop at the 7th International Conference on Multimodal Interfaces*, pages 85–92, Trento, Italy

Janyce Wiebe, Theresa Wilson and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation (formerly Computers and the Humanities)*, volume 39, issue 2-3, pp. 165-210.

Theresa Wilson and Janyce Wiebe. 2005. Annotating attributions and private states. *ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*.

Britta Wrede and Elizabeth Shriberg. 2003. Spotting "Hotspots" in Meetings: Human Judgments and Prosodic Cues. *Eurospeech 2003*, Geneva