Relationship between Utterances and "Enthusiasm" in Non-task-oriented Conversational Dialogue

Ryoko TOKUHISA
Toyota Central R&D Labs., INC.
Nagakute Aichi JAPAN
tokuhisa@mosk.tytlabs.co.jp

Ryuta TERASHIMA
Toyota Central R&D Labs., INC.
Nagakute Aichi JAPAN
ryuta@mosk.tytlabs.co.jp

Abstract

The goal of this paper is to show how to accomplish a more enjoyable and enthusiastic dialogue through the analysis of human-to-human conversational dialogues. We first created a conversational dialogue corpus annotated with two types of tags: one type indicates the particular aspects of the utterance itself, while the other indicates the degree of enthusiasm. We then investigated the relationship between these tags. Our results indicate that affective and cooperative utterances are significant to enthusiastic dialogue.

1 Introduction

For a non-task-oriented conversational dialogue system (e.g. home robots), we should strive for a dialogue strategy that is both enjoyable and enthusiastic, as well as efficient. Many studies have been conducted on efficient dialogue strategies (Walker et al., 1998; Litman et al., 2000; Komatani et al., 2002), but it is not clear how to accomplish a more ”human-like enthusiasm” for a conversational dialogue. The goal of this paper is to show the types of utterances that contribute to enthusiasm in conversational dialogues.

2 Corpus Annotation

We created a conversational corpus annotated with two types of tags: one type indicates particular aspects of the utterance itself, while the other indicates the degree of enthusiasm in the dialogue. This section describes our corpus and tagging scheme in detail.

2.1 Corpus Collection

As a result of previous works, several conversational dialogue corpora have been collected with various settings (Graff and Bird, 2000; TSENG, 2001). The largest conversational dialogue corpus is the Switchboard Corpus, which consists of about 2400 conversational English dialogues between two unfamiliar speakers over the telephone on one of 70 topics (e.g. pets, family life, education, gun control, etc.).

Our corpus was collected from face-to-face interaction between two unfamiliar speakers. The reasons were 1) face-to-face interaction increases the number of enthusiastic utterances, relative to limited conversational channel interaction such as over the telephone; 2) the interaction between unfamiliar speakers reduces the enthusiasm resulting from unobserved reasons during the recording; 3) the exchange in a twoparty dialogue will be simpler than that of a multiparty dialogue.

We created a corpus containing ten conversational dialogues that were spoken by an operator (thirties, female) and one of ten subjects (twenties to sixties, equal numbers of males and females). Before beginning the recording session, the subject chose three cards from fifteen cards on the following topics:

Food, Travel, Sport, Hobbies, Movies, Prizes,
TV Programs, Family, Books, School, Music,
Pets, Shopping, Recent Purchases, Celebrities

Straying from the selected topic was permitted, because these topic cards were only ever intended as a prompt to start the dialogue. Thus, we collected ten dialogues, each about 20 minutes long. For convenience, in this paper, we refer to the operator as **speaker1**, and the subject as **speaker2**.

Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue, pages 161–167,
Sydney, July 2006. ©2006 Association for Computational Linguistics

2.2 Annotation of DAs and RRs

2.2.1 Definition of tagging scheme

Dialogue Acts (DAs) and Rhetorical Relations (RRs) are well-known tagging schemes for annotating an utterance or a sentence. DAs are tags that pertain to the function of an utterance itself, while RRs indicate the relationship between sentences or utterances. We adopted both tags to allow us to analyze the aspects of utterances in various ways, but adapted them slightly for our particular needs.

The DA annotations were based on SWBD-DAMSL and MRDA (Jurafsky et al., 1997; Dhillon et al., 2004). The SWBD-DAMSL is the DA tagset for labeling a conversational dialogue. The Switchboard Corpus mentioned above was annotated with SWBD-DAMSL. On the other hand, MRDA is the DA tagset for labeling the dialogue of a meeting between multiple participants. Table 1 shows the correspondence between SWBD-DAMSL/MRDA and our DAs[1]. We describe some of the major adaptations below.

The tags pertaining to questions: In SWBD-DAMSL and MRDA, the tags pertaining to questions were classified by the type of their form (e.g. *Wh-question*). We re-categorized them into request and confirm in terms of the ”act” for Japanese.

The tags pertaining to responses: We subdivided *Accept* and *Reject* into objective responses (*accept,denial*) and subjective responses (*agree, disagree*).

The emotional tags: We added tags that indicate the expression of *admiration* and *interest*.

The overlap tags with the RRs definition: We did not use any tags (e.g. *Summary*), that overlapped the RR definition.

Consequently, we defined 47 DAs for analyzing a conversational dialogue.

Table 1: Dialogue Act Definition

SWBD-DAMSL/MRDA	Our DAs	Definition
Statement non opinion	inform objective fact	inform non opinion
Statement opinion	inform subjective element	inform opinion
Wh-Question **Yes-No-question** **Open-Question** **Or-Question**	request objective fact	request non opinion
	request agreement	request agreement opinion
	confirm objective fact	confirm non opinion
	confirm agreement	confirm agreement opinion
Accept	accept	accept non opinion
	agree	accept opinion
Reject	denial	denial non opinion
	disagree	denial opinion
not marked	express admiration	inform admiration
Summary	DEL. (mark as RR)	—————

The RR annotations were based on the rhetorical relation defined in Rhetorical Structure Theory (RST) (Mann and Thompson, 1988; Stent and Allen, 2000). Our RR definition was based only on informational level relation defined in RST because we annotated the intentional level with DAs. Table 2 shows the correspondence between the informational relation of RST and our RRs. We describe some of the major adaptations below.

Subdivide evaluation: The evaluation reflects the degree of enthusiasm in the dialogue, so we divided the *Evaluation* into three types of *evaluation (positive/negative/neutral).*

Table 2: Rhetorical Relation Definition

Mann’s RST	Our RRs	definition
Evaluation	evaluation (positive)	U2 is a positive evaluation about U1
	evaluation (negative)	U2 is a negative evaluation about U1
	evaluation (neutral)	U2 is neutral evaluation about U1
Volitional cause	volitional cause-effect	U2 is a volitional action, and U1 cause U2
Volitional result		
No Definition	addition	U2 consists of a part of U1

Integrate the causal relations: We use a directed graph representation for RR annotations, so that we integrate *Non-volitional cause* and *Non-volitional result* into *non-volitional cause-effect*, and *Volitional cause* and *Volitional result* into *volitional cause-effect*.

Add addition relation: The RRs initially represent the structure of the written text, segmented into clause-like units. Therefore, they do not cover those cases in which one clause is uttered by one speaker, but communicatively completed by another. So, we added an *addition* to our RRs. The following is an example of *addition*.

speaker A: the lunch in our company cafeteria

speaker B: is good value for money

We defined 16 RRs as a result of these adaptations.

[1] The tags listed in *italics* are based on SWBD-DAMSL while those in **boldface** are based on MRDA.

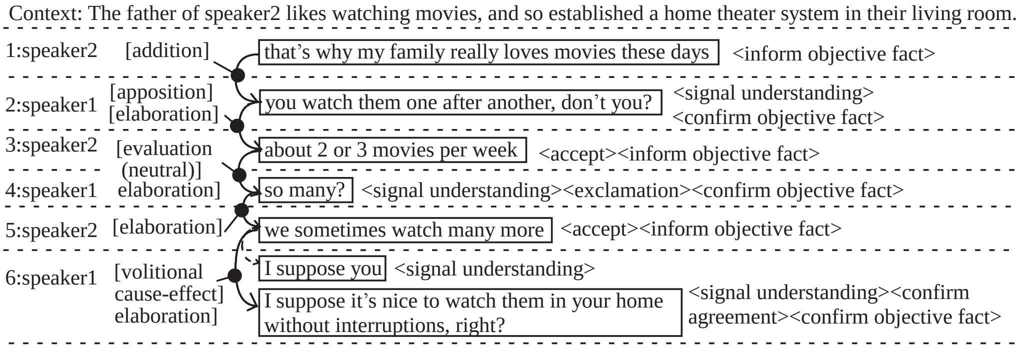
Context: The father of speaker2 likes watching movies, and so established a home theater system in their living room.
1:speaker2
[addition]
that's why my family really loves movies these days
<inform objective fact>
2:speaker1
[apposition]
[elaboration]
you watch them one after another, don't you?
<signal understanding>
<confirm objective fact>
3:speaker2
[evaluation (neutral)]
about 2 or 3 movies per week
<accept><inform objective fact>
4:speaker1
elaboration]
so many?
<signal understanding><exclamation><confirm objective fact>
5:speaker2
[elaboration]
we sometimes watch many more
<accept><inform objective fact>
6:speaker1
I suppose you
<signal understanding>
[volitional cause-effect]
elaboration]
I suppose it's nice to watch them in your home without interruptions, right?
<signal understanding><confirm agreement><confirm objective fact>

Figure 1: Example of Dialogue annotated with DAs and RRs (Originally in Japanese)

2.2.2 Annotation of DAs and RRs

DAs and RRs are annotated using the MMAX2 Annotation Tool [2] (Muller and Strube, 2003). Figure 1 shows an example of our corpus annotated with DAs and RRs. The ⟨ ⟩ symbol in Figure 1 indicates a DA, while the [ ] symbol indicates an RR. Below, we describe our annotation process for DAs and RRs.

Step 1. Utterance Segmentation: All the utterances in the dialogue are segmented into DA segments, each of which we define as an *utterance*. In Figure 1, the utterance is surrounded with a square. In this step, we also eliminated backchannels from the exchange.

Step 2. Annotation of DAs: DAs are annotated to all utterances. In those cases in which one DA alone cannot represent an utterance, two or more DAs are used (see Figure 1 line 2).

Step 3. Annotation of Adjacency Pairs: Adjacency pairs (APs) are labeled. An AP consists of two utterances where each part is produced by a different speaker. In Figure 1, the solid and dotted lines correspond to links between the APs.

Step 4. Annotation of RRs: RRs on APs are labeled. A solid line indicates an AP that is labeled with RRs, while a dotted line indicates an AP that is not labeled with RRs. If a single RR cannot represent the type of the relationship, two or more RRs are used.

2.3 Annotation of *Enthusiasm*

2.3.1 Related Work on Annotating the degree of enthusiasm

Wrede et al. annotated *Involvement* to the ICSI Meeting Recorder Corpus (Wrede and Shriberg, 2003b; Wrede and Shriberg, 2003a). In their method, a rater judges *involvement (agreement, disagreement, other)* or *Not especially involved* or *Don't Know*, by listening to each utterance without the context of the dialogue. In the experiment, nine raters provided ratings on 45 utterances. Inter-rater agreement between *Involved* and *Not especially involved* yielded a Kappa of κ=.59 (p<.01), but 13 of the 45 utterances (28.9%) were rated as *Don't Know* by at least one of the raters. For automatic detection, it is certainly effective to rate *Involvement* without context. However, the results indicate that it is quite difficult to recognize *Involvement* from a single utterance. Moreover, the fluctuation of *Involvement* can not be recognized by this method because *Involvement* is categorized into five categories only.

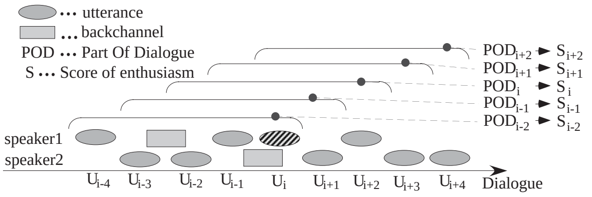

Figure 2: Rating the score of the enthusiasm

2.3.2 Our Method of Annotating *Enthusiasm*

In this section, we propose a method for evaluating the degree of enthusiasm. We describe the process for evaluating the degree of enthusiasm.

Step 1. Rating the score of enthusiasm for POD

A rater estimates a score of the enthusiasm corresponding to the part of dialogue (POD), which is a series of five utterances. As mentioned above, the backchannels are not regarded as utterances. In Figure 2, S_i denotes

[2] This supports multilevel annotation and the creation of a relationship between utterances. http://www.eml-research.de/english/research/nlp/down-load/mmax.php

the score for the enthusiasm of POD_i. The value of the score can be from 10 to 90.

90 ...	Extreme
70 ...	Moderate
50 ...	Neutral
30 ...	Low
10 ...	No

When rating the score, a rater must obey the following four rules.

1. Listen to each POD more than three times.
2. Perform estimation based on the entire POD and not just part of the POD.
3. Be sure that own ratings represented a consistent continuum.
4. Estimate as participants, not as side-participants.

We did not give any definitions or examples to rate the enthusiasm, a rater estimated a score based on their subjective determination.

Step 2. Calculate the score of enthusiasm for an utterance

The score of enthusiasm for an utterance U_i is given by the average of the scores of the PODs that contain utterance U_i.

$$V(U_i) = \frac{1}{5}\sum_{j=i-2}^{i+2} S_j \qquad (1)$$

Step 3. Calculate the degree of enthusiasm for an utterance and an adjacency pair

In this paper, we deal with all the degrees of enthusiasm as a normalized score, which we call *Enthusiasm*, because different raters may have different absolute levels of enthusiasm. Then, *Enthusiasm* for U_i is given as follows:

$$E(U_i) = \frac{V(U_i) - \overline{V(U)}}{\sigma} \qquad (2)$$

where

$$\overline{V(U)} = \frac{1}{n}\sum_{i=1}^{n} V(U_i)$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\{V(U_i) - \overline{V(U)}\}^2}$$

n denotes the number of utterances in the dialogue.

In addition, *Enthusiasm* for AP_i is given by the average of *Enthusiasm*s of the utterances where are AP_i.

$$E(AP_i) = \frac{1}{2}\{E(U_j) + E(U_k)\} \qquad (3)$$

U_j and U_k denote the utterances in AP_i.

3 Estimation of Annotated Corpus

3.1 Reliability of DAs and RRs

We examined the inter-annotator reliability for two annotators[3] for DAs, RRs and APs, using four dialogues mentioned above. Before the start of the investigation, one annotator segmented a dialogue into utterances. The number of segmented utterances was 697. The annotaters annotated them as described in steps 2 to 4 of Section 2.2.2.

DAs annotation: We can not apply the Kappa statistics since it cannot be applied to multiple tag annotations. We then apply formula 4 to examine the reliability.

$$ag. = \frac{(Agreed\,DAs) \times 2}{Total\,of\,DAs\,annotated\,by\,A1\,and\,A2} \times 100 \quad (4)$$

The result of agreement was 1542 DAs (65.5%) from a total of 2355 DAs. The major reasons for the disagreement were as follows.

- Disagreement of subjective/objective ... 124(15.3%)
- Disagreement of request/confirm ... 112(13.8%)
- Disagreement of partial/whole ... 72(8.9%)

Building APs: We examined the agreement of building APs between utterances. The result of agreement was 536 APs (85.2%) from the total of the 629 APs that were built by the annotators. This result shows that the building of APs is reliable.

RRs annotation: We also examined the agreement of RRs annotation. We applied formula 5 to this examination.

$$ag. = \frac{(Agreed\,RRs) \times 2}{Total\,of\,RRs\,annotated\,by\,A1\,and\,A2} \times 100 \quad (5)$$

As a result, we found agreement for 576 RRs (59.6%) out of a total of 967 RRs.

[3] We refer to these annotators as A1 and A2. A1 is one of the authors of this paper.

Table 3: Correlation between random rating and sequential rating

	correlation coefficient	
	speaker1	speaker2
twenties,female	0.833	0.881
twenties,male	0.971	0.950
sixties,female	0.972	0.973
sixties,male	0.971	0.958

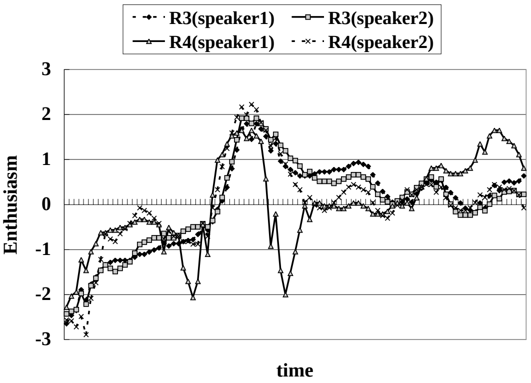

Figure 3: *Enthusiasm* of dialogue of speaker1 and speaker2(thirties,female)

3.2 Estimation Context Influence on the rating of *Enthusiasm*

In order to examine the influence of the context on the rating of *Enthusiasm*, one rater noted *Enthusiasm* under two conditions: 1) Listening to PODs randomly, and 2) Listening to PODs sequentially as dialogue. Table 3 shows the correlation between the random rating and the sequential rating. The correlation coefficient was calculated for the *Enthusiasm* of each of the two participants. The ”speaker1” shows the correlation of the *Enthusiasm* rated as speaker1, and ”speaker2” shows the correlation of the *Enthusiasm* rated as speaker2. This was found to be approximately 0.9 in both cases. These results show that *Enthusiasm* can be estimated stably and that the context has little influence.

4 Relationship between DAs/RRs and *Enthusiasm*

We investigated the relationship between DAs/RRs and *Enthusiasm*, using four dialogues. The DAs/RRs corpus annotated by A1 was used in this analysis because A1 is one of the authors of this paper and has a better knowledge of the DAs and RRs tagging scheme than A2. The *Enthusiasm* corpus annotated by R3 was used because we found that R4 rated *Enthusiasm* based on non-subjective reasons: after the examination of the rating, R4 said that speaker1 spoke enthusiastically but that it seemed unnatural because speaker1 had to manage the recording of the dialogue, which appears in the results as speaker1’s *Enthusiasm* as annotated by R4 as a notable difference (see Figure 3).

Figure 4 and 5 show the ratio of the frequency of DAs and RRs in each of the levels of *Enthusiasm* over a range of 0.5. If DAs and RRs were evenly annotated for any level of *Enthusiasm*, the graph will be completely even. However, the graph shows the right side as being higher if the DAs and RRs increase as *Enthusiasm* increases. Conversely, the graph shows the left side as being higher if the DAs and RRs fall as *Enthusiasm* increases. The number in Figure 4 and 5 indicates the average *Enthusiasm* for each DA and RR. If the average is positive, it means that the frequency of the DAs and RRs is high in that part in which *Enthusiasm* is positive. In contrast, if the average is negative, it means that the frequency of the DAs and RRs is high in that part in which *Enthusiasm* is negative.

We determined the following two points about the tendency of the DAs frequency.

Tendency of subjective and objective DAs: The ratio of the frequency of those DAs related to *subjective elements* tends to increase as *Enthusiasm* increases (see *1 in Figure 4). In contrast, the ratio of the frequency of those DAs pertaining to *objective matters* tends to decrease (see *2 in Figure 4) or equilibrate as *Enthusiasm* increases (see *3 in Figure 4) . We can thus conclude that those exchanges related to subjective elements increases in the enthusiastic dialogue, but those related to objective elements decrease or equilibrate.

Tendency of affective DAs: The ratio of the frequency of those DAs related to the *affective contents* tends to increase as *Enthusiasm* increases (see *4 in Figure 4). However, *express admiration*, which is also related to affective contents, tends to decrease (see *5 in Figure 4). We then analyzed several instances of *admiration*. As a result, we found that the prosodic characteristic of *admiration* utterance will cause this tendency.

Furthermore, we noted the following two points about the tendency of the RRs frequency.

Tendency of additional utterances: The ratio of the frequency of *addition*, which completes the

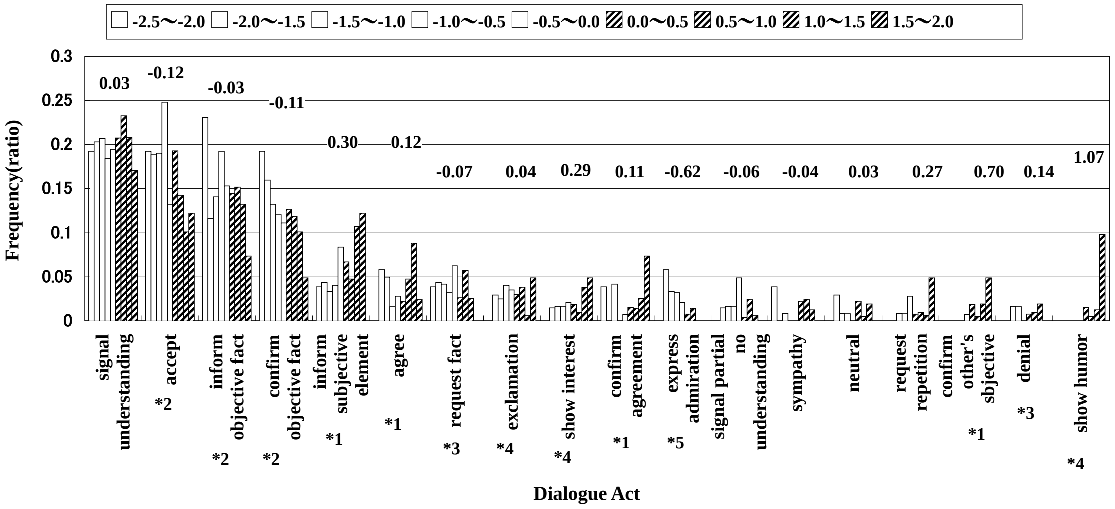

Figure 4: Frequency of DAs per *Enthusiasm*

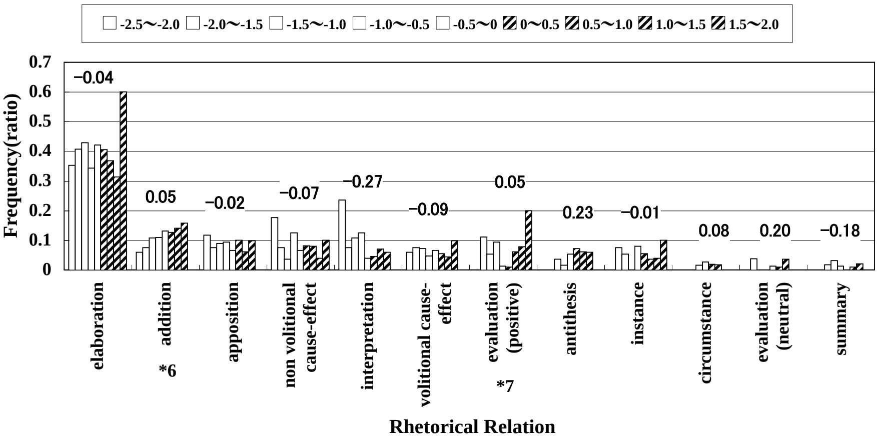

Figure 5: Frequency of RRs per *Enthusiasm*

Context:Mother of speaker2 does not cook dinner when the father is out.
1 speaker1: but if he's there then she
2 speaker2: cooks a really delicious dinner
3 speaker1: wow

Figure 6: Example of *addition*

other participant's utterance, tends to increase as *Enthusiasm* increases (see *6 in Figure 5). Figure 6 shows a dialogue example. There are *addition* relations between lines 1 and 2. This shows that the participant makes an utterance cooperatively by completing the other's utterances in enthusiastic dialogues. Such cooperative utterance is a significant component of enthusiastic dialogues.

Tendency of positive evaluation: The ratio of the frequency of *positive evaluation* tends to increase at lower *Enthusiasm* and higher *Enthusiasm* (see *7 in Figure 5). We analyzed some instances of *positive evaluation*, we then found that the speaker tries to arouse the dialogue by an utterance of *positive evaluation* at lower *Enthusiasm*, and the speaker summarizes the previous discourse with a *positive evaluation* at higher *Enthusiasm*. Figure 7 shows an example of *positive evaluation* in the enthusiastic dialogue. In this case, speaker1 ex-

Context:About a hamster and its exercise instrument.
1 speaker2: two hamsters run together in their exercise wheel
2 speaker2: they run up and down and side by side
3 speaker1: but surely they can't they run together if they aren't getting along very well?
4 speaker2: exactly
5 speaker2: one gets carried along if it stops when the other continues to run
6 speaker1: is it? does it lean forward?
7 speaker2: yes
8 speaker2: sometimes it falls out
9 speaker1: that's so cute

Figure 7: Example of *positive evaluation*

presses *positive evaluation* on line 9 about the element on line 8. The utterance on line 9 also has the function of expressing an overall *positive evaluation* of the previous discourse.

5 Conclusion and Future Research

We analyzed the relationship between utterances and the degree of enthusiasm in human-to-human conversational dialogue. We first created a conversational dialogue corpus annotated with two types of tags: DAs/RRs and *Enthusiasm*. The DA and RR tagging scheme was adapted from the definition given in a previous work, and an *Enthusiasm* tagging scheme is proposed. Our method of rating *Enthusiasm* enables the observation of the fluctuation of *Enthusiasm*, which enables the detailed analysis of the relationship between utterances and *Enthusiasm*. The result of the analysis shows the frequency of objective and subjective utterances related to the level of *Enthusiasm*. We also found that affective and cooperative utterances are significant in an enthusiastic dialogue.

In this paper, we only analyzed the relationship between DAs/RRs and *Enthusiasm*, but we expect the non-linguistic-feature related with *Enthusiasm* so that we would analyze the relationship in future research. And, we try to achieve more reliable annotation by reviewing our tagging scheme. Furthermore, we would apply the results of the analysis to our conversational dialogue system.

References

Rajdip Dhillon, Sonali Bhagat, Hannah Carvey, and Elizabeth Shriberg. 2004. Meeting Recorder Project: Dialog Act Labeling Guide. *ICSI Technical Report*, (TR-04-002).

David Graff and Steven Bird. 2000. Many Uses, Many Annotations for Large Speech Corpora: Switchboard and TDT as Case Studies. *LREC2000*.

Dan Jurafsky, Liz Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual. *www.dcs.shef.ac.uk/nlp/amities/files/bib/ics-tr-97-02.pdf*.

Kazunori Komatani, Tatsuya Kawahara, Ryosuke Ito, and Hiroshi Okuno. 2002. Efficient Dialogue Strategy to Find Users' Intended Items from Information Query Results. *In Proceedings of the COLING*.

Diane Litman, Satinder Singh, Michael Kearns, and Marilyn Walker. 2000. NJFun: A Reinforcement Learning Spoken Dialogue System. *In Proceedings of the ANLP/NAACL*.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text*, 8(3):243–281.

Christoph Muller and Michael Strube. 2003. Multi-Level Annotation in MMAX. *In Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*.

Amanda Stent and James Allen. 2000. Annotating Argumentation Acts in Spoken Dialog. *Technical Report 740*.

Shu-Chuan TSENG. 2001. Toward a Large Spontaneous Mandarin Dialogue Corpus. *In Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*.

Marilyn A. Walker, Jeanne C. Fromer, and Shrikanth Narayanan. 1998. Learning Optimal Dialogue Strategies: A Case Study of a Spoken Dialogue Agent for Email. *In Proceedings of COLING/ACL*.

Britta Wrede and Elizabeth Shriberg. 2003a. Spotting ”Hot Spots” in Meetings: Human Judgements and Prosodic Cues. *Eurospeech-03*, pages 2805–2808.

Britta Wrede and Elizabeth Shriberg. 2003b. The Relationship between Dialogue Acts and Hot Spots in Meetings. *IEEE ASRU Workshop*.