Two types of disagreement in group discussions of Japanese undergraduates

Etsuo Mizukami[a], Ikuyo Morimoto[b], Kana Suzuki[a], Hiroko Otsuka[c] ,
Hideki Kashioka and Satoshi Nakamura[a]

[a] Department of Spoken Language Research, Spoken Language Communication Research Laboratories, Advanced Telecommunications Research Institute International

[b] School of Law and Politics, Kwansei Gakuin University

[c] The Institute of Behavioral Sciences

2-2-2 Hikaridai Seikacho Sorakugun Kyoto 619-0288 Japan; email etsuo.mizukami@atr.jp

Abstract

In this study we investigated the nature of disagreement, which is a necessary component of a good discussion. We obtained 27 group discussion scenes by Japanese undergraduates. They were evaluated by two ways of impression rating and comparing them. As a result of factor analysis for the impression rating data, five factors were extracted: *activeness*, *multidirection and unification of discussion*, *relationships of participants*, *development and sophistication of discussion* and *sincerity of the participants*, and each factor score was simultaneously calculated. Each scene's rank score was also calculated by relative comparison. There was a significant positive correlation between the mean factor and the rank scores except Factor 3, *relationships of participants*. We examined four scenes of the different patterns of the factor scores and rank scores, and concluded that this difference depended on the ways of disagreement. The frequency of disagreement made Factor 3's score negative, but the *criticism* type of disagreement increased the rank scores, while its *censure* type produced lower results. The proper ways of disagreement in group discussions were discussed through the qualitative analysis.

1. Introduction

In recent years, it has been increasing that opportunities of citizens' participation in social decision-making. A major way of participation is group discussion, which has been used in various social scenes, not only in social decision-making, to find, share, solve problems, or extract the opinions and needs of people. For instance, in science communication to promote public understandings, such interactive styles of communication as workshop, science cafés, and consensus conferences have replaced the unidirectional enlightenment style [Kobayashi 2007]. In Japan, a mixed juror system will also be introduced in 2009, so venues for dialogues or discussions between experts (scientists or jurors) and non-experts (public) will progressively increase. In such scenes, the discussion outcome is regarded as important, while its process is not. However, what process that reaches what outcome is crucial. Even if a particular outcome is yielded unanimously, it's not always the thing produced through enough comprehensive discussions. Then, if what process does a discussion go through, can we say it is beneficial discussion? If one tries to assess some discussions comprehensively, to what points in discussions should he pay attention to evaluate? These problems had not been examined enough so far, although some *guidelines* or *rules* for discussions have been proposed (e.g., Hall [1971]; Kitagawa & IYO [2005]). Against such problems, we are currently developing a method to evaluate the processes of discussions ('LSSL project' supported by JST/Ristex; http://lssl.jp/).

In this paper, we report our approach and some speculations to a question caused in the course of our analysis. The question is the problem of ways of counterargument or *disagreement*. Reaching a critically-examined conclusion of a discussion confrontation between participants is inevitable. Hall [1971] argued that "differences of opinion are natural and expected. Disagreements help team decision-making because they provide a wide range of information and opinions." On the

other hand, disagreement can escalate into disintegration and disruption. What kind of disagreement is productive or necessary in group discussions? We analyzed the disagreements found in the discussion data of Japanese undergraduates. Scenes taken were rated by two evaluation methods, and then we closely examined two scenes, which were rated as identical by one method, but different by the other method. We distinguished two different types of disagreement by considering possible reasons for this discrepancy.

2. Data

Since most Japanese students do not have an opportunity to receive discussion training, they have difficulty engaging in satisfactory discussions without assistance from a *facilitator* or a *moderator*. How do they discuss? We recorded the discussion data in 2007 for the following purposes: (a) to obtain the data of typical discussions by Japanese undergraduates whose majors are different; (b) to investigate the effect of experiencing the assistance of a professional moderator; (c) to consider the definition of a good discussion by analyzing the data; and (d) to obtain examples of good and bad discussions. We obtained 27 discussion sessions from nine groups; each group had three sessions, and consisted of six people, three males and three females. Half were information science and technology majors, and the rest were humanity majors. They discussed three problems of information technology. Three groups discussed three times without the professional moderator (control condition). The other three groups discussed twice after first discussing with the moderator (first-aided condition). The remaining three groups discussed by themselves at first, then with the moderator, and finally on their own again (second-aided condition). We instructed them to discuss the following three themes concerning information technology:

(1st) Should *Yutube*, a streaming website, be regulated more strictly? If so, how?
(2nd) Do we need security cameras? If so, under what conditions should they be installed?
(3rd) Should students be allowed to use *Wikipedia*, a free encyclopedia on the web, for reports? If so, under what restrictions?

Figure 1: Snapshot of a video clip

The group members sat around a round table, and were recorded by a DVCAM recorder (SONY DSR45A) with four CCD cameras (Watec WAT 204-CX) by view-separator (3D TQS-C9) (Figure 1). Their voices were also recorded digital audio recorder (YAMAHA AW2400) with headset microphones (AKG C420). The time limits for sessions were 40 min, but they were instructed to use all the allotted time. Before the first session, they were also instructed to adhere to the following the guidelines[1] as closely as possible:

a) Make your statement positive even if your opinion is different from others.
b) Sincerely listen to other statements even they are different from yours.
c) Understand that differences of opinion are natural and expected.
d) Do not change your mind to avoid conflict and to reach agreement and harmony.
e) All members must contribute opinions during discussion to ensure diversity.
f) Describe your opinion as comprehensively as possible.
g) Be skeptical of your group's decision if it reaches an agreement quickly.

They introduced themselves at the beginning of the first session and had a break of an hour or more before the next session.

3. Methods: Evaluation of Discussion Scenes

In a previous work, we illustrated the relationships between the impressions people have toward certain scenes of focused group interviews [Vaughn et al. 1996] and the interaction of the interviewers by impression ratings, factor analysis, and interaction analysis [Morimoto et al. 2006; Suzuki et al. 2007]. We obtained four evaluation aspects for the group interviews as four factors: *conversational activeness*, *conversational sequencing*, *attitudes of participants,* and *relationships of participants*. We revealed that *conversational activeness* was related to the degrees of the reactions of observers (listener of the current speaker), and *conversational sequencing* was influenced by how the thread of discussion was built and developed. In the present study, to ascertain the content of a good discussion, we employed the same methods. It is easier to evaluate a discussion scene using some impression terms, for instance, whether it is comparatively bright or dark, active or passive, and so on. We also added that people had difficulty defining a good discussion, but it was easier to decide which of two discussion scenes was better.

Impression Rating and Factor Analysis

First, we collected candidate evaluation terms for the impression ratings of the group discussions because the 23 terms used in our previous work [Morimoto, et al. 2006; Suzuki, et al. 2007] were collected for evaluating the data of group interviews, not for group discussions. We collected 40 pairs of antonyms as evaluation terms from our previous work and the terms obtained from the contents of interviews with

[1] We determined the guidelines by modifying Hall's guideline for effective decision-making [Jay Hall 1971].

professional moderators, facilitators, and mediators. Using these terms, we demonstrated impression ratings toward four scenes of the data of 23 undergraduates who did not attend the recorded sessions. From the results of factor analysis of the impression rating data, we narrowed the terms to 31 pairs of antonyms, all of which had factor loadings of 0.4 and above after removing some kindred meaning terms that had high factor loadings in each factor. Finally the following evaluation terms were employed:
bright—dark, quiet—busy, reserved—friendly, active—passive, participating—observing, static—dynamic, natural—unnatural, open—closed, disturbed—steady, calm—desperate, blinkered—wide-scoped, flippant—serious, careful—sloppy, biased—neutral, wordy—concise, multilateral—unilateral, antipathetic—sympathetic, uniform—diverse, selfish—shared, consistent—inconsistent, linear—winding, self-centered—collaborative, unequal—equal, single—serial, developed—digressive, investigated—shallow, detailed—rough, orderly—disconnected, deep—superficial, sincere—insincere, compromising—persisting.

As the stimulus for the impression rating, we made eight-minutes video clips of the scenes taken from the discussion data. All participants who attended the recordings individually engaged in the impression ratings about two months later. Each person evaluated four video clips played on a Windows Media Player on a laptop PC with headphones, and scored each pair of terms on a 7-point scale. First, they evaluated a video clip of a discussion by the authors with the same moderator as a control stimulus. Then they evaluated three clips of other discussions randomly chosen. On balance, each of the 27 video clips was evaluated by six different persons. After the impression ratings, the participants answered questionnaires concerning the definition of a good discussion, and ranked the three clips, and justified their scores.

Table 1 shows the rotated pattern matrix (using a maximum likelihood solution, promax rotation) obtained from the factor analysis results for the impression rating data by using SPSS. Seven factors were extracted, but two were omitted because they only had one term of high factor loading. The evaluation terms shown in the left column of Table 1 are arranged so that the signs of all large factor loadings are expressed in positive values, and the antonyms are abbreviated. The bold values in Table 1 are factor loadings of 0.34 and above, i.e., large factor loadings. We interpreted five factors as *activeness*, *multidirection and unification of discussion*, *relationships of participants*, *development and sophistication of discussion,* and *sincerity of participants*.

Table 1: Pattern matrix (rotated)

Evaluation terms	Factor 1 Activeness	Factor 2 Multidirection & unification	Factor 3 Relationships	Factor 4 Development & Sophistication	Factor 5 Sincerity
bright-dark	**1.049**	-0.068	0.075	-0.177	-0.088
busy-quiet	**1.017**	-0.080	-0.051	-0.185	0.002
friendly-reserved	**0.904**	-0.032	-0.079	0.038	-0.012
active-passive	**0.816**	0.068	-0.022	0.039	-0.030
participating-observing	**0.632**	0.066	0.088	0.080	-0.049
dynamic-static	**0.621**	0.153	-0.139	-0.036	0.013
natural-unnatural	**0.596**	0.003	-0.094	0.248	0.186
open-closed	**0.538**	-0.059	-0.043	0.190	0.233
steady-disturbed	**0.457**	0.078	0.136	0.329	-0.092
calm-desperate	**0.425**	-0.083	0.093	0.127	0.301
wide-scoped-blinkered	-0.028	**0.723**	-0.090	0.090	-0.027
serious-flippant	-0.088	**0.671**	-0.031	-0.177	**0.391**
careful-sloppy	-0.110	**0.609**	0.118	0.174	-0.044
neutral-biased	-0.029	**0.587**	0.202	-0.232	0.032
concise-wordy	0.221	**0.582**	0.196	0.080	-0.198
multilateral-unilateral	0.140	**0.562**	-0.335	0.162	-0.097
sympathetic-antipathetic	0.094	0.029	**0.627**	-0.065	0.087
uniform-diverse	-0.060	-0.122	**0.575**	0.043	-0.070
shared-selfish	0.175	0.020	**0.525**	0.055	-0.010
consistent-inconsistent	0.036	0.095	**0.479**	0.031	0.130
linear-winding	-0.162	-0.010	**0.461**	0.084	0.003
collaborative-self-centered	0.042	0.046	**0.403**	0.065	0.235
equal-unequal	0.201	0.024	0.215	0.018	0.053
serial-single	0.284	-0.128	0.103	**0.604**	-0.150
developed-digressive	-0.026	0.113	0.209	**0.573**	-0.094
investigated-shallow	-0.048	0.271	-0.097	**0.534**	0.061
detailed-rough	-0.035	**0.344**	0.009	**0.514**	0.027
orderly-disconnected	-0.080	0.087	**0.401**	**0.489**	-0.036
deep-superficial	0.142	0.199	-0.136	**0.349**	0.180
sincere-insincere	0.083	0.055	0.056	-0.085	**0.686**
persisting-compromising	0.131	-0.021	0.061	0.089	0.108
Variance explained	5.777	2.704	2.194	2.094	1.044
Propotion	18.6%	8.7%	7.1%	6.8%	3.4%

Table 2: Mean factor and rank scores

session #	Factor 1 Activeness	Factor 2 Multidirection & Unification	Factor 3 Relationships	Factor 4 Development & Sophistication	Factor 5 Sincerity	Impression score	Rank score
0-1-2	0.795	0.568	0.552	0.307	0.437	0.272	120
1-3-3	0.055	-0.657	-1.474	-0.570	-0.592	-0.210	105
2-1-2	-0.016	0.518	0.332	0.689	0.192	0.119	90
0-1-1	0.323	0.213	0.169	0.011	0.168	0.097	80
1-3-1	0.209	1.049	-0.185	0.931	0.141	0.185	75
1-3-2	0.806	-0.226	-1.659	-0.277	0.173	0.000	75
2-1-1	-0.468	0.400	-0.769	-0.226	-0.503	-0.139	75
2-1-3	0.385	-0.138	-0.017	0.021	-0.099	0.057	60
0-1-3	0.279	0.643	0.676	0.188	0.390	0.182	55
1-2-3	0.555	-0.530	0.535	-0.135	0.168	0.092	55
0-3-1	0.502	-0.029	0.053	0.089	0.457	0.116	50
2-3-3	-0.767	-0.258	-0.238	-0.216	-0.055	-0.199	45
0-3-2	0.054	-0.235	0.084	0.070	-0.223	-0.007	40
2-2-2	-1.159	-0.572	0.047	-0.790	-0.845	-0.345	35
2-3-1	-1.138	-0.031	0.435	-0.568	-0.252	-0.231	35
0-3-3	0.310	-0.954	-0.151	-0.203	-0.254	-0.058	30
1-1-1	-1.196	-0.533	0.225	-0.544	-0.574	-0.309	25
2-3-2	-1.107	-0.880	0.480	-0.655	-0.481	-0.310	25
1-2-2	-0.273	-0.302	0.187	-0.582	-0.298	-0.113	20
0-2-2	-0.602	-0.447	-0.167	-0.970	-0.421	-0.243	15
1-1-2	-0.893	-0.702	-0.308	-0.733	-0.568	-0.318	10
0-2-3	-0.556	-1.309	-1.207	-1.365	-0.296	-0.405	5
2-2-3	-0.292	-0.996	0.391	-0.528	-0.267	-0.158	5
1-2-1	-1.258	0.230	-0.276	-0.787	-1.024	-0.322	0
1-1-3	-0.993	0.067	0.048	-0.921	-0.139	-0.243	-10
2-2-1	-1.205	-1.276	-0.028	-1.059	-0.587	-0.429	-10
0-2-1	-1.401	-0.917	-0.471	-0.926	-0.902	-0.467	-30

We defined Impression Score I(i) from these five factor scores as,

$$I(i)=\sum_{k=1}^{5} I_k(i)\times P_k$$

where i is the scene number (shown in the left column in Table 2), $I_k(i)$ the mean factor scores[2] of Factor k and P_k the factor proportion of each factor (shown in the bottom row in Table 1). Impression Scores, therefore, approximately indicate the total scores evaluated by the impression ratings. We show them in the second column from the right in Table 2.

Ranking of Clips

Table 2 shows the mean factor scores of each scene and their rank score R(i) arranged in the order of the rank score. Rank score R(i) is defined as

$$R(i) = N_1(i)*20+N_2(i)*5-N_3(i)*5 \quad (N_1(i)+ N_2(i)+ N_3(i)=6),$$

where i is the scene number, $N_1(i)$ the number of participants who judged scene #i as the best discussion of the three scenes they evaluated, $N_2(i)$ second, and $N_3(i)$ the worst. Therefore, if all (six) evaluators judge scene #i as the best discussion, then R(i) equals 120, e.g., scene #0-1-2 gets the highest score (top of Table 2).

Correlation Analysis

If both the results of impression ratings and rankings are sufficiently reliable, i.e., if both can be regarded as evaluation indexes for the quality of discussion, they might correlate to each other. Figure 2 shows the correlation between I(i) and R(i). It indicates that they have a strong positive correlation: rank correlation Kendall's τ =0.546, Z_0 = 4.519, p<0.001.

[2] The rating scores from 1 to 7 were converted into scores from -3 to 3 since the terms shown on the left column of Table 1 were positive and their antonyms were negative, e.g., very bright=3, bright=2, slightly bright=1, intermediate=0, slightly dark=-1, dark=-2 and very dark=-3.

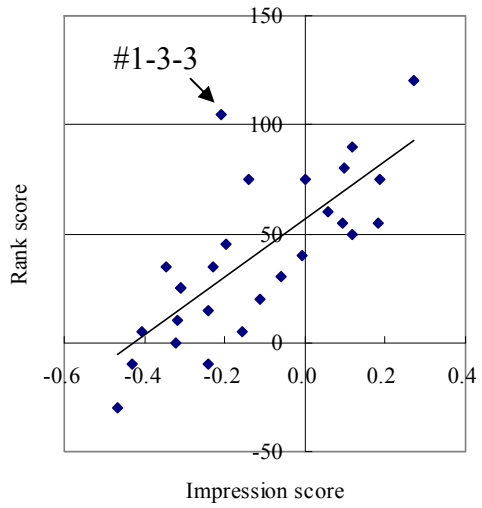

Figure 2: Plots of mean factor vs. rank scores

Therefore, the more positive a scene's impression was, the better its ranking was. The scene #1-3-3 indicated by the upper arrow, however, is distant from the correlation line. This means that even though the scene received a rather negative impression rating, it was judged a better discussion than other scenes. The mean factor score of the *relationships of participants* (Factor 3) is especially very large in a negative direction; that is, the scene was seen as *antipathetic*, *selfish*, *disconnected*, *inconsistent*, *winding*, and *self-centred*, as Table 1 indicates. Table 3 shows the correlation between the factor and rank scores with *p* values. Only the factor of *relationships of participants* does not have a correlation to rank score, while others have significant correlations. Certainly, we must be careful to regard the impression rating results as an index of discussion quality, because negative factor scores do not automatically denote a bad discussion. Does Factor 3 have nothing to do with a good discussion? Are not sympathetic or collaborative behaviours required for a good discussion?

Table 3: Correlation between factor and rank scores

	Kendall's τ	*P* (** p<0.01)
Activeness	0.527	**
Multidirection & unification	0.452	**
Relationships	0.118	n.s.
Development & Sophistication	0.591	**
Secerity	0.464	**

4. Qualitative Analysis

The fact that Factor 3 (relationships of participants) did not have a correlation to the rank scores means that some scenes were judged good discussion despite negative rating scores (good and negative; GN), and vice versa (bad and positive; BP). Scenes judged as good discussions were also rated positively (good and positive; GP), and vice versa (bad and negative; BN). Figure 3 shows the plots of these different patterns of the rank and factor scores of Factor 3. Scene #0-2-3 was rated as negative for Factor 3 and judged as a worse discussion than the others (BN). Even though the mean factor scores of Factor 3 of #0-2-3 and #1-3-3 were similar, #1-3-3 was judged a good discussion (GN). To consider this, we scrutinize the discussion process of the scene.

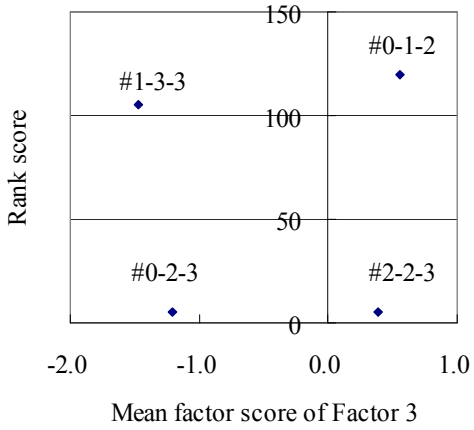

Figure 3: Plots of mean factor score of Factor 3 vs. rank score of four target scenes

Analysis of Conversational Flows

First, we would like to track the flow of statements in scene #1-3-3. The discussion theme (3rd) concerned the use of Wikipedia for writing reports. The following conversations were translated from Japanese to English and summarized.

#1-3-3 'Honesty of students'

1. E: Anybody think that we should allow the use if we regard merely viewing as use?
2. <Other participants other than B raise their hands>
3. B: Merely viewing?
4. C: Yes, only viewing not but citing.

.....

→5. B: But, I think they are differen ..... A book is published by a certain author or a publisher, and they are responsible for the contents..... But Wikipedia is written by anonymous persons, so we don't know who is responsible. That is a problem, I think.

→6. E: I think it depends on the honesty of the user. When one uses Wikipedia, and if the information looks unsubstantiated, it must be verified. Who is the responsible person is obviously another matter for the confirmation of the information. For the interpretation of laws, there are some theories such as the Agatsuma theory. The user must regard Wikipedia articles as just a way of thinking and get additional information from other sources without merely copying a Wikipedia article... (B interrupts E)

→7. B: That may be right, actually…..but..

.....

8. A: If the articles written in Wikipedia are full of mistakes, then all the information on the website is full of inaccuracies. The users should take such risks.
9. B: Yes, I agree.

.....

10. E: Using either books or Wikipedia, the person who doesn't think merely outputs the information from the

source. The problem is the users themselves…..…..
I think that we should allow the use of Wikipedia on the assumption that all users behave honestly.

11. A: On that assumption, I think so, too.

.....

→12. C: Me too.. It is a problem of honesty…But if the subject is not interesting, the student must copies word for word.

13. D: I've done that. <laughter>

14. C: To tell the truth, so have I. If I was not interested in the course, then I copied and pasted articles because I only wanted to get credit even if my grade was low.

15. B: I've done the same thing.

16. E: I'm ashamed to hear that.

17. A: Me too.

18. B: Me too.

.....

19. C: So, for this reason, I don't think that we should allow the use of Wikipedia for writing reports.

.....

20. C: Recently, I feel that most students only do what they want to do, and me too I guess. So, I think that this creates an unexpected situation such as moral decay.

21. E: Your story became a big deal.

22. C: Certainly, my story is becoming a big deal, but I think the key point is whether the students have any interest, and we must consider cases when they are not interested in…

→23. E: But, for the person giving permission, whether the students are interested in the subject has nothing to do with the permission.

.....

24. E: I think that it is high-handed to refuse permission. It's absolutely absurd to copy and paste. I'm sure that some people do copy word for word, even for books or any sources. So I think that those who regard Wikipedia as only a way of thinking may view articles on it.

Through the above discussion, participant E consistently asserted that students may use Wikipedia if using is defined as merely viewing (assertion I). To his statements, participants B and C offered counterarguments in different ways. B mentioned that merely viewing differs from using (line 5, counterargument I). She also mentioned that nobody is responsible for Wikipedia articles (assertion II). E immediately refuted B's statement (line 6, counterargument II). Although B agreed with E's assertion, she seemed to be unsatisfied (line 7). After further statements were made, E reasserted that the students may use Wikipedia if they do within reasonable limits (line 10, assertion I'). After C agreed with E once, he admitted that he had copied and pasted Wikipedia articles, which is dishonest use that E mentioned (lines 12 and 14, counterargument III). C finally asserted that one should not allow the use of Wikipedia (line 19, assertion III), and C extended his assertion to the problem of student' moral decay (lines 20 and 22, assertion IV). E pointed out his assertion's inconsistency (line 23, counterargument IV), and finally E explained his assertion again more strictly (line 24, assertion I''). The following outline of the flow of the statements:

Assertion I (E)
Counterargument I (B)
Assertion II (B)
Counterargument II (E)
Assertion I' (E)
Counterargument I' (C)
Assertion III (C)
Assertion IV (C)
Counterargument IV (E)
Assertion I'' (E)

At the utterances indicated by single arrows in the conversational data, the participants opposed or doubted the previous statement. At a glance, there are many disagreements. Therefore, the score of Factor 3 *relationships of participants* resulted in many negative directions. Despite this fact, the logic of E's assertion and the counterarguments were reasonable. Through the counterarguments to the statements of B and C, E's assertion became sophisticated and focused step-by-step on the end (assertion I''). This might explain the high score.

There were also many disagreements in scene #0-2-3, which received a negative impression and was judged worse than the others (BN). The theme of the discussion was the same as #1-3-3: using Wikipedia in writing reports.

#0-2-3 "I can't be bothered"

1. B: Merely viewing and using are the same if one doesn't cite it in references, aren't they?

2. A: It doesn't matter.

3. D: One doesn't have to cite it in the references.

.....

4. D: Then, is it using if one just viewed the site at first? I think just viewing is OK.

.....

5. D: When my teacher told us not to use Wikipedia for writing a report, what I did was, I actually referred to it, but I didn't include it in the references. Of course, I also referred to other sites, not only Wikipedia.

.....

6. C: If one doesn't cite it, just viewing isn't so bad, I think.

.....

7. C: Although it depends on the degree, it is not a problem to take a short sentence from Wikipedia…….. Copying whole is not writing report.

8. D: That is not only for Wikipedia but also for any other sites.

9. C: Yes, that's right.

.....

10. F: One goes to Wikipedia at first and gets a sense of the big picture of the idea and then examines another site. If the articles are different, then he can refer to other sites. Doesn't he gradually gather meaning this way?

→11. C: At any rate, if he finally refers to the other sites, he doesn't have to refer to Wikipedia for his report, does he?

12. D: You mean you don't agree about the use of Wikipedia for writing report, do you?

13. C: Uh..mm.. That is better, I think.

.....

→14. F: But, don't you just click on the link of the top candidate when you search an unknown word?

15. C: I do. <laughter>

16. D: Most of the top candidates are links to Wikipedia, aren't they?

17. C: Yeah, that's right.<laughter>

→18. B: If you view them, you can ignore them. The problem is that we can't see whether the article in Wikipedia is wrong. If it is inaccurate, I can't see the point of viewing it.

<long silence>

.....

19. F: We must discuss this point from our perspectives as students. There is no point to look at it from the view point of the university or teachers.

20. B: But, in my opinion, I don't think that one should use Wikipedia because it might be inaccurate. I used to use it for its creditability. We don't have to view it at all if it is inaccurate.

→21. C: But it is not all inaccurate.

→22. A: And it is organized well.

→23. D: Besides it is written compactly.

.....

24. B: Then how do I separate the correct from the incorrect?

25. C: Refer to other sites.

<All the participants except B laugh>

→26. B: .. I can't be bothered.

27. C: From the view point of students, it's up to them, it has nothing to do with whether they have permission.

→28. F: I think it is an afterthought that one shouldn't view Wikipedia due to the chance of mistakes. You mean that Wikipedia is useless because it might be inaccurate……

→29. B: An afterthought?

First, participant B claimed that merely viewing Wikipedia is regarded using (line 1, assertion I). D advocated that just viewing is not a problem (line 5, assertion II), and C made a similar assertion that only including a short sentence is not bad (lines 6 and 7, assertion III). D pointed out that it can be the same in any other sources (line 8, assertion IV). F's asserted that one had to use many sources including Wikipedia (line 10, assertion V). C's counterargument was that one does not have to use Wikipedia if he views others (line 11, counterargument V). Following this, he changed his assertion (line 13, assertion VI). F pointed out that one simply clicks the top link of the searching words (line 14, counterargument VI). B argued that one can ignore the link to Wikipedia and asserted that viewing is meaningless (line 18, counterargument VI, assertion VII). After F confirmed the discussion boundaries, B asserted that one should not view Wikipedia due to its uncertainty (line 20, assertion VII'). C immediately pointed out that most articles are not inaccurate. A and D agreed with C about the additional benefits of Wikipedia (line 21, 22, and 23, counterargument VIII'). B asked how to distinguish the correct from the incorrect. C suggested referring to other sites, and B muttered, "I can't be bothered". F resisted B's assertion by describing it as an "afterthought" (line 28, counterargument VIII'). The following is the flow of the statements of #0-2-3:

Assertion I (B)
Assertion II (D)
Assertion III (C)
Assertion IV (D)
Assertion V (F)
Counterargument V (C)
Assertion VI (C)
Counterargument VI (F)
Counterargument VI (B)
Assertion VII (B)
Reassertion VII' (B)
Counterargument VII' (C, D, A, and F)

In the first half of the discussion, some opinions were offered intermediately, but disagreement was not in the open yet. After B's statement of lines 18 and 20 (Assertions VII and VII'), the confrontation between B and the others was exposed. They were certainly antipathetic, selfish and disconnected, and their assertions were inconsistent, winding and self-centred. Such attitudes caused the negative scores of Factor 3. On the other hand, contrary to #1-3-3, the counterarguments by C, D, A, and F didn't appear so logical, and their assertions — one may use Wikipedia because not all the articles are inaccurate — also seem unpersuasive. Besides, B's utterance in line 26; "I can't be bothered" appears too emotional. These might explain the decreased score of #0-2-3.

Scenes #0-1-2 (GP) and #2-2-3 (BP) (details omitted) featured no remarkable confrontations between participants. Therefore, the factor scores of Factor 3 might be rated positive. In #0-1-2, a clever participant moderated very well. He sometimes followed up the statements of others, broadened the perspective, sought input from a silent participant, and often summarized adequately. So, this group had a varied and deep discussion without unnecessary disagreement. But the group of scene #2-2-3 lacked a definite moderator. They amicably discussed, but their assertions were almost the same. That is, they were sympathetic, uniform, shared, and collaborative, and their statements were consistent and linear. For these reasons, they failed to broaden their viewpoints and had a shallow discussion. These points might lead to low scores.

Probative Discourse Tag for Discussions

From the above analysis, we can suggest that existence of the counterarguments, i.e., disagreements, affected the Factor 3 (*relationships of participants*), and those ways of showing influenced quality of discussions. However, this suggestion is more or less visceral and subjective because of the method of analysis. In order to examine the facts more minutely and objectively, we then tried to tag all of utterances in the viewpoint of agreeableness. Tagging schemes for dialogue act or illocutionary force have been modified and improved by many researchers and working groups [e.g., Core and Allen, 1996; Araki et al, 1999]. DAMSL (Dialogue Act Markup using Several Layers; [Core and Allen, 1997]) is one of the most arranged and ordered scheme for understanding dialogue mechanism. In conversation, an utterance can simultaneously

have multiple functions, for instance, “Thank you” is an expression of thanks, positive feedback about understanding and acceptance, and an indication of dialogue closure [Bunt, 2007]. DAMSL can describe these functions by using the notion of multi-layers. Although the tag set is certainly useful for various types of dialogue scenes, we would like to focus on illocutionary aspect of utterance whether it is affirmative or not. Thus, we introduced probative Discourse Tags for Discussion (pDTD) in focusing on agreement and disagreement based on the idea of DAMSL, and tagged four target scenes to compare statistically. Galley et al [2006] proposed a method of identifying agreement and disagreement in conversation. By using explicit expressions of agreement (disagreement), such as “Yes (No),” “I agree (disagree),” and other positive (negative) expressions and prosodic features, are used as clues in their procedure. However, in the spontaneous conversation, agreement and disagreement are not always exhibited by explicit words. Pomerantz (1983) mentioned that delay or absence of respond can be a sign of disagreement. We judged the speaker’s utterance as agreement or disagreement by not only meaning of words, but contexts and some nonverbal information such as timing of reply, speech rate, vocal expression, facial expressions and gestures, etc. Agreement and disagreement can also differ in degree from weak one to strong one. Our tagging procedures are followings.

Table 4: Modified Dialogue Act

Modified DA	Descriptions
inform	talk descriptively some facts, memories, opinions
complement	add some complements, "…. Because …", "In other words, …", "Furthermore, .."
accept	"Yes", "Yes I do", "I agree"
req-agr	require someone's agreement; "Isn't it …?" "Don't you think that …?"
reject	"No", "I don't think so..", "That's wrong…", "But,
reserve	make some hesitative utterance; "Uh…", "uhmm
collaborate	two or more persons chain words to complete a sentence, or repeat same phrase
self-deny	back down on one's proposal, or tone down
vote-Q	"Does anyone agree that..?", "Anybody think that
self-accept	Vote for own vote-Q
clarify-Q	"What does that mean?", "So, if so, what do you
confirm-Q	"Is that true?", "You said what?"
response	reply to a question
emote	make affective utterance; "It's amazing!", "Oh!
accept/inform	*inform* added to *accept*
accept/complement	*complement* added to *accept*
reject/inform	diagree with some grounds
accept/req-agr	*req-agr* after *accept*
reject/complement	disagree with some gorunds
collaborate/req-agr	*req-agr* with *collaborate*
complement/req-agr	*req-agr* after *complement*

First, we segmented the conversations to the utterance units. We used J-Slash unit [JSAI, 2002] based on DFL (Disfluency Annotation Stylebook for the Switchboard Corpus) [Meteer, et al., 1995]. Slash unit is an utterance unit corresponding with a sentence in writing. Second, we annotated the exchange structures (IRF: Initiation, Response and Follow-up) [Coulthard, 1985] to identify the relationships of initiation/response in each exchange. Third, we tagged modified dialogue act characterized by the feature shown in Table 4 to each unit. This labels show the superficial function of each utterance. Fourth, we labeled pDTD to each utterance by following the rule shown in Table 5. According to Schegloff and Sacks [1973] and [Levinson [1983], a first part of adjacency pair [Schegloff and Sacks, 1973] requires a conditionally relevant second part, i.e., question and answer, and proposal and accept. We took this idea into account for our analysis.

Table 5: probative Discourse Tags for Discussion

pDTD	Descriptions
propose	new or modified proposal or assertion
reason	ground of propose or assertion
question	ask something against someone's opinion
answer	reply to the question
downgrade	tone down one's assertion
exemplify	give examples concerning one's assertion
agree	explicit agreement
disagree	explicit disagreement
reserve-agree	reservation of agreement
weak-agree	implicit agreement
weak-disagree	implicit disagreement
agree/propose	agree to propose
agree/reason	agree with reason
disagree/propose	disagree to propose
disagree/reason	disagree with ground
weak-agree/propose	weak-agree to propose
weak-agree/reason	weak-agree with reosen
weak-disagree/propose	weak-disagree to propose
weak-disagree/reason	weak-disagree with ground
substream	subsidiary interactions out of main topic

Table 6 shows the distribution of pDTD of four target scenes, #1-3-3 (NG), #0-2-3 (NB), #0-1-2 (PG) and #2-2-3 (PB). We can see a fact that disagreements appeared more frequent in the scenes of negative Factor 3’s score (#1-3-3 and #0-2-3) than in the scenes of positive (#0-1-2 and #2-2-3). Especially in the scene #0-2-3, the rate of disagreement was large, and agreements were significantly fewer than other scenes. This is a point supporting our suggestion that existence of the disagreements affected the Factor 3 (relationships of participants), and those ways of showing influenced quality of discussions.

The difference between #1-3-3 and #0-2-3 was not only this point. As mentioned above, delay or absence of second part of adjacency pair can be received as a signal of disagreement. In the case of our tag set of pDTD, ‘propose,’ ‘reason,’ ‘question’ and ‘answer’ appear to require some assessments or response. For instance, a proposal requires the assessment of agreeable or disagreeable. And the absence of assessment can be received as disagreement by at least the proposer. Table 7 indicates the rate of each type of second part for four types. The absence of second part of adjacency pair is expressed as ϕ. In the scene #0-2-3, there are six times of absence of assessment to the proposal (propose-ϕ; 37.5% to proposal) , four times of reason-ϕ (23.5%) and three answer-ϕ (75.0%). This might be a reason for downgrading the evaluation of the scene.

There can be multiple recipients corresponding with the second part in discussion. In discussion, agreements by all participants to a proposal can lead a consentient passage.

Even a question to one particular person does not disturb replying by another person who knows the answer. The right columns of each scene in Table 7 are the mean numbers of recipient. In the most target scenes other than #0-2-3, mean numbers of recipient of agreement (propose-agree, reason-agree, answer-agree) exceed those of disagreement, while those of disagreement exceed agreement in the scene #0-2-3. This might be another reason for the low evaluation. These results support our suggestion.

Table 6: Distribution of pDTD of four target scenes

pDAD	1-3-3(NG)	0-2-3(NB)	0-1-2(PG)	2-2-3(PB)
propose	31	16	25	27
reason	24	14	6	21
downgrade	2	0	2	0
exemplify	1	4	8	2
question	0	4	3	6
answer	0	4	4	7
agree	44	16	64	97
agree/propose	0	2	4	3
agree/reason	3	2	2	7
weak-agree	1	2	8	1
weak-agree/reason	1	0	0	0
weak-agree/propose	1	0	0	0
Total agreement	**50**	**22**	**78**	**108**
disagree	0	2	0	1
disagree/reason	2	1	0	0
disagree/propose	1	3	0	1
weak-disagree	1	1	1	1
weak-disagree/reason	3	0	0	0
weak-disagree/propos	0	0	1	1
reserve-agree	2	2	1	0
Total disagreement	**9**	**9**	**3**	**4**
Total utterance	**117**	**73**	**129**	**175**
Rate of Disagree	**7.7%**	**12.3%**	**2.3%**	**2.3%**
Disagree/Agree	**18.0%**	**40.9%**	**3.8%**	**3.7%**

Table 7: Distribution and rate of adjacency pair of four target scenes

	1-3-3(NG)			0-2-3(NB)			0-1-2(PG)			2-2-3(PB)		
	count	rate	recipient	count	rate	recipient	count	rate	recipient	count	rate	recipient
propose-agree	11	50.0%	1.6	8	50.0%	1.1	21	80.8%	1.9	29	90.6%	2.3
propose-disagre	5	22.7%	1.0	2	12.5%	1.5	2	7.7%	1.0	2	6.3%	1.0
propose-φ	6	27.3%	-	6	37.5%	-	3	11.5%	-	1	3.1%	-
reason-agree	21	77.8%	1.3	9	52.9%	1.4	7	87.5%	1.9	17	85.0%	1.8
reason-disagree	4	14.8%	1.0	4	23.5%	1.8	1	12.5%	0.0	0	0.0%	0.0
reason-φ	2	7.4%	-	4	23.5%	-	0	0.0%	-	3	15.0%	-
question-answer	0	0.0%	0.0	4	100.0%	1.0	4	100.0%	2.0	6	100.0%	1.3
question-φ	0	0.0%	-	0	0.0%	-	0	0.0%	-	0	0.0%	-
answer-agree	0	0.0%	0.0	0	0.0%	0.0	3	75.0%	2.3	5	71.4%	1.6
answer-disagree	0	0.0%	0.0	1	25.0%	1.0	0	0.0%	0.0	1	14.3%	1.0
answer-φ	0	0.0%	-	3	75.0%	-	1	25.0%	-	1	14.3%	-

5. Discussion and Conclusion

There were comparatively frequent disagreements in both scenes #1-3-3 and #0-2-3. It made the factor 3 scores negative. The less disagreement (#0-1-2 and #2-2-3), on the contrary, seemed to make them positive. But the disagreement of #1-3-3 helped make the discussion more reasonable, while that of #0-2-3 created a more unreasonable discussion. This difference might be due to much absence of second part of

adjacency pair and the recipients number disagreeing one proposal or reason in the scene #0-2-3. Not only these, we consider other reasons leading for the difference of two scenes.

Choice of words People formulate disagreement by expressions from many possible candidates. Many words that have negative implications were used in scene #0-2-3, but not in #1-3-3. For instance, "It doesn't matter," "I don't see the point," and "I can't be bothered" suggest comparatively aggressive impressions. It is important to point out mistakes, but one must choose words carefully.

Object of counterarguments In scene #0-2-3, some participants seemed to direct their disagreements not to B's assertion but to B herself. The discussants in #1-3-3 were comparatively careful in this regard. The counterargument must always be aimed at the argument never the person.

Treatment of minority opinions In the conversation of #0-2-3, C mentioned the same opinion as B (line 11 and 13). Despite this, C sided against B (line 21). B seemed isolated and uncomfortable within the group. Even if an assertion were offered by only one person and nobody agrees with it, they should objectively examine the argument as if it were their own.

These points seem to be the differences between the ways of thedisagreements by the members of #0-2-3 and #1-3-3. Perhaps these ways of disagreements in #0-2-3 made the evaluators of this scene feel worse. Such a disagreement can be characterized as *censure*, while #1-3-3 can be called *criticism*.

The other factor scores of scene #1-3-3 (See Table 2), Factor 2 *multidirection and unification of discussion*, Factor 4 *development and sophistication of discussion* and Factor 5 *sincerity of participants*, were also negative, except Factor 1 *activeness of floor* (nearly 0). Concerning Factor 2, the ordered and logical arguments of participant E seemed to disturb the open thinking and the viewpoints of the others. Even if the other participants tried to struggle, they eventually succumbed to E's persuasion. This probably hampered the development of the discussion and made the score of Factor 4 negative. If this assumption is true, why was the scene judged better than most other scenes? Which evaluations, the factor scores or the rank scores, should we trust? Each score shows an aspect of evaluation for discussion. Two evaluation methods showed the same results toward almost all scenes, as illustrated by the correlation analysis. However, the evaluation for scene #1-3-3 did not conform to the correlation. This means that the discussion of #1-3-3 was simultaneously good in a sense and bad in another sense. This evaluation duality is also involved in the notion of disagreement.

A proper disagreement of opinions, i.e., type *criticism*, yields fruitful discussion, but a bad one, i.e., type *censure*, can break off the discussion. A possible solution is to be conscious of the ways of making statements. We now forward the arrangement and the classification of the check points of the communication process for group discussions. We regard the five factors extracted by factor analysis as the important aspects of discussion evaluation. For example, Factor 1, *activeness of floor*, might consist of some subordinative items of positive speaking, positive listening and discussion by all members, and so on. Each item might have some lower behavioural check points; e.g., speaking positively is committing oneself to speaking autonomously without waiting to be called on, breaking the silence anyhow, and/or to replying to questions of all members, and so on. In this study, we obtained the keys to determine the items of Factor 3, *relationships of participants*. Arranging these items, we will make a communication check list for a good discussion with which students can monitor their own behaviour or that of others. This check list, however, was made by analysis of data of Japanese university students. Thus, we cannot easily adapt it to other members of any category, but five aspects of communication might apply to them. We will improve the convenience of the list in the future. Since the tagging scheme of pDTD we introduced in this study is a pilot one, it includes many inadequacy and ambiguity. We also must improve the tagging scheme to make it useful for automatic evaluation of discussions in the feature.

6. References

Bunt, H. (2007). Dialogue Act Annotation. In the presentation on ISO-SIGSEM joint working group meeting. http://www.tc37sc4.org/new_doc/ISO_TC37_SC4_N340_TDG3_DialogueActs_Bunt.pdf

Core, M. G., and Allen, J. F. (1997). Coding dialogs with the DAMSL annotation scheme. In Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines. http://citeseer.ist.psu.edu/core97coding.html

Coulthard, R. M. (1985). *An introduction to discourse analysis* (2nd ed.). Longman.

Galley, M., McKeown, K., Hirschberg, J., and Shriberg, E. (2004). Identifying agreement and disagreement in conversational speech: use of Bayesian networks to model pragmatic dependencies. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics.*

Hall, J (1971). Decisions, Decisions. *Psychology Today*, November l971.

JSAI (2002). *Japanese Slash labelling manual* ver. 2.0.1. The Working group of corpus using for discourse and dialogue study ed., JSAI (The Japanese Society for Artificial Intelligence).

Kitagawa, T., and IYO (2005). *FINRANDMESODDO NYUMON* (*Introduction to Finland Method*). Keizaikai.

Kobayashi, T (2007). *TORANSUSAIENSUNO-JIDAI* (*The Epoch of Trans-Science*). NTT Publishing Co.

Levinson, S. C. (1983). i. Cambridge University Press.

Meteer, M. and others. (1995). *Disfluency Annotation Stylebook for the Switchboard Corpus*. Linguistic Data Consortium, Revised 1995 by Taylor, A.

Morimoto, I., Mizukami, E., Suzuki, K., Otsuka, H., and Isahara, H. (2006). An exploratory study for evaluating and analyzing interactional processes of group discussion: The case of a focus group interview. In Journal of Human Interface Society, 8(1), pages 117–128.

Pomerantz, A (1984). Agreeing and disagreeing with assessments: some features of preferred/dispreferred turn shapes. *Structures of Social Action*, Atkinson, J. M. and Heritage, J. ed., Cambridge University Press, pages 57-101.

Schegloff, E. A. and Sacks, H. (1973). Opening up closings. *Semiotica*, 8, pages 280-237.

Suzuki, K; Morimoto, I; Mizukami, E; Otsuka, H; Isahara, H (2007). An exploratory study for analyzing interactional processes of group discussion: The case of a focus group interview, *AI & Society*. (in printing)

Vaughn, S., Schumm, J. S., and Sinagub, J. M. (1996). Focus Group Interviews in Education and Psychology. Sage Publications, Inc.