Analysis of Sensation-transfer Dialogues in Motorsports

Takeru Isaka, Atsushi Otsuka, Yoko Tokunaga* , Iwaki Tosima
NTT Digital Twin Computing Research Center
Tokyo Japan
{takeru.isaka, atsushi.otsuka, iwaki.toshima}@ntt.com

Abstract

Clarifying the effects of subjective ideas on group performance is essential for future dialogue systems to improve mutual understanding among humans and group creativity. However, there has been little focus on dialogue research on quantitatively analyzing the effects of the quality and quantity of subjective information contained in dialogues on group performance. We hypothesize that the more subjective information interlocutors exchange, the better the group performance in collaborative work. We collected dialogues between drivers and engineers in motorsports when deciding how the car should be tuned as a suitable case to verify this hypothesis. Our analysis suggests that the greater the amount of subjective information (which we defined as "sensation") in the driver's utterances, the greater the race performance and driver satisfaction with the car's tuning. The results indicate that it is essential for the development of dialogue research to create a corpus of situations that require high performance through collaboration among experts with different backgrounds but who have mastered their respective fields.

Keywords: Sensation, Collaborative Dialogue, Group Performance

1. Introduction

Humans are social creatures and share their innermost thoughts through dialogue. The active use of subjective ideas is essential to develop a dialogue system that improves mutual understanding and creativity among humans. Since subjective ideas often contain ambiguity, it is conceivable that poor comprehension conditions could lead to confusion or misdirection within the group. Understanding subjective information is also more difficult when the positions and roles of the interlocutors differ. Therefore, we examined the effect of conveying subjective ideas among people in different positions and roles on group performance during collaborative work.

For such verification, it is essential to have a dialogue resource where several people collaboratively work and actively express their subjective ideas for solving problems that cannot be solved from objective facts alone. An example of such a rare dialogue is between drivers and engineers in motorsports. Group performance in motorsports can be replaced with race performance. Motorsport gives a limited amount of test-driving time before a race. The driver and engineer discuss whether the car is tuned appropriately for the track conditions on the basis of the driver's sensations and decide the optimal tuning. We defined "sensation" as a unique event unfolding in the mind of a speaker, i.e., the driver. In this situation, two experts who have mastered their respective fields can finally solve a problem by agreeing on the same idea.

Professional racing drivers can accurately control their cars under extreme conditions (Reid and Lightfoot, 2019; Reid, 2022), under which the maximum speed of the car can reach over 300 km/h. They exhibit different neuroscientific (Bernardi et al., 2013), cognitive (Land and Tatler, 2001; Lappi, 2022), and sensorimotor (Van Leeuwen et al., 2017; Nishizono et al., 2023) characteristics while driving than non-racing drivers. They can understand the increase or decrease in lap time in 0.05-s increments, depending on how good or bad their driving is [1].

The dialogue we focus on has the following four advantages from the perspective of research execution.

- High Resolution - High-resolution sensations that cannot be expressed through basic emotion classification are expressed.
- Reproducibility - The same person has the same feeling when placed in the same situation. ∵Drivers and engineers are consistent in their utterances as experts.
- Evaluability - Some indicators can be used to evaluate the results of sensation transfer objectively. ∵We can objectively evaluate dialogue results with race results.
- Availability - Sufficient data are available to identify the characteristics of human communication.

Although many language resources have been reported (Scherer and Wallbott, 1994; Gonçalo Oliveira et al., 2022; Dong et al., 2018;

*Currently affiliated with NTT DOCOMO, INC., Tokyo, Japan (E-mail: yoko.tokunaga.az@nttdocomo.com).

[1] https://orientstar-watch.com/blogs/articles/blog11 (Japanese article, accessed March 19, 2024)

LREC-COLING 2024, pages 876–886
20-25 May, 2024.

Zhong and Huang, 2018; Zahiri and Choi, 2018; Anjaria and Guddeti, 2014; Sun et al., 2022; Pegoraro, 2010; Hambrick et al., 2010; Narayan-Chen et al., 2019; Ichikawa and Higashinaka, 2022; Willemsen et al., 2022; Okahisa et al., 2022), to the best of our knowledge, no corpus combines these four advantages. We investigated the impact of the differences in the following driver's utterances on race performance by using dialogues with these advantages.

- 「コーナーが、ちょっと気になる [2]」
  "I'm a little concerned about the corner."
- 「コーナーの出口で、唐突に、リアがゆっくり滑る」
  "Suddenly, the rear end slides slowly on the exit of a corner."

These are expressions by which the driver attempts to convey the car's condition as they perceive it. From an emotional viewpoint, both utterances may be categorized as "disgust," and it is difficult to distinguish between them. However, comparing these two utterances in terms of the resolution of the specific event, the latter expresses the sensation of corner driving in more detail and contributes more to group performance than the former. We consider the detailing and sharing of sensations as the essence of dialogue and verified this in this study. We refer to "sensation" as the specific event in the speaker's mind. A person who has mastered something can have a high-resolution sensation of the related event that they have mastered (Yarrow et al., 2009). The contributions of this study can be summarized as follows.

- To the best of our knowledge, this is the first study in which dialogue results, which are mainly subjective exchanges, were evaluated both subjectively by the speaker and objectively through group performance.
- We collected and analyzed dialogues for the domain of "motorsports."
- In a dialogue corpus consisting mainly of subjective exchanges between people in different positions, the results of the analysis suggest that when the speaker's utterance contains many sensation words, the speaker's satisfaction and objective performance of the competition improve.
- The verification results suggest that it is essential for the development of dialogue research to create a corpus of situations in which experts with different backgrounds who have mastered their respective fields collaborate and perform at a high level.

This study also contributes to the sporting genre of motorsports. The verification results suggest the importance of the quality of dialogue between drivers and engineers in the sport, where almost all efforts to improve race performance have been based on analysis of the mechanical characteristics of the car (Milliken et al., 1995). There are attempts to improve driver performance physically (Lappi, 2018). The verification results in this study indicate the possibility of improving race performance through dialogue intervention.

2. Related Work

2.1. Emotion-Text Corpus

Our goal was to quantitatively evaluate the sensations contained in utterances. Emotion and sentiment have been widely used as general indicators of the internal states of humans.

One of the most widely used corpora that combines text and emotion is the ISEAR dataset (Scherer and Wallbott, 1994). This dataset is a collection of emotionally evocative texts to which annotators assigned seven labels (Joy, Fear, Anger, Sadness, Disgust, Shame, and Guilt), which are an extension of Ekman's six basic emotions (Ekman, 1992). The data were obtained from approximately 3,000 respondents in 37 countries. The dataset is intended to infer average and objective human emotions toward text, thus does not record high-resolution sensations.

Most of the available corpora have an unspecified person as the source of the text (Gonçalo Oliveira et al., 2022), e.g., through crowdsourcing. Some corpora contain sensations in the collected utterances by restricting the provider's characteristics. Dong et al. (2018); Zhong and Huang (2018) confirmed that the Chinese have unique linguistic expressions that signify distinctive tastes.

EmoryNLP (Zahiri and Choi, 2018) contains data collected from dialogues between TV drama actors. The words spoken in the TV dramas were selected by a scriptwriter, an expert in handling words. The utterances that contain sensations are similar to the dialogues collected in this study. However, the collected utterances are not subjective expressions, and the annotated information is limited to six categories in the emotional layer.

Anjaria and Guddeti (2014) collected texts posted on Twitter (now known as X) about a specific election and constructed a corpus with three sentiment labels (Positive, Negative, and Neutral) assigned to each text. The senders of the texts had subjective

[2] The dialogues collected in this study are in Japanese and translated into English by one of the authors.

ideas, such as agreeing or disagreeing with the political claims of the election candidates. In addition to multiple people in different positions conveying subjective information, objective evaluation of election results was possible. However, the senders were not guaranteed to have a high-resolution sensation. The annotated representations of subjective information were also limited to three categories.

Although there are language resources from professional athletes during non-competition, such as answers to interviews (Sun et al., 2022) and Twitter posts (Pegoraro, 2010; Hambrick et al., 2010), there are no language resources during the decision-making phase of a competition.

2.2. Collaborative Dialogue

Narayan-Chen et al. (2019) used Minecraft and set up a task in which one dialogue participant, who knew the correct structure to build, instructed the other dialogue participant to do the work. In such a task, conveying objective information on the basis of the correct answer rather than subjective ideas in the dialogue will likely improve performance as a group.

Ichikawa and Higashinaka (2022) also used Minecraft and tasked dialogue participants with building a garden that they thought would be good for them. This study used a creative task with no correct answer. However, the dialogue participants were not necessarily gardening experts and did not convey high-resolution, reproducible sensations.

Willemsen et al. (2022) used a sorting game, a task in which dialogue participants sort multiple images on the basis of a rationale. The study focused on objectively recording the relationship between visual references and utterances. The collected dialogues were based on objective visual information and did not include much of the speaker's sensations.

Okahisa et al. (2022) collected dialogues from interviews that actively elicited knowledge from professionals. However, the study focused on the mechanisms by which the interviewer elicits knowledge from the interviewee and did not evaluate the dialogue results regarding group performance.

2.3. Social Decision-making

Researchers in social decision-making have investigated the relationship between dialogue and group performance. Bahrami et al. (2010) asked two participants to identify low-order visual stimuli using a Gabor patch (a simple visual pattern used in perceptual experiments). They found that the percentage of correct responses was higher when communication between them was allowed than when not. The effect was verified on the basis of the dichotomous value of whether dialogue occurred. The dialogues contained rough granularity information about whether the participants saw the target stimulus. Although the study showed a relationship between the presence or absence of dialogue and performance, it did not clarify the relationship with dialogue quality.

Tohyama and Shirouzu (2018) analyzed the characteristics of dialogue when children solve mathematics problems while discussing them. They concluded that clarifying the meaning of numbers and expressions given as problems through dialogue increases the percentage of correct answers. They showed that the quality of dialogue affects group performance. However, there were textbook solutions and correct answers to a given task. Accurately understanding objective facts rather than subjective information leads to improved group performance.

3. Data Collection

Pairs of driver and engineer	2
Dialogues	38
Utterances of Driver 1	547
Utterances of Driver 2	532
Utterances of Engineer 1	1036
Utterances of Engineer 2	805
Ave. utterances of Driver 1 per dialogue	28.79
Ave. utterances of Driver 2 per dialogue	28
Ave. letters of Driver 1 per utterance	30.34
Ave. letters of Driver 2 per utterance	28.17

Table 1: Statistics of collected dialogues

To collect dialogues containing high-resolution sensations, we recorded and transcribed dialogues between two pairs of drivers and engineers belonging to the same team over radio communication during five rounds of All Japan Championship Super Formula (Super Formula) held in a specific year. Super Formula is the highest motorsports category in Japan. Since the drivers and engineers who were the targets of the dialogue collection have achieved a certain level of continuous success in Super Formula, they are equipped with a wealth of knowledge to win races. Since the drivers and engineers belong to the same team and handle the same specifications of cars, they equally share a certain amount of knowledge and skills.

A round of Super Formula consists of free practice, a qualifying race, and a final race. During the time between free practice and the start of the qualifying race, engineers tune the car to run as fast as possible. One of the factors used by the engineers to determine the car's condition is the driver's comments after driving the car during the free practice.

Such comments are important [3] for the engineers to determine the car's condition. Table 1 lists the statistics of the collected dialogues. We refer to the two drivers as Driver 1 and Driver 2 and refer to Driver 1(2)'s interlocutor as Engineer 1(2).

3.1. Qualitative Characteristics of Collected Dialogues

We present a case in which the sensations contained in the utterances of the collected dialogues significantly affected the race results. Figure 1 shows the dialogues between the two pairs of drivers and engineers before the same final race. The horizontal axis represents time, and the utterances of the drivers and engineers are shown in chronological order. Utterances expressing dissatisfaction with the car's condition are shown in red. The drivers and engineers had discussions about determining the optimal tuning until the start of the race.

For the upper pair of Figure 1, Driver 2 repeatedly uses the same informative word, "dizzy." He also utters, "I don't know what to do," which contains almost no sensations. These words provide little information for Engineer 2 to make effective tuning decisions. Therefore, the race result was weak.

For the lower pair, Driver 1 reported the car's condition in a more detailed and multifaceted manner than Driver 2, such as "slips out a little," "slip quite abruptly," and "a left corner." Driver 1 provided more information for Engineer 1 to decide on the car' s tuning than Driver 2. Therefore, the race result was stronger.

On the basis of the relationship between the qualitative characteristics of the dialogue and race performance, we hypothesized that the more the driver verbalizes sensations about the car's condition, the more likely the engineer will have more information to make decisions to improve the condition. Therefore, the better the car performs as the driver hoped, the better the race performance. We conducted a quantitative verification of this hypothesis, as discussed in the next section.

4. Verification of Effect of Sensation Amount on Race Performance

The hypothesis described in the previous section was verified by decomposing it into the subjective satisfaction of the drivers and objective race performance. Among the dialogues collected (for five rounds), we sampled the dialogues during the free practice and qualifying race for three of the rounds (Rds. A, B, and C) for verification. We excluded the other two rounds from the verification because those rounds had no free practices. We also excluded the final race from the analysis. The qualifying race is a time attack, and the final race is for the finishing order. The competitors easily influence the results of the final race. Therefore, we adopted the qualifying race time as an appropriate indicator of race performance to verify our hypothesis described in Section 4.5.

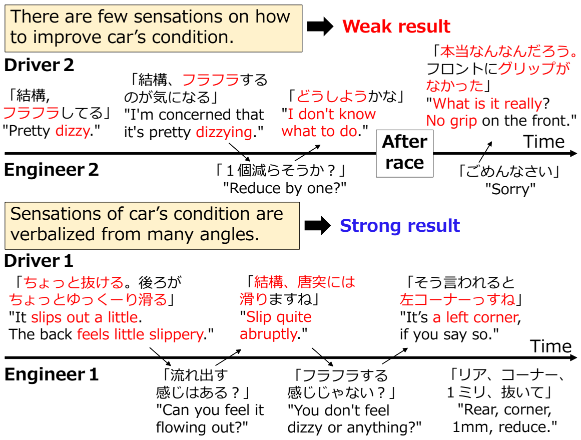

Figure 1: Qualitative characteristics of collected dialogues. Excerpts of dialogues in same race, with lower-performing pair shown at top and higher-performing pair at bottom.

For quantitative verification, we defined the amount of sensations in the driver's utterance as the information score. There are three different information scores depending on how it is calculated.

4.1. Calculation of Information Score using Term Frequency-inverse Document Frequency

The first information score, denoted as $I_{\text{score-T}}$, is quantitatively defined as

$$I_{\text{score-T}} = \sum_{i=1}^{n} \text{TF-IDF}_i, \qquad (1)$$

where n represents the total number of morphemes within all utterances of the driver in the specified dialogue segment. Here, TF-IDF_i is term the frequency-inverse document frequency (TF-IDF) for the i^{th} morpheme. The TF-IDF (Salton and McGill, 1983) is a measure of the importance of a morpheme in a document and used as an indicator of textual emotion classification (Cahyani and Patasik, 2021). It is calculated as the product of the number of occurrences of a morpheme in a document (TF) and the log of the value obtained by dividing the total number of documents by the number of documents in which the morpheme appears (DF). One document is defined as the set of

[3] https://www.espn.co.uk/f1/story/_/id/22433084/inside-vital-relationship-f1-driver-engineer (Accessed March 19, 2024)

morphemes contained in the dialogue data of a pair of a driver and engineer for one run (free practice or qualifying race). The number of documents used to calculate DF was 89, consisting of 38 documents of all the dialogue data of the two pairs of drivers and engineers for all 5 rounds and 51 documents of group-task dialogue data unrelated to motorsports. The TF-IDF of words unique to motorsports is calculated higher, while that of words commonly used in daily conversation is calculated lower by adding group-task dialogues.

We used MeCab (Kudo, 2005) as a library for morphological analysis of utterances and used mecab-ipadic-NEologd (Sato, 2015) as a system dictionary within MeCab. If many morphemes with small TF-IDFs are contained in the morphemes used to calculate the $I_{\text{score-T}}$, morphemes with large TF-IDFs will have less weight in this $I_{\text{score-T}}$, and the amount of sensations in the utterance will not be represented appropriately. For example, even if a person speaks a clerical utterance at length that contains many objective facts with a small amount of sensations, the $I_{\text{score-T}}$ will be large. Therefore, morphemes with a DF of 50 or more, which is more than the majority of the total number of documents, were excluded from the TF-IDF calculation.

Table 2 (a) lists the top 20 morphemes of TF-IDF of all the utterances of Driver 1 during Rd. B, and Table 2 (b) lists the bottom 20 morphemes of TF-IDF. Table 2 (a) contains many morphemes expressing the parts and condition of the car, and Table 2 (b) contains more general terms and fillers that are less related to the car's condition than those in Table 2 (a). Thus, the TF-IDF can extract the desired morphemes, and the $I_{\text{score-T}}$ is appropriate as an indicator of the amount of sensations.

(a) Top 20 morphemes of TF-IDF	(b) Bottom 20 morphemes of TF-IDF
後ろ (back), リア (rear)	ちゃんと (properly)
ギャップ (gap), 方向 (direction)	しよう (do), 多分 (probably)
ブレーキ (brake)	ずっと (the whole time)
コーナー (corner)	他 (other), いける (good)
ナーバス (nervous)	最初 (first), じゃ (well then)
介入 (intervention)	なかなか (quite), やつ (thing)
ロールオーバー (rollover)	辺 (vicinity), 来る (come)
硬い (firm), 広げる (expand)	なんだろう (wonder)
戻す (restore), インナー (inner)	最後 (last), あのー (errr ...)
フロント (front), 高 (high)	逆 (reverse), ここ (here)
ポジ (position), ウォーム (warm)	そんなに (so much)
オプション (option)	わかる (understand)
踏む (step), 感じる (feel)	行ける (look good)

Table 2: Top 20 (a) and bottom 20 (b) morphemes of TF-IDF in all utterances of Driver 1 in Rd. B

4.2. Calculation of Information Score using Subjective Keywords

The $I_{\text{score-T}}$ described in the previous section uses an objective index as TF-IDF. We introduce the second information score directly reflecting human subjectivity, $I_{\text{score-S}}$, which is defined by the expression:

$$I_{\text{score-S}} = \sum_{i=1}^{m} K_i, \tag{2}$$

where m indicates the total number of subjective keywords identified across all utterances of the driver within the specified dialogue segment, and K_i represents the occurrence of the i^{th} subjective keyword. We also define the subjective keywords as morphemes extracted by us from the five perspectives listed in Table 3 regarding tuning decisions among all collected dialogues. These five perspectives were confirmed by interviewing the engineers from which the dialogues were collected as valid perspectives to be kept in mind when tuning the car.

Perspective	Concrete example
Problem event	アンダーステア (understeering), ボトミング (bottoming)
Car part with pronounced problems	タイヤ (tire), フロント (front)
Course area with pronounced problems	カーブ (curve), 立ち上がり (standing up)
How problem arises	唐突 (sudden), ゴリゴリ (gorigori: Onomatopoeia for abrasive sounds)
Other expression useful for judgment	迷う (be puzzled), ない (not)

Table 3: Perspectives when extracting subjective keywords

4.3. Calculation of Information Score using BERT

Building upon the set of subjective keywords detailed in Section 4.2, we introduce the third information score, denoted as $I_{\text{score-B}}$ and defined as

$$I_{\text{score-B}} = \sum_{i=1}^{p} \text{TF-IDF}_{\text{BERT},i}, \tag{3}$$

where p denotes the number of morphemes extracted via BERT (Devlin et al., 2019) from all utterances of the driver within the specified dialogue segment. The term $\text{TF-IDF}_{\text{BERT},i}$ signifies the TF-IDF of the i^{th} morpheme, identified using BERT. We fine-tuned a Japanese-trained BERT [4] with a dataset for named entity recognition. This dataset specifies which section of the driver's utterances contains the subjective keywords discussed in the previous section. We created the dataset using 1157 sentences of the driver's utterances and divided it into 694 sentences of training data, 231 sentences of validation data, and 232 sentences of

[4] https://github.com/cl-tohoku/bert-japanese (Accessed March 19, 2024)

Perspective	Evaluation value
Traction	Five levels from 1 to 5 1: Bad 2: Slightly bad 3: Neither good nor bad, or not mentioned 4: Slightly good 5: Good
Grip	
Brake	
Bottoming	
Steering	
Load balance of car body	
Tire internal pressure	
Aerodynamic characteristics	
Suspension	Average of evaluation values from 11 perspectives on left is defined as driver satisfaction.
Stabilizer	
Overall good or bad	

Table 4: Perspectives to evaluate driver satisfaction

test data. We conducted fine-tuning with the number of epochs of 30 and batch size of 256. The test results of fine-tuning indicate a precision of 0.679, recall of 0.586, and F value of 0.629.

These are the evaluated values when the inferred interval containing the subjective keywords and the interval in the test data match perfectly. When we considered the inferred interval was correct under the condition that it encompassed the interval of the test data, precision was 0.899, recall was 0.781, and F value was 0.836. For example, if the inference result of BERT was “front wing” and the first half of this inference result, “front,” was present in the test data, we considered it correct. These evaluations showed that the training results are accurate enough to automatically extract subjective keywords in the utterance without needing to omit many subjective keywords. Using these three information scores, we examined the effects of the amount of sensations in the driver’s utterances on driver satisfaction and race performance.

4.4. Verification and Results with Subjective Evaluation Values

We conducted a subjective evaluation to verify our hypothesis. Between free practice and the start of the qualifying race, the drivers and engineers update the car’s tuning by repeating the following three-step cycle as time permits.

1. Test driving
2. Feedback on car’s condition
3. Consideration and determination of tunings

The feedback of the car’s condition provided by the driver in Step 2 has two roles: to evaluate the tuning determined in the previous cycle and provide sensations that can be used to decide how to improve the tuning. If our hypothesis holds, the greater the amount of sensations the driver verbalizes in one cycle, the greater the driver’s satisfaction with the tuning will be in the next cycle. To verify this, we calculated driver satisfaction with the tunings for each cycle.

Table 4 lists the perspectives the drivers used to evaluate the cars’ conditions. These perspectives were confirmed by interviewing the engineers whose dialogues were collected and generally covered what the drivers were concerned about. For the driver’s utterance in Step 2 (Feedback on car’s condition), one of the authors rated the car’s performance from each of the 11 perspectives on a 5-point scale from 1 to 5 (the higher the rating, the higher the driver’s evaluation). For example, if the driver said, “losing traction,” the author assigned 1 as the evaluation value of traction. Since the driver did not mention all the 11 perspectives in each cycle, a uniform evaluation value of 3 was given to those not mentioned. We define the average of the evaluation values of the 11 perspectives for each cycle as the driver’s satisfaction for the cycle.

We verified our hypothesis by using driver satisfaction and the three information scores introduced in Sections 4.1 to 4.3. If the hypothesis holds, the information scores in one cycle would positively correlate with the driver-satisfaction gain in the next cycle. The driver-satisfaction gain represents the increase or decrease in driver satisfaction compared with the previous cycle. If the satisfaction in cycle t refers to S(t), Equation 4 expresses the satisfaction gain ∆S as

$$\Delta S = S(t) - S(t-1) \quad (4)$$

4.4.1. Results of Hypothesis Verification

Figure 2 shows the results of verifying our hypothesis. The horizontal axis shows the normalized $I_{\text{score-T}}$s for all cycles included in Rds. A, B, and C. The $I_{\text{score-T}}$s are normalized across all samples so that the maximum value is 1. The vertical axis represents the driver-satisfaction gain. A total of 42 samples were obtained for the two drivers. The analysis revealed a statistically significant positive correlation between the normalized $I_{\text{score-T}}$ and the driver-satisfaction gain, with a correlation coefficient (r) of 0.499 and a 95% confidence interval (CI) of [0.165, 0.661]. The $I_{\text{score-S}}$ analysis also indicated a statistically significant positive correlation, with r = 0.464 and a 95% CI of [0.187, 0.673]. Similarly, the $I_{\text{score-B}}$ results showed a significant positive correlation, with r = 0.518 and a 95% CI of [0.254, 0.710].

Figure 3 shows the $I_{\text{score-T}}$ and Driver 2’s satisfaction per cycle in Rd. B. Each point on the horizontal axis represents a cycle, with the blue circles representing the $I_{\text{score-T}}$ for a cycle and the red squares representing the driver’s satisfaction. Let us focus on the two consecutive cycles in the time series enclosed with the green oval. We can see the following trend regarding our hypothesis. A higher (lower) $I_{\text{score-T}}$ in the previous cycle increases (decreases) the driver’s satisfaction in the next cycle.

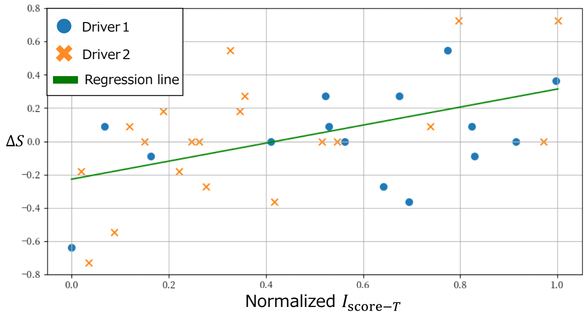

Figure 2: Relationship between $I_{\text{score-T}}$ and driver-satisfaction gain (ΔS)

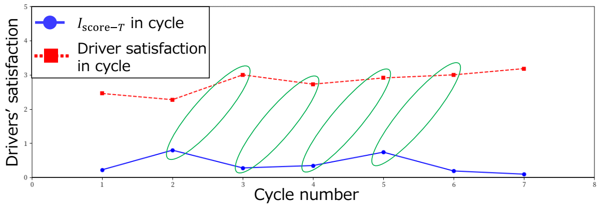

Figure 3: Time-series changes in $I_{\text{score-T}}$ and Driver 2's satisfaction in Rd. B

4.5. Verification and Results with Objective Evaluation Values

We then verified the effect of the amount of sensations in a driver's utterance on race performance. We needed to score the results of the qualifying races for this verification. Since motorsports are not always held on the same course, using race time as an indicator would make it impossible to separate whether the car's condition or course-related factors, such as course length and difficulty, affects the time.

To remove course-related factors, we therefore used the ratio of the target driver's time to the time of the driver who placed first in the qualifying race to measure race performance, as shown in Equation 5. We refer to this ratio as the time score. The time score is between 0 and 1, with a higher score indicating better race performance. Our hypothesis was verified on whether the information scores and time scores positively correlated within the same round.

$$(\text{Time score}) = \frac{(\text{Top qualifying driver's time})}{(\text{Subject driver's time})} \tag{5}$$

4.5.1. Results of Hypothesis Verification

Figure 4 shows the verification results. The horizontal axis shows the normalized $I_{\text{score-T}}$s for Rds. A, B, and C. The $I_{\text{score-T}}$s were normalized across all samples so that the maximum value is 1. The vertical axis represents the time scores in all three rounds. A total of six samples were obtained for the two drivers. The correlation coefficient was r = 0.694, indicating a positive correlation. The $I_{\text{score-S}}$ result was r = 0.737, and $I_{\text{score-B}}$ result was r = 0.701. However, the results are not statistically reliable due to the insufficient sample size.

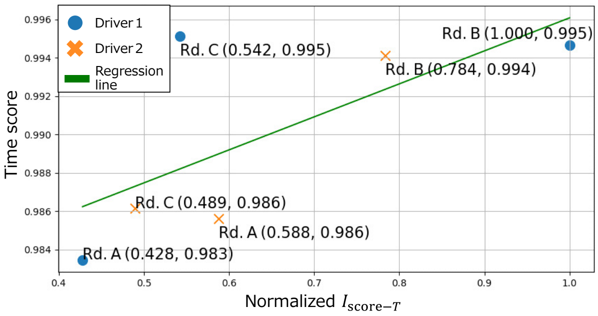

Figure 4: Relationship between $I_{\text{score-T}}$ and time score

5. Discussion

5.1. Case Study for Verification Results

The results of the subjective and objective verifications in the previous section support our hypothesis. We examined whether the drivers' actual utterances confirm these verification's results. In Figure 4, the sample of Driver 1 for Rd. B had the highest $I_{\text{score-T}}$ and the second-highest time score. Figure 5 shows the primary utterances of Driver 1 in chronological order, from free practice to the start of the qualifying race in Rd. B. The horizontal axis represents time.

The utterances expressing satisfaction regarding the car' s condition are shown in blue, and those expressing dissatisfaction are shown in red. Driver 1 verbalized the car's condition from various angles from the beginning to the middle, such as “seems to lock,” “feeling of my hip being crushed,”

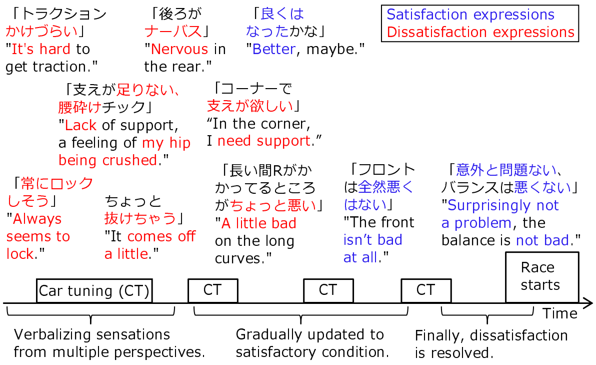
「トラクションかけづらい」
"It's hard to get traction."
「後ろがナーバス」
"Nervous in the rear."
「良くはなったかな」
"Better, maybe."
Satisfaction expressions
Dissatisfaction expressions
「支えが足りない、腰砕けチック」
"Lack of support, a feeling of my hip being crushed."
「コーナーで支えが欲しい」
“In the corner, I need support.”
「常にロックしそう」
"Always seems to lock."
ちょっと抜けちゃう」
"It comes off a little."
「長い間Rがかかってるところがちょっと悪い」
"A little bad on the long curves."
「フロントは全然悪くはない」
"The front isn’t bad at all."
「意外と問題ない、バランスは悪くない」
"Surprisingly not a problem, the balance is not bad."
Car tuning (CT)
CT
CT
CT
Race starts
Time
Verbalizing sensations from multiple perspectives.
Gradually updated to satisfactory condition.
Finally, dissatisfaction is resolved.

Figure 5: Time series of Driver 1’s utterances in Rd. B

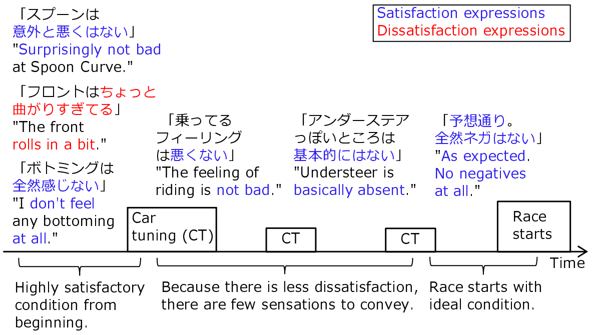

Figure 6: Time series of Driver 1's utterances in Rd. C

and "need support", then gradually increased the number of statements such as "better", "not bad", and "not a problem", indicating that he was satisfied with the tuning. This driver placed 3rd out of 11 in the qualifying race. The actual driver's utterances suggested that as the amount of sensations included in the utterance increased, the driver's satisfaction and race performance improved.

We also investigated another sample of Driver 1 in Rd. C. In Figure 4, this sample appears to be an outlier, as its time score is the largest among all samples, despite its relatively small $I_{\text{score-T}}$. Figure 6 shows the primary utterances of Driver 1 in chronological order, from free practice to the start of the qualifying race in Rd. C. For the car's condition, utterances expressing satisfaction are shown in blue, and utterances expressing dissatisfaction are shown in red. Driver 1 expressed that there was no concern regarding the car's condition, such as "not bad" or "not feel at all" from the beginning. Furthermore, "as expected" or "no negatives" just before the race indicated that he had no complaints about the car's condition. The result was excellent, placing first out of 10. On the basis of these utterances and the race results, this sample can be interpreted as follows. Since the car was in perfect condition from the beginning, Driver 1 had little sensation to improve the car's condition, and the race result was excellent.

The verification results in Section 4 and the two case studies described above support our hypothesis. However, we must carefully determine the cause-and-effect relationship by considering the initial condition of the car, actual utterances, and the environment. The same trend as our hypothesis will appear, particularly when cars still need a large amount of tuning.

5.2. Differences in Results with Method of Calculating Information Scores

Regarding the verification results in Section 4, the $I_{\text{score-T}}$ did not differ significantly from the $I_{\text{score-S}}$ and $I_{\text{score-B}}$ for the correlation coefficient. Thus, the TF-IDF reflects human subjectivity.

The purpose of using BERT in Section 4.3 was to extract latent subjective keywords that humans are unaware of and reflect them in the information score. As expected, the correlation coefficient between the $I_{\text{score-B}}$ and driver-satisfaction gain exceeded $I_{\text{score-T}}$ and $I_{\text{score-S}}$, as described in Section 4.4.

BERT also extracted words that were not included in the correct set of subjective keywords but should be included. Two examples are shown below. Words in the driver's utterances that were not included in the set of subjective keywords but inferred with BERT are shown in bold.

- 「なんかずーっとフロントが**ふわふわふわふわ**ずっとしているのが気になるんですよね」
  "I feel like the front end is **fluffy** and **fluffy** all the time."
- 「タイヤが動いているからなのか 、うーん、**収まり**がすごい悪い」
  "Maybe it's because the tires are moving, umm, it's ill-**fitting**."

The words "fluffy" and "fitting" inferred with BERT are essential expressions for the car's tuning. These words convey to the engineer the subtle nuances of the driver's experience. Although the training data set used for BERT fine-tuning was small (1157 sentences), the above examples suggest that automatic extraction of sensations may be feasible with accuracy exceeding that of humans by enhancing the training data.

5.3. Influence of Utterance Length and Driver Personality Traits

The longer the utterance, the more sensations it is considered to contain. However, as the time available to speak in a race is limited and drivers try to speak as much as possible in that limited time, we consider that the length of utterances varies little between drivers. Relatedly, each driver is potentially chatty or reserved as an innate human trait. However, in professional motorsports, the influence of the number of words spoken as a personality trait is small. For drivers, their honor, life, and livelihood are at stake in a race. To drive the car as fast as possible, drivers speak up without hesitation about anything bothering them. We have also heard from the engineer from whom we collected dialogues that "drivers will point out anything they feel like pointing out without hesitation."

5.4. Correlation between Driver's Satisfaction and Performance

To reinforce the verification described in Section 4.4, analyzing the correlation between driver satisfaction/dissatisfaction and driving time is useful. However, analyzing this relationship during free practices is difficult because drivers do not always drive at full speed (sometimes they dare to meander to check grip). Once we have collected enough data with the car running at full speed during free practices, we will work on verifying this correlation.

6. Summary and Future Work

We defined "sensation" as a unique event unfolding in a speaker's mind, and clarified that the amount of a speaker's sensations in their utterances during collaborative work contributes to improving group performance. We collected dialogues on motorsports, a rare situation in which multiple top professionals collaboratively work and actively exchange their innermost thoughts. As a result of the observation of the collected dialogues, we hypothesized that the more sensations the driver provides about the car's condition, the more the engineers can make decisions to improve the condition. Therefore, the more the car's condition meets the driver's satisfaction, the more race performance is improved. We scored the sensations of the driver's utterances, satisfaction, and race performance to verify our hypothesis. The correlation coefficients between each score and the case studies support this hypothesis.

We analyzed dialogues from two drivers in three rounds. Augmenting the data set from more drivers and rounds is a top priority for future work for statistically reliable confirmation.

We also need to consider more plausible methods of sensation calculation. The main focus of this study was to verify whether there is a correlation between sensation and performance. To do this, we scored sensations as scalar quantities using basic methods such as TF-IDF, the number of subjective keywords, and BERT, as mentioned in Section 4. When discussing the degree of effectiveness in the future, it will be worthwhile to scrutinize multiple methods. Also, collecting sufficient data will enable us to express sensations in richer formats, such as vector representations.

6.1. Application to Support Communication

The information score can be used to evaluate the extent to which the speaker's sensations are included in the text, in the context of text generation (Brown et al., 2020) and textual intent estimation (Frank and Goodman, 2012). For example, one application of the findings of this study could suggest questions to the engineer to elicit the driver's sensations.

As mentioned in Section 1, ambiguities in sensations may cause confusion among interlocutors or lead a group in the wrong direction. Particularly in motorsports, many digital sensors are installed to measure the car's precise conditions. Relying solely on these sensors' data without the driver's sensations would result in accurate tuning. Nevertheless, our verification results suggest that driver's sensations improve race performance and driver's satisfaction. Humans can be highly excellent sensors with a resolution surpassing digital sensors. They can also be decoders capable of accurately conveying the sensations. This study aimed to achieve communication-support technology that will strengthen top professionals and create a world where people who met yesterday and today can communicate as if they were old friends who have known each other for ten years.

7. Limitations

It remains to be discussed whether the findings obtained in this study can be applied to other collaborative dialogues. We considered that the same trends as the hypotheses verified in Section 3.1 will appear for interactions between high-resolution and reproducible sensation holders. Experts (e.g., professional athletes) who have mastered a specific field, not limited to motorsports, possess high-resolution and reproducible sensations (Yarrow et al., 2009). How abundantly such experts' sensations can be shared with others will help improve group performance and creativity. Verifying whether this hypothesis holds for other use cases is for future work. Since the speakers in this study were top professionals with precise sensations, the verification results may have supported our hypothesis. If the sensations were incorrect or not reproducible (e.g., novices or amateurs), conveying them would have worsened group performance.

Bahrami et al. (2010) reported that when two people with different abilities work cooperatively while interacting, their performance tends to be worse than when they work alone. In contrast, their performance improves when their abilities are equal. In this study, although the speakers' roles (i.e., driver and engineer) were different, as described in Section 3, they were equally capable of understanding the car's condition. Therefore, the verification results are consistent with those reported in a previous study (Bahrami et al., 2010).

There may be collaborative work in which performance is improved by conveying only objective facts (e.g., solving mathematics problems). Note that the scope of the present study was limited to research subjects that satisfy the four advantages described in Section 1.

8. Acknowledgements

We thank the anonymous reviewers for their insightful suggestions and discussions.

9. Bibliographical References

Bahador Bahrami, Karsten Olsen, Peter E Latham, Andreas Roepstorff, Geraint Rees, and Chris D Frith. 2010. Optimally interacting minds. *Science*, 329(5995):1081–1085.

Giulio Bernardi, Emiliano Ricciardi, Lorenzo Sani, Anna Gaglianese, Alessandra Papasogli, Riccardo Ceccarelli, Ferdinando Franzoni, Fabio Galetta, Gino Santoro, Rainer Goebel, et al. 2013. How skill expertise shapes the brain functional architecture: an fmri study of visuo-spatial and motor processing in professional racing-car and naïve drivers. *PloS one*, 8(10):e77764.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Denis Eka Cahyani and Irene Patasik. 2021. Performance comparison of tf-idf and word2vec models for emotion text classification. *Bulletin of Electrical Engineering and Informatics*, 10(5):2780–2788.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sicong Dong, Yin Zhong, and Chu-Ren Huang. 2018. How do non-tastes taste? a corpus-based study on chinese people' s perception of spicy and numbing food. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 25th Joint Workshop on Linguistics and Language Processing*.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Hugo Gonçalo Oliveira, Patrícia Ferreira, Daniel Martins, Catarina Silva, and Ana Alves. 2022. A brief survey of textual dialogue corpora. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1264–1274, Marseille, France. European Language Resources Association.

Marion E Hambrick, Jason M Simmons, Greg P Greenhalgh, and T Christopher Greenwell. 2010. Understanding professional athletes' use of twitter: A content analysis of athlete tweets. *International Journal of Sport Communication*, 3(4):454–471.

Taku Kudo. 2005. Mecab: Yet another part-of-speech and morphological analyzer. https://taku910.github.io/mecab. Accessed March 19, 2024.

Michael F Land and Benjamin W Tatler. 2001. Steering with the head: The visual strategy of a racing driver. *Current biology*, 11(15):1215–1220.

Otto Lappi. 2018. The racer's mind—how core perceptual-cognitive expertise is reflected in deliberate practice procedures in professional motorsport. *Frontiers in Psychology*, 9.

Otto Lappi. 2022. Egocentric chunking in the predictive brain: a cognitive basis of expert performance in high-speed sports. *Frontiers in Human Neuroscience*, 16:822887.

William F Milliken, Douglas L Milliken, and L Daniel Metz. 1995. *Race car vehicle dynamics*, volume 400. SAE international Warrendale.

Ryota Nishizono, Naoki Saijo, and Makio Kashino. 2023. Highly reproducible eyeblink timing during formula car driving. *iScience*, 26(6):106803.

Ann Pegoraro. 2010. Look who' s talking—athletes on twitter: A case study. *International journal of sport communication*, 3(4):501–514.

Michael B Reid. 2022. Redox implications of extreme task performance: the case in driver athletes. *Cells*, 11(5):899.

Michael B. Reid and Joshua T. Lightfoot. 2019. The physiology of auto racing: A brief review. *Medicine & Science in Sports & Exercise*.

Gerard Salton and Michael McGill. 1983. *Introduction to modern information retrieval*.

Toshinori Sato. 2015. Neologism dictionary based on the language resources on the web for mecab. https://github.com/neologd/mecab-ipadic-neologd. Accessed March 19, 2024.

Sayaka Tohyama and Hajime Shirouzu. 2018. Proposal for an assessment framework for collaborative problem solving skills. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society*, 24(4):494–517.

Peter M Van Leeuwen, Stefan De Groot, Riender Happee, and Joost CF De Winter. 2017. Differences between racing and non-racing drivers: A simulator study using eye-tracking. *PLoS one*, 12(11):e0186871.

Kielan Yarrow, Peter Brown, and John W Krakauer. 2009. Inside the brain of an elite athlete: the neural processes that support high achievement in sports. *Nature Reviews Neuroscience*, 10(8):585–596.

Yin Zhong and Chu-Ren Huang. 2018. Pleasing to the mouth of pleasant personality: A corpus-based study of conceptualization of desserts in online chinese food reviews. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 25th Joint Workshop on Linguistics and Language Processing*.

10. Language Resource References

Malhar Anjaria and Ram Mohana Reddy Guddeti. 2014. Influence factor based opinion mining of twitter data using supervised learning. In *2014 sixth international conference on communication systems and networks (COMSNETS)*, pages 1–8. IEEE.

Takuma Ichikawa and Ryuichiro Higashinaka. 2022. Analysis of dialogue in human-human collaboration in Minecraft. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4051–4059, Marseille, France. European Language Resources Association.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in Minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415, Florence, Italy. Association for Computational Linguistics.

Taro Okahisa, Ribeka Tanaka, Takashi Kodama, Yin Jou Huang, and Sadao Kurohashi. 2022. Constructing a culinary interview dialogue corpus with video conferencing tool. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3131–3139, Marseille, France. European Language Resources Association.

Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

Hanfei Sun, Ziyuan Cao, and Diyi Yang. 2022. SPORTSINTERVIEW: A large-scale sports interview benchmark for entity-centric dialogues. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5821–5828, Marseille, France. European Language Resources Association.

Bram Willemsen, Dmytro Kalpakchi, and Gabriel Skantze. 2022. Collecting visually-grounded dialogue with a game of sorts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2257–2268, Marseille, France. European Language Resources Association.

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Workshops at the thirty-second aaai conference on artificial intelligence*.