Multimodal Technologies and Interaction

MDPI

Article

Observing Collaboration in Small-Group Interaction

Maria Koutsombogera *,† **and Carl Vogel** †

Trinity Centre for Computing and Language Studies, Trinity College, the University of Dublin, Dublin 2, Ireland

* Correspondence: koutsomm@scss.tcd.ie

† These authors contributed equally to this work.

Received: 30 April 2019; Accepted: 24 June 2019; Published: 28 June 2019

Abstract: In this study, we define and test measures that capture aspects of collaboration in interaction within groups of three participants performing a task. The measures are constructed upon turn-taking and lexical features from a corpus of triadic task-based interactions, as well as upon demographic features, and personality, dominance, and satisfaction assessments related to the corpus participants. Those quantities were tested for significant effects and correlations they have with each other. The findings indicate that determinants of collaboration are located in measures quantifying the differences among dialogue participants in conversational mechanisms employed, such as number and frequency of words contributed, lexical repetitions, conversational dominance, and in psychological and sentiment variables, i.e., the participants' personality traits and expression of satisfaction.

Keywords: collaboration; group interaction; turn-taking; dominance; personality; lexical repetition

1. Introduction

Analysis of small-group interaction and group dynamics has been an important research topic because of the multitude of everyday professional and personal activities performed by small groups. From a computational perspective, in recent years there has been a significant amount of work aimed at developing methods that automatically analyze group interaction, and models able to predict information about the participants as well as the state of the interaction, based on both spoken words and nonverbal channels. Related work reports on identification of group dynamics [1–3], group performance [4–6] and participation styles [7], and inferring behavioral aspects in dialogue such as competitiveness [8,9], dominance and leadership [10–12], affect [13], cohesion [14] and personality [15]. The aforementioned works focus mainly on nonverbal and turn-taking cues, while a smaller amount includes the investigation of verbal cues, reporting the usefulness of linguistic features [4] and linguistic alignment (repetitions) [16] for predicting group performance.

Research in small-group interaction has been facilitated by multimodal corpora in two-party (HCRC Map Task Corpus [17]) and in multiparty settings, such as the AMI corpus [18], the Mission Survival Corpus-2 [15], the Canal 9 political debates database [19], the Idiap Wolf Corpus [9] and the GAP (Group Affect and Performance) corpus [20], corpora which investigate various aspects of group interactions.

Small-group interactions are considered pervasive and quintessential of collaborative work [3]. A crucial goal within task-based interactions is to establish cooperation, since task-oriented small groups are most effective when all group members can participate and be heard [21]. However, most of the existing approaches have not explicitly investigated aspects of collaboration nor considered collaboration as an achievement itself, independently of task success. Studies that build models for automatically predicting how well a group performs on a well-defined task focus on performance, and performance is synonymous to task success, i.e., defined on score terms. In these studies, group performance is clearly estimated within datasets with objective task scores [4–6].

In this respect, collaboration is a presupposition to task implementation and completion, but its aspects are not evaluated. Successful completion of a task implies that there has been interaction among task participants, and interaction entails collaboration, yet the level of collaboration is not assessed. Although task completion may be achieved with low collaboration levels, there are tasks (e.g., professional group meetings, learning tasks in an educational context) focusing on social aspects of the interaction, such as the development of social skills and cooperation, where maximized collaboration is more important than achieving a correct task score. For example, in task-based interactions where one of the participants is a dominant speaker and manages to go through the task correctly and quickly, the task may be successful (in the sense that all task items are correctly provided or ordered); however collaboration has not been successful, as the participants have not actually shared their ideas nor participated equally in the discussion.

In this work, we inspect the factors that we think are constituents of group collaboration in a corpus of triadic interactions. In each group session in the corpus two participants collaborate with each other to solve a quiz while assisted by a facilitator. Their task is to provide the three most popular answers to three quiz questions, according to the responses that external groups have provided. In this setup, collaboration refers to the process where the two players coordinate their actions to achieve their shared goal, i.e., find the appropriate answers to the quiz questions and rank them in terms of popularity, and not suitability.

In this dataset, similarly to other task-based interactions, participants share common ground, i.e., are aware of the goals of the interaction, they are equally competent to talk about the topic at hand and collaborate to produce responses to quiz questions, i.e., they share similar communicative goals. Although the task is short-term, participants build and maintain social bonds, as they are asked to collaborate. In informing prospective participants about the research and the multimodal corpus that we planned to construct for our own research and to release to the wider research community, we indicated, "this dataset collection will enable us to study the way people discuss and collaborate with each other to complete a task." Our purpose in the present work is to observe and report what participants do in response to a request that they discuss the task with each other, and collaborate. The task was constructed not as a competitive game, but one that required consensus responses and linguistic acts to achieve those. In this paper, we report on what participants were observed to do to achieve collaboration. This entails that we focus on several isolated qualities, quantifying operationalized measures of those qualities, and to some extent, relating them.

Contrary to other tasks reported in the small-group literature, correctness of answers in this game is mainly a matter of estimating popular responses, i.e., attunement to the popular opinion. Players guess and express options that can be reasonable possible answers to the specific question, but if those options are not among the 3 most popular answers according to the game database, then they keep trying until they achieve to identify the appropriate answers. In this respect, the criterion of correctness in this game is a subjective one. Nevertheless, players are not discouraged by this subjectivity, they rather seem intrigued to discover what the popular opinion is, and for that reason they may feel more motivated to collaborate with their co-player to discover the appropriate answers. Suggesting or agreeing on answers that are incorrect in this sense does not entail a knowledge gap for the participants; rather it entails a distinct perspective from the individuals polled. Therefore, learning that the answers put forward are not correct is expected to prime discussion about the locus of perspectival differences.

To observe how participants attempted to achieve collaboration we explore measurable turn-taking dialogue aspects, lexical features, and behavioral and psychological variables of group members to identify the quantities that comprise measuring aspects of collaboration. Turn-taking analysis is used to identify the level of balanced participation of the speakers, through the number of words and turns they use, their speaking rate, and the timings of their contributions. Lexical features considered important to aspects of collaboration are explored, i.e., word similarity between participants, lexical expressions of attunement, and personal pronouns.

Personality trait scores from both self and informants are used as tools to investigate effects of personality on aspects of participants' behavior that relate to collaboration. Asymmetry aspects are also considered using features of perceived dominance scores attributed to the participants. Perception of the facilitator's practices during the game by the quiz players are also taken into account, to examine whether the players' satisfaction levels affect the way they collaborate. We also take demographic aspects into account (i.e., gender, nationality and native language), as these have been proven to have effects on team creativity [22], as well as information related to the level of familiarity among the speakers, to examine the extent to which these features interpret the speakers' collaborative behavior. Finally, we will use the standard measures of task success, i.e., the task scores and the amount of time groups need to complete a task.

We have specific hypotheses about a range of quantities that we think capture aspects of collaboration, and we are conducting a naturalistic examination of the dialogues in light of these quantities, reporting where the effects are significant, by observing how participants talk and behave in the dialogues, without experimentally controlling a way in which they should collaborate, and without a gold standard assessment of collaboration by the participants themselves, or by external raters. We argue that by constructing measures of linguistic behaviors of individuals who have been asked to collaborate, these measures will reflect important aspects of collaboration and will result in the definition of determinants which ultimately interpret levels of collaboration among group participants.

Additionally to linguistic behaviors, we also study personality traits (using the "Big Five" model) assessed "internally", by participants using a standard inventory of preference assessments and "externally", by independent judges. We also analyze the relation of independent assessment of dominance of the participants in the conversations. We examine linguistic behaviors in relation to these psychological profiles. In some cases, we consider the profile variables of the individuals, and in other cases, we consider the balance of the value for the variable between participants in an interaction. For example, given that participants are collaborating, it is interesting to know what behaviors are visible when participants are rated via independent observation with the same level of dominance or where there is an imbalance of dominance.

The results show that there are quantities that have strong potential as determinants of collaboration, and these are located in measures quantifying the differences among dialogue participants in conversational mechanisms employed, such as number and frequency of words contributed, lexical repetitions, conversational dominance, and in psychological and sentiment variables, i.e., the participants; personality traits and expression of satisfaction. On the contrary, features such as task scores or demographics do not capture, as much as expected, aspects of collaboration.

The next sections describe the materials and methodology employed in this study, followed by the reporting and the discussion of results.

2. Materials and Methods

The materials used for this study involve the dataset of triadic dialogues, turn-taking and lexical features coming from the transcripts, features from surveys (personality and experience assessment), perceived dominance features, and "correct" answers to the task questions.

2.1. Dataset

The study presented here exploits the MULTISIMO corpus [23], a corpus which records 23 group triadic dialogues among individuals engaged in a playful task akin to the premise of the television game show, *Family Feud*. All subjects gave their informed consent to participate in the dialogue recordings. The recordings and the study were conducted in accordance with the EU General Data Protection Regulation 2016/679 (GDPR), and the protocol was approved by the Trinity College Dublin, School of Computer Science and Statistics Research Ethics Committee. Two of the participants in each dialogue were randomly partnered with each other, and the third participant serves as the

facilitator for the discussion. Individuals who participate as facilitators are involved in several such discussions (in total three facilitators), while the other participants participate in only one dialogue. Each dialogue has the same overall structure: an introduction, three successive instances of the game, and a closing. Each instance of the game consists of the facilitator presenting a question. The 3 questions were: (a) *name a public place where catching a cold or a flu bug is likely*; (b) *name 3 instruments you can find in a symphony orchestra*; and (c) *name something that people cut*. Each question was followed by a phase in which the participants have to propose and agree to three answers, and that followed in turn by a phase in which the participants must agree to a ranking of the three answers. The rankings are based on perceptions of what 100 randomly chosen people would propose as answers. This survey was conducted independently, and rankings are reported in a database related to the game (http://familyfeudfriends.arjdesigns.com/, last accessed 19.04.2019) and therefore the"correct" ranking is in the order of popularity determined by this independent ranking.

The sessions were carried out in English. Overall, 49 participants were recruited, and the pairing of players was randomly scheduled. 46 were assigned the role of players and were paired in 23 groups. The remaining 3 participants shared the role of the facilitator throughout the 23 sessions. While in other corpora the familiarity among group participants is a controlled variable [17], in our case there was no attempt to pair players based on whether they know each other or not. This was decided to ensure that the least number of constraints is imposed on the experimental design, allowing for more flexibility in the formulation of research questions. In most of the sessions the participants do not know each other, although there are a few cases (i.e., in four groups) where the players are either friends or colleagues. The average age of the participants is 30 years old (min = 19, max = 44). Furthermore, gender is balanced, i.e., with 25 female and 24 male participants, randomly mixed in pairs. The participants come from different countries and span 18 nationalities, one third of them being native English speakers. Tables 1 and 2 present the details about the gender of the participants, the number of groups whose players are familiar with each other or not, the number of native and non-native English speakers, as well as their different nationalities. Tables 1 and 2 were originally included in [23] , where the MULTISIMO corpus was introduced.

The players expressed and exchanged their personal opinions when discussing the answers, and they announced the facilitator the ranking once they reached a mutual decision. They were assisted by the facilitator who coordinated this discussion, i.e., provided the instructions of the game and confirmed participants' answers, but also helped participants throughout the session and encouraged them to collaborate. Facilitators were selected in advance and were briefly trained before the actual recordings, i.e., they were given the quiz questions and answers and they were instructed to monitor the flow of the discussion and, if necessary, intervene to help players or to balance their participation. Before the recordings the participants filled in a personality test (cf. Section 2.5). After the end of each session the participants filled in a brief questionnaire to assess their impression of the facilitator's behavior (cf. Section 2.6).

The corpus consists of synchronized audio and video recordings, and its overall duration is approximately 4 h, with an average session duration of 10 min.

Table 1. Distribution of corpus participants per gender, language, and familiarity among group players.

Gender		Language		Familiarity (Groups)	
Male	24	EN Native	16	Familiar	4
Female	25	EN Non-native	33	Non-familiar	19

Table 2. Distribution of participants' nationalities.

Participants' Nationalities			
Greek	13	Croatian	1
Irish	13	Egyptian	1
French	4	Italian	1
Brazilian	2	German	1
British	2	Kazach	1
Indian	2	Mexican	1
Pakistani	2	Romanian	1
American	1	Slovenian	1
Chinese	1	Thai	1

2.2. Turn-Taking Features

The corpus audio files were transcribed by 2 annotators using the `Transcriber` tool (http://trans.sourceforge.net/, last accessed 19.04.2019). The annotators listened to the audio files, segmented the speech in turns and transcribed the speakers' speech. Transcription thus includes the speakers' identification and the words they utter, as well as the marking of silence and overlapping talk. For each corpus speaker, the following turn-taking features were computed from the annotations:

- **Number of Words**: The total number of words uttered by a participant during the conversation.
- **Number of Words per Minute**: The number of words uttered by a participant on average over the course of one minute of his/her speaking time.
- **Total Number of Turns**: The cumulation of the number of turns a participant takes over the course of the session.
- **Average Turn Duration**: The speaking time per participant divided by the participant's number of turns to discover the average length of turns in the session.

Figure 1 presents a sample of an annotated file in the ELAN editor (https://tla.mpi.nl/tools/tla-tools/elan/, last accessed 19.04.2019) with information about speaking turns and speech transcription.

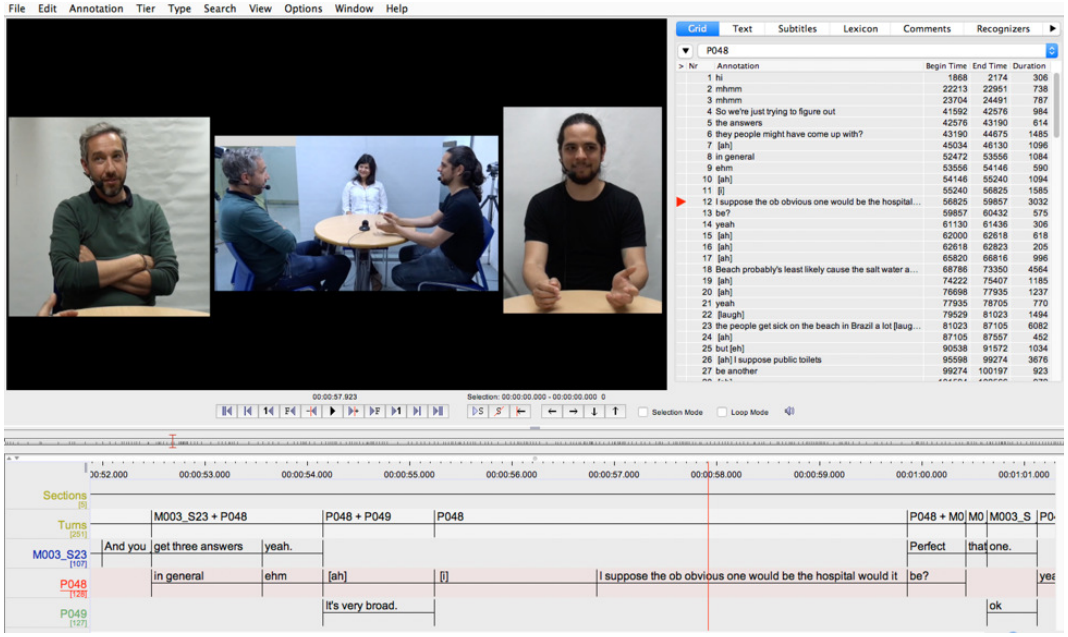

Figure 1. Screenshot of a sample file in ELAN, where transcript annotations for words and turns for each speaker may be viewed.

Although overlapping talk occurs in the data, and it can be automatically spotted, overlaps were not considered in this study, as they are not necessarily connected to collaboration. Simultaneous talk

may have a lot of functions, i.e., interruption, delayed completion, feedback expression, indication of transition relevant points, but also collaborative completions. Therefore, a more comprehensive study of overlaps (currently in progress) is needed to identify specific instantiations that are related to collaboration.

2.3. Word Similarity

Word similarity between participants was computed for each session. Specifically, we computed the similarity among words within turns of the three participants in the same dialogue, to measure other-repetition. The transcripts were preprocessed in that broken turns (because of overlaps) were fused, artificially fused items were separated, and transcripts were represented in the vector space. To calculate how dissimilar the three participants' word sets were, we computed the unweighted Jaccard distance as follows:

$$d_J(words_{P1}, words_{P2}) = 1 - \frac{|words_{P1} \cap words_{P2}|}{|words_{P1} \cup words_{P2}|} \quad (1)$$

That is, 1 minus the number of wordtypes used by pairs of participants, divided by word types used by either participant within a pair. The measure ranges between 0 (no distance) and 1, a score towards 0 indicating a lot of other-repetitions, and a score towards 1 indicating few other-repetitions.

The Jaccard distance between participants is consistent in weighting all items equally and was computed for each session.

In parallel, the Jaccard distance to self was computed. In this case, the word sets compared are those of turns owned by the same participant, hence there is a focus on self-repetition, as opposed to other-repetition, measured by the distance among distinct participants. In this respect, if a person's Jaccard distance to his/her own contributions approaches 0, then that person is engaged in a lot of self-repetition, and if it approaches 1, that person is engaged in a little self-repetition.

2.4. Dominance Assessment

Perceived dominance levels of the players involved in the dialogues was assessed through a perception experiment, where external raters provided their scores after watching the dialogue videos. The implementation of this experiment received too the approval of the School of Computer Science and Statistics Research Ethics Committee. Five annotators were recruited, and were asked to watch the corpus videos twice, observe the behavior of participants and, for each video file, rate the two quiz players (excluding the facilitators) on a scale from 5 (highest) to 1 (lowest), depending on how dominant they think that the video speakers are. The annotators spent up to one hour per day on the task and completed the ratings after watching the full videos. Raters were given concise guidelines that included a definition of conversational dominance as "a person's tendency to control the behavior of others when interacting with them".

The intraclass correlation coefficient, calculated with the presumption of random row (subject) and column (annotator) effects (ICC(A,5) = 0.776) is significant ($F(35, 24.2) = 6.68$, $p = 3.57 \times 10^{-06}$), suggesting a reasonable level of agreement among annotators, a level sufficient to warrant use of the aggregation of annotations (using the median annotation score for each individual) as a response variable, even though annotation disagreements exist. Using this dataset, [24] report the association of high dominance scores to the large number of words a speaker utters per minute, as well as to high extroversion and high Openness personality traits, and the association of low dominance scores with high Agreeableness.

2.5. Personality Traits

Personality variables are an important tool for the interpretation of social behavior. The related literature widely acknowledges the necessity to have an accepted classification scheme to categorize

empirical findings and that the 5-factor model is a robust and meaningful framework enabling the formulation and testing of hypotheses related to individual differences in personality [25]. To have an objective estimation of the participants' personality traits and investigate the effect of their personality on their collaborative behavior and task success, we opted for the Big Five personality inventories that assess the OCEAN traits: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Before the recordings, participants completed the Big Five Inventory (BFI-44), a self-report inventory designed to measure the Big Five traits [26,27]. The test consists of 44 items (statements) and the participants rated each statement to indicate the extent to which they agree or disagree with it. As a result, a list of scores per personality trait and per participant was created. The percentile rank of each participant across the five personality traits was then calculated using local norms (i.e., upon the groups population). Although dialogue participants were engaged in this self-assessment of OCEAN traits before the session recordings, partner matching did not take the test scores into account.

In addition to the self-reports, after the recordings a separate perception experiment was run with the aim to collect ratings from eight independent informants, who listened to audio clips of 36 corpus participants and responded to a 10-scale questionnaire including statements about the participants' personalities, i.e., the BFI-10 questionnaire [28]. The implementation of this experiment received too the approval of the School of Computer Science and Statistics Research Ethics Committee. The BFI-10 is a 10-item scale measuring the OCEAN personality traits and is an abbreviated version of the BFI-44 scale employed in the self-reports; it was designed for contexts in which respondents' time is limited. It has been reported that the BFI-10 items predict almost 70% of the variance of the BFI-44 scales, with a Fisher-Z-corrected overall mean correlation between the dimensions of the BFI-10 and those of the BFI-44 of r = 0.83 [28]. This experiment was implemented as a remote survey in the Qualtrics platform (https://www.qualtrics.com/, last accessed 19.04.2019). Each survey item consisted of an audio clip followed by the BFI-10 questionnaire; the informants were asked to rate each of the 10 statements to indicate the extent to which they agree or disagree with them. The audio clips ranged from 4 to 10 s and each clip included speech content from a single corpus participant. A list of scores per personality trait and per clip was created. Since more than one clip corresponded to each speaker, the mean score per speaker per trait was computed. Again, percentile ranks across the five personality traits were computed based on local norms.

2.6. Experience Assessment

After the recordings, participants that had the role of player were asked to fill in an experience assessment questionnaire (EAQ). Through this questionnaire participants expressed their opinion about the interaction they had with their facilitator and their satisfaction levels with the facilitator's characteristics. The EAQ consisted of 17 pairs of contrasting characteristics that may apply to a certain degree to the facilitator, e.g., helpful vs. unhelpful, polite vs. impolite, or supportive vs. obstructive, and was based on similar work related to user experience [29]. The full list of the questionnaire items can be found in the Appendix A. For each of the 17 statements, participants ticked a number on a seven-point scale.

2.7. Ranking Scores

As mentioned earlier in Section 2.1, rankings are based on perceptions of what 100 randomly chosen people answered and were recorded in the game database. In this respect, we argue that correct scores in this dataset measure the group's attunement to popular opinion: participants attempt to guess and rank items that have most probably been identified as important by most people.

Participants first identified and agreed on the set of correct answers, and then proceeded to their ranking. A ranking score was then computed for each session, following three approaches.

In the first approach, a 3-point scoring system is used: if participants provide correct answer ranking to one question, then they score 1 point, if they correctly rank answers to a second question

they score 2 points, and if they correctly rank answers for all questions, they achieve the best score, i.e., 3 points. Thus, a score for each session is the sum of points achieved for all three questions in the session. This approach is the strictest one, in that it requires that for a given question, all items must be correctly ranked. It thus marks as 0 cases where one answer is appropriately ranked, while the other two are not.

The second approach calculates scores on a 9-item scale: for each session, the participants' ranking of 9 answers (3 answers for each of the 3 questions) is compared to the most popular rankings; 1 point is scored each time the ranking of an answer matches the ranking provided by the 100 people in the game database.

The third approach calculates correctness based on the relative order of the ranked answers in the list supplied by the participants and by the independently surveyed responses. We count the difference between ranked pairs that agree (concordant pairs) and pairs where the lists show disagreement (discordant pairs) and divide this by the total number of pairs that could agree. This is exactly the formulation of Kendall's rank correlation coefficient. Thus, we compute Kendall's τ for the ranked lists supplied for each of the three questions within each session and use the arithmetic mean of these as the success score for the session.

By testing the correlation among the scores that are the outcome of the three approaches, we noted that the third approach, the mean τ score, ranks the 23 sessions with a rank ordering that is very similar to the order derived from both the first ($\tau = 0.969$, $p = 4.064e-09$) and the second ($\tau = 0.848$, $p = 1.113e-06$) approach. Similarly, there is an almost perfect correlation among the scores of the first and the second approach ($\rho = 0.939$, $p = 3.556e-11$).

All three approaches to the ranking score were used in the experiments that followed, wherever ranking score is involved (cf. Section 3); however, by observing the high rank similarity among the different approaches, we do not anticipate noteworthy differences among measurements that use either of those approaches.

3. Methodology

To investigate the quantification of aspects of collaboration, we assumed that there are rational measures of collaborative communication which show whether there is a balance in the presence of certain features among the dialogue participants. We therefore constructed such measures of collaborative communication based on the various feature sets and we tested the relationship of these measures with other variables, as well as the mutual relationship of these measures. We consider that a critical concept is that of *balance* among the participants in the various aspects of collaboration examined. Below are definitions of different types of balance variables that have been computed for each session:

- **WordBalance**: the absolute value of the difference in the number of words produced by the 2 players, divided by the sum of words from both players.
- **TurnBalance**: the absolute value of the difference in the number of turns contributed by the 2 players, divided by the sum of turns from both players
- **SpeedBalance**: the absolute value of the difference in the number of words produced by the 2 players, divided by the dialogue duration.
- **TimeBalance**: the absolute value of the difference in the number of turns contributed by the 2 players, divided by the dialogue duration.
- **WpMBalance**: the absolute value of the difference in the speaking rate of the 2 players, i.e., in the number of words per minute that they produce within their speaking time.
- **DomBalance**: the absolute value of the difference of the median dominance scores attributed to the 2 players.
- **EAQBalance**: the absolute value of the difference of the scores assessing each of the facilitator's characteristics by the 2 players in the experience assessment questionnaires.

- **Personality-Self-Balance**: the absolute value of the difference of each of the percentile scores per trait provided by the players in their self-assessed personality tests.

Furthermore, we defined variables descriptive of the participants' demographic data, i.e.,

- **GenderMatch** defines whether the gender of the 2 players is the same.
- **NationMatch** defines whether the nationality of the 2 players is the same.
- **OneEnglish** defines whether player pairs in each session include one native English speaker.
- **AllEnglish** defines whether both player in each session are native English speakers.

Further variables that were constructed are the following:

- **RankingScore**: number of correctly ranked answers provided in a session, computed by either of the 3 approaches listed in Section 2.7.
- **TaskSuccess**: number of correctly ranked answers provided in a session (i.e., ranking score) per the overall session duration; the higher this quantity is, the more successful is considered the task.
- **Familiarity** defines whether players in a group are familiar with each other.
- **PronounSg**: number of personal pronouns in singular (*I, me, my, mine, myself*), divided by the total word count.
- **PronounPl**: number of personal pronouns in plural (*we, our, us*), divided by the total word count.
- **Attunement**: To approach attunement to popular opinion from a lexical perspective, we tracked the mentions of the word *people* in the 23 dialogues, and the distribution of this word per speaker role (facilitator and player), relativized by the total turn count. We assume that by using this word facilitators aim to stress the way ranking should be performed (i.e., according to popular, not personal opinion), and the players themselves to justify their choices and to establish a common ground with their co-player.
- **pJaccard**: the mean Jaccard distance computed between turns of distinct participants in the same dialogue, taking into account only turns where one participant is compared to another participant (other-repetition).
- **sJaccard**: the mean Jaccard distance computed between the turns of the same participant (self-repetition).
- **Personality-Self-Level**: binary classification of self-assessed personality scores (i.e., high, low) treated as an ordinal variable, defined as the sum of scores of the two players for each trait in each session.
- **Personality-Other-Level**: binary classification of other-assessed personality scores (i.e., high, low) treated as an ordinal variable, defined as the sum of scores of the two players for each trait in each session.

Correlation tests (Spearman's rank correlation) were then performed to compute the extent of correlations among the aforementioned quantities that approach the balance aspects in dialogue participant features, such as e.g., the relationship between the balance of dominance scores and balance of turns, to investigate which quantity best capture aspects of collaboration. Also, other non-parametric tests (Wilcoxon rank sum test, Kruskal-Wallis rank sum test) were performed to find significant differences between collaboration measures and variables related to demographic features, lexical features and other dialogue features, as described above. All tests were run in the R environment [30]. Results are reported in the next section.

4. Results

Tests between measures of personality trait scores (classes) assessed by informants and other quantities show significant correlations. The lower the word imbalance among participants is, the higher the participants score in both levels of Extraversion and Conscientiousness

(negative correlations of $\rho = -0.809$, $p < 0.003$, and $\rho = -0.905$, $p < 0.0001$ respectively). On the contrary, the higher the word imbalance, the higher participants score in Neuroticism levels (positive correlation of $\rho = 0.739$, $p < 0.009$). Neuroticism is positively correlated with DomBalance ($\rho = 0.685$, $p < 0.019$), showing that the greater the differences in dominance median scores are, the higher the Neuroticism levels are. Furthermore, as the ratio of words contributed per the duration of dialogue decreases, higher scores are achieved in Conscientiousness (i.e., negative correlation of Conscientiousness with SpeedBalance, $\rho = -0.653$, $p < 0.029$) Finally, the higher the imbalance is in the ratings of the participants in assessing *Confusing* aspects of the facilitator in the EAQ questionnaire, the higher are their Openness levels (positive correlation of $\rho = 0.665$, $p < 0.026$).

Correlation tests with self-assessed personality traits indicate a very good positive correlation of Neuroticism with DomBalance ($\rho = 0.718$, $p < 0.001$), showing that the greater the differences in dominance median scores are, the higher the Neuroticism levels are, as is also the case with informant-assessed Neuroticism. Finally, the more participants agree on their ratings in assessing *Impatient* aspects of the facilitator in the EAQ questionnaire, the higher are their Conscientiousness levels (negative correlation of $\rho = -0.499$, $p < 0.015$).

In addition to self- and informant-assessed personality classes, distinctness of perceptions of the participants, regarding the facilitator, as expressed in the EAQ questionnaire fit well together with the measure of SpeedBalance (balance in frequency, i.e., words contributed per the duration of dialogue); where the SpeedBalance measure is balanced, participants tend to disagree more in their ratings of *Obstructive* and *Impoliteness* aspects of the facilitator, as shown from negative correlations of $\rho = -0.542$, $p < 0.007$, and $\rho = -0.413$, $p < 0.050$ respectively.

Lexical tests of collaboration were also employed by examining the correlations between the ratio of singular and plural personal pronouns (cf. *PronounSg*, *PronounPl*) per total word count, and other measures. The related results show that the more singular personal pronouns the participants utter, the higher they are assessed on the Conscientiousness level (positive correlation between the *PronounSg* and informant-assessed Conscientiousness classes, $\rho = 0.613$, $p < 0.045$). A negative correlation was found among singular personal pronouns and self-assessed Agreeableness classes ($\rho = -0.523$, $p < 0.010$), showing that the more agreeable the participants consider themselves, the less pronouns they express. Tests about personal pronouns and the TaskSuccess measure present no correlation with pronouns in singular, and a negative correlation of $\rho = -0.382$ when pronouns in plural are involved, i.e., the relativized ranking score decreases when pronouns in plural increase, however, this correlation does not hit significance ($p = 0.072$).

Correlation tests exploring the Jaccard distance between turns of distinct participants within a dialogue (*pJaccard*) show that where dominance imbalance is high, so is Jaccard distance (a positive correlation of $\rho = 0.501$, $p < 0.034$); that is, the higher the difference between dominance scores among participants, the less likely it is for them to engage in other-repetitions. This holds true for players repeating their co-player, and players repeating the facilitators. This is supported by the fact that repetitions of other players and repetition of facilitators are closely correlated ($\rho = -0.782$, $p = 1.523e-05$). While tests among *pJaccard* and other measures, such as SpeedBalance, WordBalance, TurnBalance and TaskSuccess did not yield significant correlation results, *pJaccard* seems to fit well with the self-assessed Extraversion ($\rho = 0.430$, $p < 0.040$) and Openness ($\rho = 0.469$, $p < 0.024$) personality traits; where there are higher differences in the percentile scores of Extraversion and Openness among the participants, the Jaccard distance is also higher, hence the other-repetitions fewer. Other-repetitions among players and facilitator are related to differences in ratings of facilitators by the players; the higher the other-repetitions are, the more participants tend to disagree in their ratings of *Obstruction* ($\rho = -0.432$, $p = 0.040$), *Impoliteness* ($\rho = -0.492$, $p = 0.017$) and *Unhelpfulness* ($\rho = -0.430$, $p < 0.041$).

As regards self-repetitions, tests involving the *sJaccard* measure, i.e., distance between turns of the same participant, show that there is a significant positive correlation of self-assessed Openness percentiles with *sJaccard* ($\rho = 0.448$, $p < 0.032$), indicating that the higher the imbalance in Openness

is, the more the distance from self increases in the dialogue pairs, i.e., there are fewer self-repetitions. *sJaccard* also shows relatively good correlations with Extraversion and Agreeableness, but with no significance ($\rho = 0.379$, $p = 0.074$ and $\rho = 0.343$, $p = 0.109$ respectively). In addition, distance to self shows a non-significant positive correlation with DomBalance, and no correlations with the measures of SpeedBalance, WordBalance, and TurnBalance. Interestingly, self-repetition in facilitators relates to the degree of difference ratings of facilitators by the players, i.e., the more the facilitators engage in self-repetitions, the more participants tend to disagree in their ratings of *Obstructive* and *Impoliteness* aspects of the facilitator, as shown from negative correlations of $\rho = -0.405$, $p = 0.055$, and $\rho = -0.460$, $p < 0.027$ respectively.

Tests examining the relation of players' lexical expressions of attunement to popular opinion (i.e., mentions of the word *people* per turn) with the correct ranking score, computed by all three approaches, show an inverse relation, i.e., the more players mention *people* in their turns, the less good they are ranking items, i.e., at knowing what other people might think (3-point approach: $\rho = -0.493$, $p < 0.017$; 9-item approach: $\rho = -0.442$, $p < 0.035$). On the other hand, mentions of *people* correlate positively with the ratio of first person plural pronouns ($\rho = -0.561$, $p < 0.006$), i.e., with the participants thinking as a group.

To examine standard task success metrics, we introduced the variable TaskSuccess (session ranking score per overall session duration) and tested it in all its variations, i.e., including each of the three approaches of calculation the ranking score. TaskSuccess was tested with all the turn-taking-related quantities. The results show that there are good correlations with TurnBalance, varying from $\rho = 0.362$ to $\rho = 0.433$, depending on the scoring approach, but the only significant correlation holds with the first scoring approach (3-point scoring system), i.e., $\rho = 0.433$, $p < 0.04$. On the other hand, TaskSuccess shows no correlation with TimeBalance (all correlations are below $\rho = 0.06$), and very low and non-significant correlations with the WordBalance (all correlations below $\rho = 0.13$) and the SpeedBalance quantities (all correlations below $\rho = 0.18$).

Related Mann-Whitney-Wilcoxon tests looking for differences in TaskSuccess between groups whose players are of the same—and those whose players are of different gender, nationality, and nativeness, show non-significant results for the demographic variables. The same holds for TaskSuccess between familiar and unfamiliar pairs of participants, although a directed hypothesis that the measure of TaskSuccess is maximized with familiarity with significance, independently of the approach followed to compute the ranking score.

Familiarity among participants seems to be an informative feature regarding aspects of collaboration. To test if there are significant differences in WordBalance (words contributed), SpeedBalalnce (words per minute) and TimeBalance (time owning the floor) between groups whose players are familiar with each other, and groups where players do not know each other we used the Mann-Whitney-Wilcoxon test. No significant difference was found for WordBalance ($W = 17$, $p = 0.097$) and TimeBalance ($W = 29$, $p = 0.50$), while differences in SpeedBalance between familiar and unfamiliar pairs of participants were found significant ($W = 5$, $p < 0.039$). In addition, there is a significant effect of Familiarity in interaction with the DomBalance measure, shown by a significant difference in DomBalalnce between familiar and unfamiliar pairs of participants ($W = 7.5$, $p < 0.039$).

On the contrary, no significant effects of gender matching ($W = 21$, $p = 0.10$) or nativeness matching ($W = 55$, $p = 0.12$) were observed with respect to the DomBalance. Our hypotheses related to dominance were formulated around assumptions that the greatest difference in dominance balance will coincide with the greatest difference in words uttered (WordBalance) or turns owned (TurnBalance), as well as with frequency measures, i.e., SpeedBalance (words per minute) and TimeBalalnce (turns per minute). The related tests exhibit indeed that when dominance imbalance is high, so are word imbalance and speed imbalance, as shown from good correlations of DomBalance with WordBalance ($\rho = 0.557$, $p < 0.016$) and SpeedBalance ($\rho = 0.501$, $p < 0.034$); however our hypotheses were not confirmed as regards turn quantity and frequency aspects. This result is in line with findings of [24],

a study which analyzed dominance in the MULTISIMO corpus and reports on the association of high dominance scores to the large number of words a speaker utters per minute.

The current article builds upon this previous study in that, with respect to the examination of dominance as a collaboration determinant, it uses the perceived dominance median scores to construct the balanced quantity of dominance, and tests this new quantity with respect to other quantities. Thus, rate-related features such as words and turns per minute in the present study are also constructed as balanced quantities. Furthermore, the current study takes into account externally assessed personality scores, in addition to self-reported ones, as well as self- and other-repetitions. Finally, contrary to findings of [24] that the features of Extraversion, Openness, and Agreeableness are good predictors of dominance median scores, dominance quantities in the current study do not seem to correlate significantly with those personality quantities; instead, they correlate well with both self- and informant-assessed Neuroticism quantities.

Table 3 lists the measures that have significant mutual correlation.

Table 3. Significant correlations among measures of collaboration. The first column includes the measures tested, the second column reports the Spearman's rank correlation coefficient (ρ) value and the third column the p-values.

Test	Correlation (ρ)	Significance (p-Value)
WordBalance and DomBalance	0.557	0.016
SpeedBalance and DomBalance	0.501	0.034
WordBalance and informant-assessed Extraversion classes	−0.809	0.003
WordBalance and informant-assessed Neuroticism classes	0.739	0.009
WordBalance and informant-assessed Conscientiousness classes	−0.905	0.0001
TurnBalance and TaskSuccess	0.433	0.039
DomBalance and self-assessed Neuroticism classes	0.718	0.001
DomBalance and informant-assessed Neuroticism classes	0.685	0.019
SpeedBalance and informant-assessed Conscientiousness classes	−0.653	0.029
SpeedBalance and Obstructive score balance	−0.542	0.007
SpeedBalance and Impolite score balance	−0.413	0.050
Confusing score balance and informant-assessed Openness classes	0.665	0.026
Impatient score balance and self-assessed Conscientiousness classes	−0.499	0.015
Personal pronouns singular (ratio) and informant-assessed Conscientiousness classes	0.613	0.045
Personal pronouns singular (ratio) and self-assessed Agreeableness classes	−0.523	0.010
Personal pronouns plural (ratio) and Attunement	0.561	0.006
Attunement and Ranking Score	−0.493	0.017
DomBalance and pJaccard	0.501	0.034
ExtraBalance and pJaccard	0.430	0.040
OpenBalance and pJaccard	0.469	0.024
OpenBalance and sJaccard	0.448	0.032
sJaccard and Impolite score balance	−0.460	0.027
pJaccard and Obstructive score balance	−0.432	0.040
sJaccard and Impolite score balance	−0.492	0.017
sJaccard and Unhelpful score balance	−0.430	0.041

5. Discussion

In this work we expressed specific hypotheses about a range of quantities that we think capture aspects of collaboration, and we conducted an examination of the dialogues in light of these quantities, reporting where the effects are significant. The construction of those quantities was based upon the dialogues' turn-taking and lexical features, demographic features, and personality, dominance, and satisfaction assessments related to the corpus participants. An alternative way to assess aspects of collaboration would be to have a gold standard measurement of collaboration through independent assessments of the conversations by external raters, or through assessments of the interactions by the participants themselves. In the first case, external raters would evaluate the degree of collaboration as a holistic measure of the dialogue, in a similar way that external assessments of dominance and personality to individual participants were performed. Our intuition is that external raters would still be making judgements on observable features in the individual speakers, therefore we opted for directly assessing this pool of features. As regards the second case, it would be extremely informative

to have assessments of collaboration by participants themselves; however, this was not included in the experimental design.

From a methodological perspective, the aim of this work was to construct blocks of a theory of collaboration (as opposed to having a gold standard of collaboration scores), based on features related to language, turn-taking, and self-assessed and perceived psychological variables. We thus exploit the linguistic content, the structure of the dialogues, as well as the multimodal behavior of the dialogue participants, given that perception assessments are performed based on audio and visual features. It is by no means an exhaustive study of all multimodal features in a conversation and their potential to serve as collaboration factors. As the corpus will keep being populated with more manual and automatic annotations, we plan to quantify factors related to gaze, gestures, and head-nodding features. We argue that the quantities constructed and described in this work, a summary of which is reported below, reflect aspects of collaboration, and contribute to developing a theory of collaboration that interprets levels of collaboration among group participants.

Differences among participants in their contributions (number of words, i.e., WordBalance) appear to constitute a measure that fits well with differences in dominance levels, as well as differences in externally assessed personality traits, leading to the assumption that the quantity of words contributed affects the perception of informants when assessing dialogue participants. We explored two quantities considering dialogue duration, i.e., TaskSuccess (correct answers score per dialogue duration) and SpeedBalance (difference in the number of words per dialogue duration). Collaboration is approached by maximizing the first measure, and minimizing the second one. The results indicate that SpeedBalance shows significant correlations with other quantities (i.e., differences in contributions, dominance levels, personality trait levels, satisfaction assessment scores), while TaskSuccess does not, thus leading in the assumption that correctness of answers is less relevant to collaboration than balance of contributions among participants.

Conversational dominance is an aspect that is by definition linked to collaboration, in the sense that participants may wish to control or dominate the other participants' behavior, resulting in asymmetry in participation, and thus in collaboration. Our findings suggest that imbalance in perceived dominance scores can effectively determine aspects of collaboration. Except for the rational assumptions that the greatest difference in dominance will coincide with the greatest difference in words and word frequency, which were confirmed, dominance differences were also well correlated with self and externally assessed personality traits, as well as with lexical other-repetitions. Lexical repetition was investigated among participants within a dialogue (other-repetition) and within participants in a dialogue pair (self-repetition) by measuring the mean Jaccard distance among participant turns, and the mean Jaccard self-distance, respectively. Repetition is strongly connected to the process of developing common ground and can also signal engagement or involvement in an interaction, where involvement is defined as the action to achieve mutual understanding. The results in this study show that although the mean Jaccard distances, i.e., repetitions, do not correlate with collaboration measures related to turn-taking balance aspects, other-repetitions are associated with asymmetry aspects in the dialogue (dominance), to personality aspects of the participants (Extraversion and Openness traits), and to differences in the perception of the facilitator, while self-repetitions are associated with Openness and to perceptions of the facilitator as well. Results showed that externally assessed personality traits are efficient as measures of collaboration. As mentioned in Section 2.5, external assessment of personality was performed on thin audio slices. The results support the face validity of the thin-slice approach [31], that people form impressions of humans very quickly and on very little data, and that one can discover observables in thin slices of the data that provide evidence for what triggers the impressions.

Differences in the perceptions of the participants in assessing the facilitators, i.e., in expressing their satisfaction from collaborating and being guided by the facilitator in the EAQ questionnaire, emerge where there are differences in contribution frequency (words contributed per the duration of dialogue), but also when participants present differences in the personality traits of Openness

and Conscientiousness. Given that significant correlations involving this questionnaire were found, we acknowledge that additional survey materials allowing for peer assessment would be important to construct and include in studies of collaboration, to capture the players' perceptions of each other and be further investigated in addition to other features related to the players' performance, linguistic, and conversational features. In terms of group composition and demographic data of the participants, tests involving gender matching, native language or nationality matching did not yield significant results.

A hypothesis that was not confirmed, was that a higher frequency of lexical expressions of attunement would lead to higher ranking scores, in the sense that players need to answer what most people think as opposed to what they think themselves. The results show that the more frequent these lexical expressions are, the lower the score is, i.e., the worse players are at knowing what other people might think. There are two possible interpretations of this result. First, when participants mention *people*, they seemingly reason as if they are not among people themselves, i.e., that other people might be different from themselves; a second interpretation is that if participants are truly attuned with each other, they do not need to reason about what other people say. In this work, we argue that attunement is a component of collaboration: correct scores in the dialogue actually measure the group's attunement to popular opinion, i.e., to groups outside the conversation. On the other hand, attunement with the people inside the conversation is reflected by the balance quantities we introduced, i.e., expressing the balance of input or alignment of linguistic contributions. The fact that TaskSuccess (ranking scores relativized per task duration) correlates poorly with balance measures, except for TurnBalance, and thus lacks correlation with group-internal attunement, provides an empirical demonstration that these components of collaboration are independent of each other.

The results show that the data has some aspects that conform to our hypotheses. We argue that the measures which have face validity and show correlation with each other are capturing elements of collaboration, and the fact that they do not show perfect correlation indicates a corresponding degree of independence, demonstrating that collaboration is a multi-dimensional construct.

Author Contributions: Conceptualization, M.K. and C.V.; methodology, M.K. and C.V.; software, C.V.; validation, M.K. and C.V.; formal analysis, C.V. and M.K.; investigation, M.K. and C.V.; data curation, M.K.; writing–original draft preparation, M.K.; writing–review and editing, M.K. and C.V.; visualization, M.K. and C.V.; supervision, M.K. and C.V.; project administration, M.K.; funding acquisition, M.K. and C.V.

Funding: The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 701621 (MULTISIMO).

Acknowledgments: The authors would like to thank Rachel Costello for her contribution in the experimentation related to conversational dominance.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

EAQ	Experience Assessment Questionnaire
BFI	Big Five Inventory
OCEAN	Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism

Appendix A

This Appendix lists the set of the 17 items of the Experience Assessment Questionnaire (cf. Section 2.6) that the dialogue players filled in to assess their interaction with the facilitator. Each item consists of a pair of contrasting characteristics, and participants ticked a number on a seven-point scale to grade their impressions of the facilitators' characteristics.

Annoying	vs.	Enjoyable
Not understandable	vs.	Understandable
Unfriendly	vs.	Friendly
Obstructive	vs.	Supportive
Impolite	vs.	Polite
Unhelpful	vs.	Helpful
Inefficient	vs.	Efficient
Confusing	vs.	Clear
Impatient	vs.	Patient
Not motivating	vs.	Motivating
No feedback	vs.	Feedback
Does not meet expectations	vs.	Meets expectations
Not consistent	vs.	Consistent
Does not listen	vs.	Carefully listens
Does not check my understanding	vs.	Checks my understanding
Does not hold our attention	vs.	Holds our attention
Does not know what to do	vs.	Knows what to do

References

1. Gatica-Perez, D.; McCowan, L.; Bengio, S. Detecting group interest-level in meetings. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, PA, USA, 18–23 March 2005; Volume 1, pp. 489–492. [CrossRef]
2. Lai, C.; Carletta, J.; Renals, S. Detecting Summarization Hot Spots in Meetings Using Group Level Involvement and Turn-Taking Features. In Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013), Lyon, France, 25–29 August 2013; pp. 2723–2727.
3. Gatica-Perez, D.; Aran, O.; Jayagopi, D. Analysis of Small Groups. In *Social Signal Processing*; Burgoon, J.K., Magnenat-Thalmann, N., Pantic, M., Vinciarelli, A., Eds.; Cambridge University Press: Cambridge, UK, 2017; pp. 349–367, doi:10.1017/9781316676202.025.
4. Murray, G.; Oertel, C. Predicting Group Performance in Task-Based Interaction. In Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI '18), Boulder, CO, USA, 16–20 October 2018; ACM: New York, NY, USA, 2018; pp. 14–20. [CrossRef]
5. Avci, U.; Aran, O. Predicting the Performance in Decision-Making Tasks: From Individual Cues to Group Interaction. *IEEE Trans. Multimed.* **2016**, *18*, 643–658. [CrossRef]
6. Dong, W.; Lepri, B.; Kim, T.; Pianesi, F.; Pentland, A.S. Modeling conversational dynamics and performance in a Social Dilemma task. In Proceedings of the 5th International Symposium on Communications, Control and Signal Processing, Rome, Italy, 2–4 May 2012; pp. 1–4. [CrossRef]
7. Nakano, Y.I.; Nihonyanagi, S.; Takase, Y.; Hayashi, Y.; Okada, S. Predicting Participation Styles Using Co-occurrence Patterns of Nonverbal Behaviors in Collaborative Learning. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15), Seattle, WA, USA, 9–13 November 2015; ACM: New York, NY, USA, 2015; pp. 91–98. [CrossRef]
8. Chowdhury, S.A.; Riccardi, G. A Deep Learning approach to modeling competitiveness in spoken conversations. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 5680–5684. [CrossRef]
9. Hung, H.; Chittaranjan, G. The Idiap Wolf Corpus: Exploring Group Behaviour in a Competitive Role-playing Game. In Proceedings of the 18th ACM International Conference on Multimedia (MM '10), Firenze, Italy, 25–29 October 2010; ACM: New York, NY, USA, 2010; pp. 879–882. [CrossRef]
10. Jayagopi, D.; Hung, H.; Yeo, C.; Gatica-Perez, D. Modeling dominance in group conversations from non-verbal activity cues. *IEEE Trans. Audio Speech Lang. Process.* **2009**, *17*, 501–513. [CrossRef]
11. Nakano, Y.; Fukuhara, Y. Estimating Conversational Dominance in Multiparty Interaction. In Proceedings of the 14th ACM International Conference on Multimodal Interaction (ICMI '12), Santa Monica, CA, USA, 22–26 October 2012; ACM: New York, NY, USA, 2012; pp. 77–84. [CrossRef]
12. Sanchez-Cortes, D.; Aran, O.; Mast, M.S.; Gatica-Perez, D. A Nonverbal Behavior Approach to Identify Emergent Leaders in Small Groups. *IEEE Trans. Multimed.* **2012**, *14*, 816–832. [CrossRef]

13. Tian, L.; Moore, J.D.; Lai, C. Recognizing Emotions in Spoken Dialogue with Acoustic and Lexical Cues. In Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents (ISIAA 2017), Glasgow, UK, 13–17 November 2017; ACM: New York, NY, USA, 2017; pp. 45–46. [CrossRef]
14. Hung, H.; Gatica-Perez, D. Estimating Cohesion in Small Groups Using Audio-Visual Nonverbal Behavior. *IEEE Trans. Multimed.* **2010**, *12*, 563–575. [CrossRef]
15. Mana, N.; Lepri, B.; Chippendale, P.; Cappelletti, A.; Pianesi, F.; Svaizer, P.; Zancanaro, M. Multimodal Corpus of Multi-party Meetings for Automatic Social Behavior Analysis and Personality Traits Detection. In Proceedings of the 2007 Workshop on Tagging, Mining and Retrieval of Human Related Activity Information (TMR '07), Nagoya, Japan, 15 November 2007; ACM: New York, NY, USA, 2007; pp. 9–14. [CrossRef]
16. Reitter, D.; Moore, J.D. Predicting Success in Dialogue. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007), Prague, Czech Republic, 23–30 June 2007; Association for Computational Linguistics: Prague, Czech Republic, 2007; pp. 808–815.
17. Thompson, H.S.; Anderson, A.; Bard, E.G.; Doherty-Sneddon, G.; Newlands, A.; Sotillo, C. The HCRC Map Task Corpus: Natural Dialogue for Speech Recognition. In Proceedings of the Workshop on Human Language Technology (HLT '93), Plainsboro, NJ, USA, 21–24 March 1993; Association for Computational Linguistics: Stroudsburg, PA, USA, 1993; pp. 25–30. [CrossRef]
18. Carletta, J.; Ashby, S.; Bourban, S.; Flynn, M.; Guillemot, M.; Hain, T.; Kadlec, J.; Karaiskos, V.; Kraaij, W.; Kronenthal, M.; et al. The AMI Meeting Corpus: A Pre-Announcement. In Proceedings of the Workshop on Machine Learning for Multimodal Interaction (MLMI'05), Edinburgh, UK, 11–13 July 2005; Renals, S., Bengio, S., Eds.; Springer: Berlin, Germany, 2005; pp. 28–39.
19. Vinciarelli, A.; Dielmann, A.; Favre, S.; Salamin, H. Canal9: A database of political debates for analysis of social interactions. In Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, Amsterdam, The Netherlands, 10–12 September 2009; pp. 1–4. [CrossRef]
20. Braley, M.; Murray, G. The Group Affect and Performance (GAP) Corpus. In Proceedings of the Group Interaction Frontiers in Technology (GIFT'18), Boulder, CO, USA, 16–20 October 2018; ACM: New York, NY, USA, 2018; pp. 2:1–2:9. [CrossRef]
21. Sunstein, C.; Hastie, R. *Wiser: Getting beyond Groupthink to Make Groups Smarter*; Harvard Business Review Press: Brighton, MA, USA, 2015.
22. Reiter-Palmon, R.; Sinha, T.; Gevers, J.; Odobez, J.M.; Volpe, G. Theories and Models of Teams and Groups. *Small Group Res.* **2017**, *48*, 544–567. [CrossRef]
23. Koutsombogera, M.; Vogel, C. Modeling Collaborative Multimodal Behavior in Group Dialogues: The MULTISIMO Corpus. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018; Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., et al., Eds.; European Language Resources Association (ELRA): Paris, France, 2018; pp. 2945–2951.
24. Koutsombogera, M.; Costello, R.; Vogel, C. Quantifying Dominance in the MULTISIMO Corpus. In Proceedings of the 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2018), Budapest, Hungary, 22–24 August 2018; Baranyi, P., Esposito, A., Földesi, P., Mihálydeák, T., Eds.; 2018; pp. 141–146.
25. Goldberg, L.R. The development of markers for the Big-Five factor structure. *Psychol. Assess.* **1992**, *4*, 26–42. [CrossRef]
26. John, O.P.; Donahue, E.M.; Kentle, R.L. *The Big Five Inventory Versions 4a and 54*; Technical Report; University of California, Berkeley, Institute of Personality and Social Research: Berkeley, CA, USA, 1991.
27. John, O.P.; Naumann, L.P.; Soto, C.J. Paradigm shift to the integrative big five trait taxonomy. *Handb. Personal. Theory Res.* **2008**, *3*, 114–158.
28. Rammstedt, B.; John, O.P. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *J. Res. Personal.* **2007**, *41*, 203–212. [CrossRef]
29. Laugwitz, B.; Held, T.; Schrepp, M. Construction and Evaluation of a User Experience Questionnaire. In *HCI and Usability for Education and Work*; Holzinger, A., Ed.; Springer: Berlin, Germany, 2008; pp. 63–76.

30. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2008; ISBN 3-900051-07-0.
31. Ambady, N.; Bernieri, F.J.; Richeson, J.A. Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. In *Advances in Experimental Social Psychology*; Academic Press: Cambridge, MA, USA, 2000; Volume 32, pp. 201–271, doi:10.1016/S0065-2601(00)80006-4.

Sample Availability: With complete compliance with the terms of consent provided by the participants, 18 dialogues are represented in the version of the MULTISIMO corpus available under a non-commercial license at http://multisimo.eu/datasets.html.