electronics

MDPI

Article

Aspects of Dynamics in Dialogue Collaboration

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

Trinity Centre for Computing and Language Studies, Trinity College Dublin, The University of Dublin, D02 KF66 Dublin, Ireland

* Correspondence: vogel@tcd.ie (C.V.); koutsomm@tcd.ie (M.K.)

† These authors contributed equally to this work.

Abstract: Collaborative dialogue is an important category of human interaction and is widely studied in the literature, especially in fields that attempt to develop new technologies that enable wider varieties of collaborative dialogues. The ingredients of collaboration in dialogue are less thoroughly addressed. We describe the theoretical framework within which we are working and our approach to the construction of a theory of what may make dialogue collaborative. We study a multimodal dialogue corpus (MULTISIMO) testing for positive and negative correlations between dialogue content features and interaction features that one might reasonably imagine are related to assessments of degrees of collaboration. The duration before the second speaker's first turn and degree of imbalance in the number of words produced by speakers negatively correlate with collaboration assessments (that is, imbalances of content and a delay in the first speaker yielding the floor lead to diminished perceptions of collaboration), while a monotonically increasing duration of focus in successive dialogue sections (rather than overall dialogue duration) correlates positively (that is, when participants are deemed to be extending the duration of the task rather than increasing speed with experience, this is perceived as collaborative).

Keywords: collaboration; dialogue; collaboration assessment; annotation disagreement; interactional semantics; cognitive infocommunications

Citation: Vogel, C.; Koutsombogera, M.; Reverdy, J. Aspects of Dynamics in Dialogue Collaboration. *Electronics* **2023**, *12*, 2210. https://doi.org/10.3390/electronics12102210

Academic Editors: Adam B. Csapo and Mika Luimula

Received: 22 March 2023
Revised: 27 April 2023
Accepted: 3 May 2023
Published: 12 May 2023

1. Introduction

Our thoughts on collaboration are informed by the contemplation of fire. One might measure the success of a fire by the heat generated, the light emitted, the height of the flames, the duration of the flames, the aural quality of the crackling, or by other qualities. There is not a fixed notion of a successful fire. One might try to set a fire by holding a lit match to a dry piece of wood, removing the match when the fuel has a flame. This act may be successful in obtaining an initial flame. However, the flame is likely to quickly extinguish unless the initial piece of wood is in close proximity to additional fuel, with some gap that supports the circulation of oxygen. In such a configuration, a sustained fire is more likely, as the flames on each source "feed off each other". We think that the analogy of a fire provides insight into the nature of collaboration. One does not ascribe collaboration to a successful fire. What is missing from that scenario is intentionality within the fuel. If the pieces of wood in the hearth individually intended to create a fire, one would be inclined to judge that the success of the fire depended upon their collaboration.

It is useful to analyze the essential components of collaboration. Unlike the case of a fire, collaboration among people occurs in contexts in which it is not necessary to achieve a successful outcome. However, in some situations, collaboration is definitional of success. Collaborative dialogue is an example. In these situations, the analogy of the fire provides apt distinctions: one fuel source might have a higher degree or duration of flammability than the other, or they might be equal. Similarly, one party to the dialogue might be more active than the other in alternation or for the duration of the dialogue, or the parties may have equal inputs. Consider comedy teams in which one party has the "comic" role and the

other the "straight" one. Even silence on the part of the latter may lead to consistent effects, which lead to the team being deemed successful comic collaborators. In the case of dialogue, the "gap that supports the circulation of oxygen" is easy to supply, since in general, people do not know what is in each other's head (and when they do, they frequently engage in dialogue to confirm that, or, more rarely, have no dialogue at all). Dialogue is the analogue of the flame. Moreover, as in the case of fire, there are many ways that one might measure success: achieving a task, developing or maintaining a social bond, filling silence, etc. Dialogue partners may intend to collaborate, and actually collaborate, but fail to fill the silence. In attempting to measure collaboration, it is necessary to be clear about whether one is measuring collaboration itself or the outcomes of collaboration. The importance of isolating within scrutiny of collaboration preconditions, processes and outcomes has been highlighted before [1].

Measures of collaboration in dialogue associate with measurements of teams as team efficacy may be seen partly through efficacy in dialogue, stretched over time. Others have studied cognate matters, for example "potency" of collaboration with reflective questionnaires addressed by participants [2]. Behavioral measures have been the subject of cluster analysis, categorizing group outcomes into clusters of behavior types [3] . In the domain of language learning, researchers sometimes focus on tasks that elicit specific collaborative phenomena—"language-related episodes"—in which dialogue participants discuss nuances of the natural language being acquired [4]. It is common to examine the content of negotiation [5], including deploying computational approaches to identification of content-based cues, for example, of "team decision making" [6]. However, para-linguistic behaviors are also studied, for example, gaze [7] and intonation patterns [8]. Others note situations for which it is desirable for the behaviors of interest be amenable to automated measurement [9,10]. We note these dimensions of past works in order to highlight dimensions in which others have studied comparable problems with related interests.

We report here on our approach to building a theory of collaboration in dialogue. Within this work, we think partly of the judgments of independent observers of dialogue interaction and partly of the observable behaviors that might influence those judgments. We consider behaviors that may be classified or quantified. Given dimensions of such measurement that might rationally be thought to inform a judgment of whether dialogue has been collaborative, we explore specific ways of articulating and conducting those measurements. We then seek to quantify how those dimensions aggregate in providing support for ordinal categories of collaboration. Regression analysis allows the estimation of the influences provided by the dimensions of analysis. This is a "bottom-up" approach to theory building, where ultimately the specific theory produced is a regression equation, and by construction, fits the facts of the data analyzed. Our approach uses both observational methods and theory testing (see [11]). The latter is most visible in identifying dimensions of interaction that inform judgments of collaboration. Elements of the "top-down" theory-testing approach are in our decision to focus on observations that do not require deep knowledge of the content of interaction. We largely ignore what, exactly, interlocutors discuss and whether they express steadfast disagreement with their words. (We do attend to the pronouns they use, for example.) Rather, we seek to understand aspects of the *dynamics* of their interaction that inform judgments of observers of the interactions about the nature of collaboration within them. Others have also pursued "content-free" analyses of dialogues, even for purposes of indexing interactions [12], so we do not see this as a controversial theoretical exclusion. It is also easy to imagine technologies that would benefit from being able to assess whether interlocutors are in a collaborative mode of interaction at any point in time—for example, a privacy-preserving mediator intended to intervene with discreet prompts when interlocutors who should be collaborating are not. (The cognitive infocommunications literature, which systematically addresses technologies that can extend human cognitive capabilities [13–18], frequently addresses robots and the nature of robotic interactions. Sensitivity to the dynamics of interaction is part of the social

knowledge that humans have and which robots would need to obtain if they are to be useful in social contexts [19]. Similarly, in that literature, researchers try to understand the dynamics of interaction between humans and other humans and with (and through) computer interfaces [20–23].) Deciding to adopt or ignore a dimension with potential influence over a response variable is a "top-down" theory-driven distinction.

Our decision to seek a data- and theory-driven regression equation as an articulation of the current state of the theory is an homage to the theory of politeness in social interaction provided by Brown and Levinson [24]. In that work, the theory of politeness is presented in a form that evokes idealized theories of physics, addressing perceptions of politeness in terms of power, distance and rank of imposition in a dialogue act (some of the details of this are recalled below in Section 2). However, the terms of the equation are not matched with operationalized means of measurement, but this opens theoretical and empirical vistas as researchers explore those possibilities. We seek both terms and procedures for measuring them, not in order to close off possibilities but the contrary, because theory building in an age of "big data" evidently requires more concrete starting points within the scientific community in attracting interest to joining the endeavor of theory building and testing in relation to particular phenomena. Patently, selecting the form of measurement in a response variable and selecting both the "explanatory" variables and their forms of measurement are theory-driven. They are also data-driven in that some forms of measurement are possible and others not, and some relationships, depending on the measurements, carry more information than others (for example, if x is a count of interruptions, d is a duration in seconds and t is a count of turns, that $\frac{x}{t}$ is a more informative relationship than $\frac{x}{x}$ can be determined from a top-down perspective, and this is because it will contribute no more than the constant, 1, to any relationship that invokes the relationship, $f = \frac{x}{s} = 1$, but whether $\frac{x}{t}$ is more or less interesting than $\frac{x}{s}$ depends on observing what happens in the data). In this task, turning the theoretical notion of interest, i.e., collaboration, into a response variable that can be measured is itself an interesting process.

In related work [25], we have analyzed the MULTISIMO corpus, identifying participants' behaviors (such as repetition and balance of contributions) and features (personality traits, derived from participation in a big-five personality trait inventory, and dominance, as assessed by independent observers) that may be understood as contributing to collaboration. We also addressed their success in the dialogue task, which, we argue, depends on attunement to popular thought and reasoning outside the confines of their own "best answers" to questions. In that work, we were keen to identify what people do when asked to collaborate. The present paper expands on that work by relating some of these observable quantities to direct ratings of collaboration in the dialogues. We address only some of the observable quantities that were considered before and also consider additional quantities, such as variables that capture effects of experience, for instance, whether from one section to the next in the dialogues participants become quicker with the task or more "accurate" in their responses. More importantly, here, we consider the quantities in combined form and identify which, when taken together, provide a parsimonious model of collaboration judgments. The resulting regression equation amounts to a theory of the aggregation of the component variables.

The paper proceeds, firstly, by expanding on the theoretical framework we use in thinking about dialogue collaboration and theory building (Section 2) and then considering past approaches in the literature to operationally quantifying collaboration (Section 3). We then describe a multimodal corpus of dialogues appropriate for the purpose of analyzing positive collaboration and the collaboration assessments of the dialogues within that corpus. We detail the dialogue content and interaction features that we explore in relation to collaboration and the method of assessing those relations (Section 4). Empirical hypotheses are detailed in Section 5. The results are presented in Section 6, and from these results, a partial theory of collaboration in dialogue emerges, which we discuss in Section 7. We conclude (Section 8) with reflections on prospects for generalizing the theory.

2. Theoretical Framework

We see the presentation of a theoretical framework as a researcher's clearest insights into the ontology and epistemology that they assume. Foundational questions in any science are somewhat messy to answer. Interestingly, one can evidently advance sciences without clear answers to those foundational questions. However, one can try to make true statements while aspiring to be relatively clear about the sort of things one believes exist and how one might have access to knowledge about those things through measurement.

Sheridan [26] (p. 90) argues "Collaboration is not absolute; it is not a concrete product, mechanism, or technique. Rather, it is a dynamic process…". Furthermore, it is claimed that "parity" is essential to collaboration, (p. 90): "Parity within a relationship, however, should be interpreted as meaning equal in decision-making status, not equal in content or process expertise". As this discussion is focused on success in educational processes, it must be considered whether this is a necessary element of collaboration in all contexts. In some contexts, collaborators do not have parity of responsibility for decisions: a manager and subordinate may collaborate on a decision, but up the command chain, the manager will typically be held responsible for the decision and its impacts. Parity may well be an ingredient of collaboration but probably not in decision-making status. Instead, parity might more universally apply to candidacy for supplying ideas that may ultimately be acted upon. In this sense, parity is necessary for collaboration but not sufficient: there must also be actual distribution in the production of ideas that are considered for adoption. The "Abbott and Costello" test may be applied: a noncomic reaction to comic antics creates a contrast that places the comic in relief—although complementary, both categories of input are essential to the overall effect.

Wood and Gray [1] have analyzed a range of definitions used by researchers and identified a number of factors upon which definitions draw, which we thusly summarize (cf. [1] (Table 4, p. 147, ff.)):

1. Voluntary membership
2. Shared goals
3. Perspectival complementarity
4. Agency
5. Constructivism
6. Shared norms
7. Temporary structure
8. Interactivity
9. Problem focus
10. Solution seeking
11. Individual autonomy
12. Individual-transcending
13. Decision making

Wood and Gray [1] (p. 149) appear to regard these factors as jointly necessary to achieve a state of collaboration: "Many collective forms, however, explicitly excluded by the definition, including blue ribbon panels that are appointed but never meet, mergers of formerly independent organizations, legal entities such as corporations whose participants have no autonomy with respect to the entity's decision making, and clubs or other groups that meet regularly but have no specific problem solving objectives." However, an alternative is to take these as traits that may be shared to a greater or lesser extent, along the lines of family resemblances. A rational approach to measuring collaboration, given these factors, would be to treat each as binary and to count (possibly in a weighted manner) the number of these qualities that are present within a scenario. An alternative is to assess the degree to which each of these factors is satisfied within the scenario and to aggregate (again, possibly in a weighted manner) those individual scalar quantities.

Thomson and Perry [27] (p. 23) emphasize a contrastive approach to collaboration by settling on a definition that draws on a number of the features above (autonomy, interaction, solution seeking and shared norms), with additional emphasis on "jointly creating rules

and structures governing their relationships and ways to act…" The contrastive approach is in seeing collaboration as distinct from cooperation or coordination, building on work by Gray [28]. We agree that it is right to analyze relationships among cognate concepts but also think it is important to be aware of the intensity of synonymy avoidance within human cognition [29] and the consequent risk of specious distinctions. The argument of Thomson and Perry [27] (p. 23) is that collaboration is a higher-order construct than cooperation or coordination, in that in order to collaborate, collaborators must not only intend to collaborate but also "understand the multidimensional nature of collaboration." While it may be true that agents who reason about the component factors of collaboration are better collaborators than those who do not, it seems unlikely that in order for a person to function as a collaborator the person must have a conscious and complete working theory of collaboration. For many processes in which human cognition impinges, people are able to know how to do things without knowing what it is that they know that enables doing those things.

The MULTISIMO corpus [30] was constructed as a multimodal record of individuals participating in collaborative dialogue. Participants were asked to collaborate with each other, without providing a decompositional account of what is entailed by collaborating. The dialogues each include a facilitator who was instructed to guide the other two participants in each as they address a task, while encouraging collaboration of the participants. Again, the facilitators were not provided with instructions on what is entailed by collaboration. The MULTISIMO dataset thus enables the scrutiny of collaboration from a perspective that is atypical in the literature on collaboration in that it supports the study of what people do when they are asked to collaborate. If one makes the reasonable assumption that experimentally they more or less do what they are asked to, and if one respects the methodological principle that one must analyze the data as if participants have followed instructions unless there is impeccable evidence to the contrary [31], then one can conclude that the MULTISIMO dataset includes a subset of behaviors that participants engage in when they are collaborating. This is the approach that we have taken in past work [25].

It is an alternative approach to analyze this data from an independent standard of collaboration, scoring each pair of participants on their success in collaborating (or each participant on their success in contributing to a collaborating pair). To embark on this strategy in a post hoc manner, when it is no longer possible to ask participants how they rate their collaboration (individually or as a partnership), there are two main alternatives. One may put the interactions to scrutiny by raters who are asked to determine the level of collaboration in each, again without specifying the internal composition of a theory of collaboration that should inform their judgments. One may have interactions rated with reference to a relatively established framework, typically multidimensional, that specifies quantities and qualities to monitor.

Here, we take both approaches. One is to judge the interactions as total multimodal constructs using the unanalyzed holistic method. The other is to judge the linguistic modality alone, that is, based upon transcripts, using a multidimensional construct. A different judge embarks on each of these tasks. We take both of these ratings to provide "gold standards". We seek to identify the analytic and empirical relations between them. It is true that this research design could be more robust if it involved more judges and applied both methods to both modalities. However, we note that the wider literature on collaboration does not tend to dwell on communication channels. Furthermore, we note that if an aggregate judgment based on a transcript alone differs in kind from a holistic judgment based on the multimodal view, then the difference is explained either by the difference in judges or by the additional modalities. Our method of annotation allows us in the first instance to identify and resolve differences of opinion, but more importantly, to learn what conditions give rise to those differences. Thus, we assess annotator agreement, and we preserve annotation of disagreement. In reasoning about disagreement, we use the absolute value of the difference between the two assessors' scores for any item; Table 1 indicates the range of the raw values of these assessments (see [32] for an argument regarding the value

of studying correlations with variables that encode levels of disagreement in judgment alongside values that aggregate judgments of independent assessors).

The considerations so far do not resolve the scales that apply to measuring collaboration, whether as a holistic construct or the aggregation of multiple factors. Again, one may relate this situation to measuring properties of fires, for example, temperature. Quotidian decisions based on temperature make reference to a zero, where water freezes, and make reference to degrees below freezing and degrees above freezing. Some reference points above and below zero have theoretical and empirical relevance beyond day-to-day existence. The temperature at which water boils is a matter of everyday consequence for those who drink coffee, but the temperature at which books combust is of more seasonal interest, since it is also related to temperatures that keep fires going. The temperature at which the motion of molecules is arrested does not impinge on day-to-day human awareness. Furthermore, when thinking about the temperature of a fire, one tends to ignore the possibility of temperatures below zero. If collaboration or its constituents is to be measured in a scalar manner, it must be determined whether it has an absolute minimum or maximum. However, the theoretical scale appropriate to collaboration need not coincide with the range of values required by ordinary human behaviors.

If one adopts the view that the degree of collaboration involved in a situation is determined by the relative number of conditions (1–13) that are met, then there is a natural maximum value (1) and a natural minimum (0). However, such a perspective neglects scenarios where one would deem there to be anticollaboration. Of course, it is not obvious what anticollaboration might comprise. It is reasonable for a theory of conflict to presuppose that mutual conflict entails collaboration (otherwise, people would not say of blame in conflict situations that "it takes two to tango") but that it is destructive rather than constructive (or some other modification). It is sometimes witnessed that political interviews and debates descend into a pattern of interaction that is highly interlocked but without any evidence of progress towards the ostensibly shared discourse goal of scrutinizing the relative merits of positions or revealing information. In such cases, one might wish to assert that collaboration is visible but that it has a negative value. However, the relative quantity of (1–13) does not directly provide for the possibility of negative values.

In the case of dialogue that may be adversarial, one might consider measuring collaboration with reference to quantities such as those in (1–13) and the issues that they raise. However, in dialogue that is designed to be collaborative, reference to "obstructionist acts" might be contrived. On the other hand, it is intuitive that two dialogues that are both collaborative might differ in degree. It is less clear what exactly might give rise to graduated levels of dialogue collaboration. This is what we explore in this paper.

1. $\frac{supportiveacts}{obstructionistacts}$—undefined when there are no obstructionist acts;
2. $\frac{supportiveacts - obstructionistacts}{supportiveacts + obstructionistacts}$—negative when there are more obstructionist acts than supportive acts; positive when there are more supportive acts than obstructionist acts; maximum = 1 and minimum = -1; silent on acts that are neither obstructionist nor supportive, and 0 does not discriminate the absence of positive or negative collaboration from the cancellation of positive and negative collaboration;
3. $\frac{supportiveacts - obstructionistacts}{totalacts}$—maximum = 1 and minimum = -1; does not discriminate the absence of positive or negative collaboration from the cancellation of positive and negative collaboration.

In what follows, we present more detail of the operationalization of various qualities as measurable quantities. Now, we return to a point made in the introduction about theory building in relation to the process of identifying these quantities and evaluating their weights and significance within theories that aggregate them. In particular, we use simple linear regression as a tool that supports theory building. We note a passage from Brown and Levinson [24] (pp. 74–76) in a presentation of the theoretical construct of the "weightiness" of "face threatening acts" (FTA) in relation to a speaker (S) and hearer (H). Within this quotation, D is understood as some notion of social distance separating the speaker (S)

and hearer (H); P is a representation of the relative power of the speaker and hearer in the situation; and R is understood as a culture-specific ranking of the amount of imposition conveyed by some act (indexed by the subscript x) that threatens face. These notions are put forward in an intuitive sense. The quotation follows (emphasis, etc., is as in the original, except that the original used lowercase Roman numerals in parentheses to enumerate):

> . . . we argue that the assessment of the seriousness of an FTA (that is, the calculations that members actually seem to make) involves the following factors in many and perhaps all cultures:
>
> 1. The "social distance" (D) between S and H (a symmetric relation);
> 2. The relative "power" (P) of S and H (an asymmetric relation);
> 3. The absolute ranking (R) of imposition in the particular culture.
>
> An immediate clarification is in order. We are interested in D, P and R only to the extent that the actors think it is mutual knowledge between them that these variables have particular values. Thus, these are not intended as *sociologists' ratings* of *actual* power, distance, etc., but only of actors' assumptions of such ratings, assumed to be mutually assumed, at least within certain limits.
>
> Our argument here has an empirical basis. . .
>
> . . . So let us say that the weightiness of an FTA is thus calculated thus:
> $W_x = D(S, H) + P(H, S) + R_x$

We note that this deeply influential theoretical approach to the analysis of politeness in language aspires to the precision of set theory and multivariate analysis of response variables in exposition. While Brown and Levinson [24] explicitly allow for a distinction between a social scientist's measurements of the underlying constituent factors and dialogue participants' calculations of the same, we note that they may coincide, and the likelihood of a strong positive relationship is embedded in the interlocutors access to "mutually assumed" values for the relevant quantities. In other words, there is every reason to think that independent observers, including social scientists, are able to approximate the interlocutors' assumptions about the relevant terms, if interlocutors are assumed, in general, to share assumptions about those terms. We focus on this example from the literature on politeness in communication because we see the epistemic and ontological approach to be aligned with what we report in this paper—not about the weightiness of individual face-threatening acts during a communication but about the perception of collaboration in a stretch of dialogue, including its entirety.

3. Approaches to Assessing Collaboration

In the wider literature, beyond interest in collaborative dialogues in particular, collaboration is a topic of scholarly scrutiny in a range of disciplines—education, health care, management, etc. Some metrics that have been developed vary, sometimes intended for "first-person" assessment of participants' perceptions of their experiences and sometimes for "third-person" assessment by parties external to the relevant situations.

Collaboration measures in the research domain are useful for comparing the degree of collaboration in scientific fields and providing evidence about trends towards multiple authorships in a discipline. The Collaborative Index calculates the mean number of authors per paper [33] and the Degree of Collaboration the proportion of multiple-authored papers [34], while the Collaboration Coefficient uses merits of the aforementioned measures but differentiates among levels of multiple authorships [35]. Relatedly to this domain, Scientific Collaboration Networks examine collaborative relations between researchers and quantify them using graph-based methods [36]; to measure collaboration among grant partners, the Levels of Collaboration Scale was used, i.e., a scale ranging from 0 to 5, with the lowest level representing little or no collaboration and the highest level representing full collaboration [37].

The medical literature often addresses collaboration in the spirit of shared decision making for clinical practice and the multidimensional aspects of collaboration among

nurses and physicians. A model of shared decision making among clinicians and patients was constructed by [38] and is based on choice, option and decision talk.

A concise review of instruments related to nurse–physician collaboration is performed by [39], listing, among others, the Collaborative Practice Scale [40], assessing the perception of assertiveness and collaboration, with related items measured on a 6-point Likert scale; the Collaboration and Satisfaction about the Care Decisions instrument [41] measuring nurse–physician collaboration and satisfaction about care decisions in intensive care units; and the Jefferson Scale of Attitudes Toward Physician–Nurse Collaboration [42], measuring physician and nurse attitudes toward authority, autonomy, responsibility, collaborative decision making and role expectations based upon 15 questions answered on a 4-point Likert scale.

Educational and psychological research has highlighted collaborative problem solving (CPS) as a 21st century skill, a skill necessary for the emerging workforce, addressing shifting workplace requirements. In this respect, [43] describe a scoring process for measuring CPS in online environments. The generalized scoring process consists of seven development steps and relies on Hesse's CPS framework [44] to interpret the construct.

A model of collaboration was constructed by [45] upon four Contributing Factors to Collaboration, physical space, social space, team mental model and organizational goals, to explore a professional environment and to assess day-to-day collaboration of a professional team. Ref. [46] presents a general-purpose self-evaluation tool assessing collaboration in groups. This tool allows groups to rate their collaboration on key factors (such as communication, sustainability, evaluation, leadership, community development, etc.) in the form of a checklist, where group members mark their opinion on a 1–5 Likert scale.

While not directly assessing collaboration as a quantity on its own, a number of researchers have tried to assess behaviors in dialogue that lead to success in achieving tasks through dialogue [47–58]. However, it seems clear that it is possible to collaborate extensively without achieving mutual understanding or task-related success in task-based dialogue.

Although we perceive that the literature on collaboration, including systems for supporting collaboration in dialogue, is vast, we are not aware of prior work that attempts to articulate what constitutes collaborative dialogue outside the healthcare domain.

4. Materials and Methods

We analyzed the full set of 23 dialogues collected within the MULTISIMO multimodal corpus of dialogues [30]. The corpus was constructed following research ethics review and approval by the Research Ethics Committee of the School of Computer Science and Statistics of Trinity College, the University of Dublin. Participants all consented to audio and video recording of interactions involving, in each case, a facilitator and two others for the purpose of research in dialogue interactions. Some individuals did not consent to their data being shared with others, and therefore, only 18 of the 23 dialogues are available for others in the research community. As indicated above, in each dialogue, facilitators asked the other two participants to collaborate in identifying what an independently sampled group of 100 people might have identified as answers to questions such as "what are three locations where people are likely to catch a cold?", "what are three instruments in a symphony orchestra?", "what are three things that people cut?", and furthermore, to rank the three responses in terms of popularity according to the independent sample. The task approach is inspired by a popular television game, "Family Feud". The questions were the same for each of the groups, selected by us on the basis of our impression that participants would each have ideas but not necessarily the same idea of what "most people" would think of as answers to the questions. Neither participants nor facilitators were instructed about how to collaborate nor what constitutes collaboration. However, facilitators were asked, in isolation from the others, to encourage participants to contribute equally if they perceived a gross imbalance of contributions. Table 2 provides an indication of the duration of dialogue sessions and the sections of them in which named items are ranked for imagined popularity.

Table 1. Range of scores provided by each accessor, the median score for each session, and the disagreement distance for each session

Assessment	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Assessor 1	1.000	2.000	3.000	2.652	3.500	4.000
Assessor 2	1.000	2.000	3.000	2.783	3.000	4.000
Aggregate	1.000	2.000	3.000	2.717	3.250	4.000
Difference	−2.0000	0.0000	0.0000	0.1304	1.0000	1.0000

Table 2. Durations, measured in seconds, of the ranking phases of the dialogues and the dialogue sessions as a whole.

Duration	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Ranking	79.27	130.18	150.86	184.98	198.40	426.49
Overall Session	335.80	469.70	569.60	589.40	705.20	994.50

In analyzing collaboration in the dialogues, we give primary attention to the two individuals in each session who are tasked with answering the questions, the "players", rather than the facilitator. Table 3 indicates the durations of the three ranking subsections across the dialogues. The players were presented with the same questions in the same order in each of the dialogue sessions.

Table 3. Range of durations, in seconds, for the successive ranking sections of each dialogue.

Ranking Discussion	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
1 (Infection)	28.90	46.82	68.39	81.57	96.47	215.59
2 (Music)	11.34	21.08	35.13	40.97	46.15	178.17
3 (Cut)	21.53	37.44	49.24	62.44	78.22	173.31

In some sense, the durations indicate the relative difficulty of the questions, in that the ranking of the most frequently named musical instruments used within symphony orchestras—the second of the three questions in each case—is evidently easier (or, at any rate, took less time) than the other two. Table 4 provides another view of difficulty through mean agreement between the players' ranking of items for each of the three questions with the judgments of the same in a pool of 100 independently surveyed people. We measure agreement using nonparametric correlation, namely correlation of ranks [59], given that the values available, since they are not normally distributed, do not support the perhaps more familiar Pearson correlation. The method of scoring is based on Kendall's τ statistic, which matches two rankings of n items by counting the "concordant pairs" (c), the pairs of items that are ranked in the same relative order in the two rankings, and the "discordant pairs" (d), the pairs of items for which the two rankings disagree, taking the ratio of the difference between the concordant and discordant pairs to the total number of pairs, as in (1). The aggregate score for a session is the mean of the τ scores for each of the three ranking sections. The final line of Table 4 indicates the distribution of mean-τ scores (the value in the bottom line is **not** the mean of the values in the preceding three lines). Thus, Table 4 demonstrates that in addition to deciding the most quickly on rankings for the second question, the players were generally the best attuned to popular opinion in their responses to this item as well. Participants were least accurate in their response to the third item.

$$\tau = \frac{(c-d)}{\left(\frac{n(n-1)}{2}\right)} \tag{1}$$

We constructed three Boolean conditions for sessions in relation to the duration of the ranking sections of each (see Equation (2)): true, in each case just if the left ranking section for a pair of players was strictly shorter in duration than the right ranking section.

Truth for these relations for a pair of players has two possible interpretations: decreased speed in the task with increased experience, or increased conversational engagement with increased experience. Table 5 indicates the number of sessions for which each of these relations is true or false. We also constructed three Boolean conditions for sessions for the measures of accuracy in matching popular opinion between sections—see Equation (3). Truth for these relations encodes an increase in task "correctness" with experience. Strict inequality is used in the duration conditions, because durations were measured with millisecond precision, and therefore chances of equality are vanishingly small anyway; similarly, with only three items in the rankings, only a small set of values for Kendall's τ are arithmetically possible, and equality is rather more likely. Table 6 indicates the number of sessions for which each of these relations is true or false. We tested whether Gestalt assessments of collaboration and levels of collaboration assessor disagreement differed according to these Boolean conditions.

$$1 \overset{d}{<} 2, 1 \overset{d}{<} 3, 2 \overset{d}{<} 3 \tag{2}$$

$$1 \overset{a}{\leq} 2, 1 \overset{a}{\leq} 3, 2 \overset{a}{\leq} 3 \tag{3}$$

Table 4. Kendall's τ as a measure of players' ranks' conformity to opinion of a 100-person reference sample in each of the three ranking sections and the mean of the three τ scores within each session.

Ranking Discussion	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
1 (Infection)	−0.3333	0.3333	0.3333	0.4783	1.0000	1.0000
2 (Music)	−0.3333	0.3333	1.0000	0.7681	1.0000	1.0000
3 (Cut)	−1.0000	−0.3333	0.3333	0.1884	0.6667	1.0000
Mean − τ	−0.1111	0.3333	0.5556	0.4783	0.6667	1.0000

Table 5. Counts of sessions in which Boolean duration relations between ranking sections hold true and false.

	$\mathbf{1 \overset{d}{<} 2}$	$\mathbf{1 \overset{d}{<} 3}$	$\mathbf{2 \overset{d}{<} 3}$
FALSE	20	17	5
TRUE	3	6	18

Table 6. Counts of sessions in which Boolean accuracy relations between ranking sections hold true and false.

	$\mathbf{1 \overset{d}{<} 2}$	$\mathbf{1 \overset{d}{<} 3}$	$\mathbf{2 \overset{d}{<} 3}$
FALSE	3	9	15
TRUE	20	14	8

The dialogues were annotated using two approaches to collaboration assessment. One, informed by the literature discussed in Section 3, addresses the totality of the multimodal corpus in a Gestalt fashion, with two raters addressing each of the 23 dialogues with the question "how much do the players collaborate?" using a five-point scale ("not at all", "a little", "moderately", "highly" and "very highly"). Thus, the rating scale is based on a rank order of levels, with a minimum and maximum possible value. We adopted this range because, from our knowledge of the dialogue corpus and past scrutiny of dominance within it [60], values below "zero" would not have been triggered. We used this low-granularity rating scheme because the total number of data points (23) seemed unlikely to us to support a more finely graduated scheme. We coded the levels of collaboration assigned from 0 ("not at all") to 4 ("very highly") using the natural ordering of labels provided above. The assessors both observed the sessions as videos with synchro-

nized audio playback and provided a rating at the end of each video in answer to the question above.

Table 1 summarizes the ratings of the two assessors: an aggregate of their judgments (the median score for each session) and the difference in judgments. The intraclass correlation coefficient shows significant agreement: $ICC(A,1) = 0.696, F(22,22.8) = 5.52$, $p = 6.92 \times 10^{-5}$. Computing Pearson's correlation, one obtains a coefficient of 0.711, $p < 0.001$. Thus, it is reasonable to use an aggregate score in studying other quantities in relation to perceived collaboration among participants in the sessions. We also analyze the difference between the assessors' scores in relation to other quantities of interest.

The assessors were among the authors of this paper. After completing the annotations, they reported on the cues they relied upon to perform their assessments. Raters' impressions were formed on the basis of visual, audio and linguistic cues. Specifically, raters reported on signs of physical engagement, as manifested in a range of participants' body movements: head nods signaling active listening and feedback and torso leaning towards the partner and eye contact among partners to show engagement. Addressing the facilitator frequently through gaze or verbally, instead of focusing on the partner, was considered a sign of decreased collaboration. On the contrary, verbal content including argumentation about partners' choices and opinions, as well as questions checking the agreement of the partner, were considered constituents of collaborative activity. Raters also acknowledged that because of the nature of the task, the ranking parts of the dialogues elicit more collaboration-related activity by the participants.

The holistic measures above addressed the sessions in multiple modalities. We also adopted an alternative that uses a few dimensions of rating but with respect to a single modality of the dialogue sessions, namely the linguistic content of the dialogues as recorded in transcripts and their metadata, such as indication of utterance start times. Within the transcripts, we isolated the sequence of turns that comprise the ranking discussions for each of the three questions within a dialogue game. We focus on the players. We count the number of utterances from the dialogue that comprise the ranking discussion and the number of those that involve chat from the players, as well as the number of turns in the ranking discussion that constitute questions. As discussed in Section 3, we deem offering declarative content to be an indication of participation. We discuss some variables separately [25]: TURN-BALANCE indicates the degree to which participants are equal in the turns they take across the session; WORD-BALANCE indicates the degree to which participants contribute equal numbers of word tokens (WORD-BALANCE may be relativized to the total number of word tokens uttered by the players or to the total duration of the time spent ranking items); and TIME-BALANCE indicates the degree to which the players had equal durations of speaking, given the total session duration. We have not measured overlap here, and therefore cannot subtract it in order to measure the relative amount of time each speaker "holds the floor". We pay attention to the first moments of the players' contributions in each session. In particular, we study the temporal distance between the completion of the facilitator's opening to the players and the first utterance from either player (START LAG), the temporal distance between the onset of the first utterance from either player and and the onset of the first utterance of the other player (INITIAL FLOOR HOLDING), as well as a combination of those two distances. It would be natural to explore this ratio over each of the three questions of the dialogue, but for this work, we are interested mainly in the very first moments of the players' participation, in isolation from their participation for the second and third questions, because we think these first moments will contribute very much to the assessors' views of collaboration within the sessions as wholes, and because we imagine that the participants themselves would have formed first impressions from those moments as well. Questions demonstrate an involvement of oneself in a dialogue and ones' interlocutor. Therefore, we also count the number of ranking section turns that are questions. We look at the chat count and question counts relative to the total ranking section count (the complement consists of the contributions of the facilitator) and duration. Inspecting the transcript-based measures detailed above, it is evident that they

are mostly "content-free" features that reveal patterns of interaction [12,61]. We also attend to a small number of content-based features, counting the explicit mentions of "people" and also the use of first-person singular and first-person plural pronouns.

Each participant completed a personality assessment instrument using the Big Five inventory [62,63]. Their results were computed to achieve percentile ratings for each participant with respect to each of the attributes: openness, conscientiousness, extraversion, agreeableness and neuroticism. From these, as each session involved exactly two players, we noted the absolute value of the difference in their percentile score for each attribute.

Of the quantities that are demonstrated to correlate with aggregate collaboration assessment, we construct a linear regression model using those quantities as predictors of the aggregate collaboration measure. We seek to balance maximizing adjusted R^2, the amount of variation in collaboration assessments, with model simplicity, reducing the model in a stepwise fashion by eliminating the least significant contributing factor. A model constructed to maximize R^2 fully to 1 (accounting for 100% of the variation in the regression target value) will nearly certainly overfit the data at hand and will therefore have limited value in approaching distinct datasets. Therefore, we do not expect all of the terms of demonstrable independent interest to factor into the model.

5. Empirical Hypotheses

We expect that Gestalt judgments of collaboration made with reference the multimodal content of the dialogues to correlate positively with the relativized question count, chat count and turn balance. We expect the balance of dominance perceived among the participants to correlate negatively with Gestalt judgments of collaboration. It seems unlikely to us that in this context assessors will have been influenced by correctness of responses to questions but likely that duration of deliberations in ranking will have influenced assessors of collaboration; we expect that where the ranking deliberation increases from section to section, collaboration will be assessed as greater. We also expect that where correctness is monotonically increasing, no impact on aggregate collaboration assessment will be seen.

We expect disagreement in the Gestalt judgment of collaboration to correlate with player personality type mismatches. Our intuition is that as observers of the dialogues, the assessors may make conscious or unconscious adjustments to their view of the individual contributions of the players on the basis of their own view of each player's personality. A difference in collaboration assessment between the assessors may have a link to the different degrees to which the players may be ascribed particular personality traits.

Within any session, a shorter START LAG, we think, will be indicative of a more pronounced degree of collaboration than a longer START LAG; longer START LAG values may suggest a reluctance to "jump in" to the conversation. However, this value measures only one player. Longer INITIAL FLOOR HOLDING values are consistent with players allowing their partners to dominate the first turn, and such dominance is likely to correlate inversely with perceived collaboration. We expect effects of inordinately long START LAG values to be mitigated by shorter INITIAL FLOOR HOLDING values and effects of longer INITIAL FLOOR HOLDING values to be mitigated by shorter START LAG values. The sum of these two values is a simple aggregation model and corresponds directly to the *start lag* for the second player in any session. Thus, we expect aggregated collaboration assessment to be inversely related to the second player's start lag.

Where features of the dialogues can be understood as demonstrating collaboration when "in balance", such that the feature takes on greater magnitude with greater imbalance, we expect negative correlations between the feature and the assessments of collaboration. For example, we expect negative correlations between TURN BALANCE, WORD BALANCE and TIME BALANCE, respectively, and the aggregated assessment of collaboration; conversely, we expect positive correlations between these quantities and measures of the assessors' disagreement.

Repetition features, such as those recorded in the Jaccard scores, differ as a function of source and target: in the context of the MULTISIMO dialogues, self-repetition may be

easily understood as a negative persistence, while other-repetition may be understood as explicit signaling of grounding. We note that in other contexts, self-repetition might be cooperative and that other-repetition may be destructively sarcastic.

Jaccard distances are computed pairwise among turns (apart from turn–identity pairs) in dialogues. We consider the mean Jaccard distance associated with the sessions for the players, where players in each pair are distinct (i.e., "other-repetition") and where the players are the same (i.e., "self-repetition"). The Jaccard distance is given in (4), where $i \neq j$ are indices of dialogue turns, and "words$_i$" refers to the set of word types used in the ith turn.

$$\text{Jaccard-distance}(\text{words}_i, \text{words}_j) = 1 - \frac{|\text{words}_i \cap \text{words}_j|}{|\text{words}_i \cup \text{words}_j|} \tag{4}$$

Among team-building word use, we anticipate that players who make reference to "people" are therefore reasoning explicitly about their attunement to others in the task, and that those who do this more will be deemed more collaborative and thus positively correlate with assessments. We expect that where there is a proliferation of first-person singular pronoun use, collaboration will be thought correspondingly less, and where there is a proliferation of first-person plural pronoun use, collaboration will be thought greater.

In forming a linear model that integrates significant features, we anticipate that the best model will be sensitive to the second player's START LAG, players' questions relativized to the ranking duration, successive increase in ranking section duration, balance of contributions and aspects of players' personality traits.

6. Results

6.1. Collaboration in Relation to Individual Variables

6.1.1. Discourse Acts

The aggregated Gestalt judgments of collaboration correlate positively with the number of questions asked in the ranking phases of each game, whether that count is relativized to the total number of turns in ranking sections (Spearman's $\rho = 0.428, p = 0.04147$) or to the total number of turns in the dialogue (Spearman's $\rho = 0.482, p = 0.01997$). Aggregate collaboration assessment also correlates positively with the count of player questions in ranking sections relativized either to the ranking section durations (Spearman's $\rho = 0.5086, p = 0.0132$) or the overall session durations (Spearman's $\rho = 0.4915, p = 0.0172$). In contrast, the number of player chat turns in the ranking sections is not correlated with judgments of collaboration (neither relativized to the ranking section nor to the overall dialogue). Aggregate collaboration assessment is not significantly correlated with the chat count relativized to ranking section duration or dialogue duration. The difference in the judgements of collaboration made by the assessors of each session does not correlate with the ranking section chat counts, either.

6.1.2. Task Experience

There was no correlation between aggregate collaboration assessment or assessor disagreement and aggregate accuracy in sessions. Aggregated collaboration assessments were partly influenced by successive ranking section durations. The assessors indicate significantly higher collaboration levels when the third ranking session is longer than the second ranking session ($2 \overset{d}{<} 3$; Wilcoxon $W = 4, p = 0.002158$, two-tailed). The other two Boolean duration conditions did not elicit significant effects, nor were there significant interactions between any of the Boolean duration conditions and assessment disagreement. The aggregate collaboration assessments had no significant interaction with the Boolean accuracy relations. A significant interaction was evident between the assessors' disagreement and accuracy differences between the second and third ranking sections ($2 \overset{a}{\leq} 3$; Wilcoxon $W = 29, p = 0.02401$, two-tailed): there was higher disagreement with regard to sessions where the third ranking section was more accurate than the second ranking section (means of 0.875 vs. 0.267).

6.1.3. Turn-Taking Balance

Neither the aggregated collaboration assessments nor the measures of collaboration assessment disagreements correlate with the balance of turn taking by participants. Aggregated collaboration assessment correlates negatively with the balance of contributions measured in words (Spearman's $\rho = -0.4860, p = 0.0187$)—that is, as the imbalance increases, the assessment decreases. The difference between the assessors' judgments and word balance is positively correlated (Spearman's $\rho = 0.4701, p = 0.0236$). If the word balance is relativized not to the total number of words of the players but the total duration of ranking discussion, then the negative relationship with the aggregate assessment merely approaches significance (Spearman's $\rho = -0.3951, p = 0.0620$), while the positive relationship to the differences in assessors' judgments remains significant (Spearman's $\rho = 0.4534, p = 0.0290$). Equality of speaking time relative to the total dialogue duration does not correlate significantly with the aggregate assessment of collaboration or disagreement.

6.1.4. Turn Durations

There is not a significant correlation between aggregated collaboration assessment of either START LAG or INITIAL FLOOR HOLDING as isolated quantities, but there is a negative correlation between their sum and aggregated collaboration assessment (Spearman's $\rho = -0.4499, p = 0.03126$). Furthermore, collaboration assessors' disagreement correlates positively with INITIAL FLOOR HOLDING (Spearman's $\rho = 0.4242, p = 0.04366$) and with the sum of *start lag* and INITIAL FLOOR HOLDING (Spearman's $\rho = 0.4735, p = 0.02247$), but it only approaches significance with respect to START LAG (Spearman's $\rho = 0.3844$, $p = 0.07013$).

6.1.5. Player Dominance

The aggregate assessment of collaboration among players correlates negatively with the balance of independent assessments of dominance between players (Spearman's $\rho = -0.6761, p = 0.0021$). There is not a significant correlation between collaboration assessors' disagreement and the (im)balance of dominance between players.

6.1.6. Personality Traits

With respect to personality traits, the difference in players' openness, conscientiousness and agreeableness showed no correlation with the assessors' judgments of collaboration. Differences between players' extraversion (Spearman's $\rho = 0.4521, p = 0.0303$) and neuroticism (Spearman's $\rho = 0.4306, p = 0.0402$) showed significant positive correlation with judgment differences between the assessors.

6.1.7. Linguistic Repetition

Players' self-repetition scores did not have a significant correlation with aggregate collaboration assessments but had a positive correlation with collaboration assessment disagreement that approaches significance (Spearman's $\rho = 0.3546, p = 0.0969$). Neither aggregate collaboration assessment nor assessor disagreement correlate with other-repetition.

6.1.8. Reference to Groups

With respect to word choice, none of the mentions by players of "people", use of first-person singular pronouns, or use of first-person plural pronouns correlated with either aggregate collaboration assessment or assessor disagreement.

6.2. Collaboration in Relation to Combined Variables

The tests described above do not depend on the aggregate collaboration assessments being normally distributed. In fact, testing these values for normality using the Shapiro–Wilk test reveals that it is not reasonable to reject the null hypothesis that these values follow a normal distribution ($W = 0.93973, p = 0.1773$). Therefore, we conclude it appropriate to

apply regression analysis. In building a linear model regressing on aggregate collaboration assessments, as predictor variables we first include START LAG for the second player to speak, the ratio of players' questions during the ranking sections to the ranking section durations, the Boolean conditions $2 \overset{d}{<} 3$ and $2 \overset{a}{\leq} 3$, WORD BALANCE, the difference in players' percentile extraversion and neuroticism, player self-repetition, the total number of turns for the sessions and the total duration for the sessions. We do not include the balance of dominance of the players since this is closely (if inversely) related to collaboration; as noted above, the dominance assessments are seemingly informed by factors comparable to those informing collaboration assessments. This initial model has an adjusted R^2 value of 0.6128. The final model ($R^2 = 0.6537; F = 14.84$ on 3 and 9 DF, $p = 3.234 \times 10^{-5}$) retains as terms the second player's START LAG ($p = 0.014827$), $2 \overset{d}{<} 3$ ($p = 0.000471$) and WORD BALANCE ($p = 0.021063$). In the regression equation in (5), C stands for aggregate collaboration assessment; L for the second-speaking player's START LAG; I (increasing duration) for $2 \overset{d}{<} 3$; and W for WORD BALANCE.

$$C = 2.549 - 5.86 \cdot 10^{-5} L + 1.163 I - 2.5 W \tag{5}$$

7. Discussion

Here, we return to the hypotheses (Section 5) and interpret the results in their light (Section 7.1). We outline how we think this theory impinges on the semantics of multimodal dialogue (Section 7.2).

7.1. Interpreting the Results

Some of our experimental hypotheses are supported and others are not. The number of questions players posed during the ranking sections of their sessions correlated with aggregate collaboration assessments on a number of reasonable relativizations of the question count. The amount of chat during the ranking sessions did not matter, nor did the imbalance of turn taking or speaking time, but the imbalance of words produced did matter. The amount of time before the first speech of the second player to speak in a session also mattered. Negative correlations were significant for both of these variables. We understand from these measures that an imbalance of contributed words between the participants over the duration of the conversation impacts negatively on perceptions of collaboration, and that the longer the first player to speak holds the floor in the first turn, the less collaborative the communication is perceived to be.

It surprised us that measures oriented to direct linguistic content did not interact with collaboration assessments. Reasoning about "people" did not increase collaboration assessments; use of inclusive first-person plural pronouns did not increase collaboration assessments; and use of exclusive first-person singular pronouns did not decrease collaboration assessments. Rather, the features that emerged as relevant were determined by aspects of the way things were said.

The imbalance of extraversion and neuroticism between the players showed a positive interaction with assessors' disagreement about collaboration. We speculate that this relates to the assessors' response to the players' personalities, as well as differential expectations and adjustments they made in their assessment of collaboration: if one player is high in extraversion and their partner is low, one might reset expectations regarding what counts as "high collaboration".

We thought that assessments of collaboration would not depend on the players' accuracy in identifying the same rankings of items as the independently sampled reference set; however, we see that there is a correlation between collaboration assessments and nondecreasing correctness between the second question and the third question. We thought that assessment of collaboration would depend on the overall duration of the sessions, but it seems instead to depend on a more refined notion: an increase in durations of successive session sections. That is, as one section of the dialogue occupies more time than the

prior section, greater collaboration is perceived; this is consistent with durations failing to decrease in time required being understood not as task inefficiency but as desire to interact and collaborate.

We have arrived at a partial theory of collaboration in dialogues in which participants endeavor to follow the instruction to collaborate with each other. The theory is partial because (5) it explains only 65% of the variation in the aggregated collaboration assessments, and only in the context of the MULTISIMO corpus. However, we think that at least the terms of theory, if not the exact coefficients, are transferable or adaptable to other corpora. For example, the START LAG for noninitial speakers and WORD BALANCE are likely to be relevant in other contexts. The notion of the third ranking section having a duration no less than the second is rather specific to the MULTISIMO corpus, but it may be adapted to other indications of nonmonotonically increasing durations of shared focus.

7.2. Ramifications for Dialogue Semantics

Discourse semantics, for all its challenges, appears to be more straightforward than dialogue semantics. There is much to be gained from reflection on the conditions that must hold in the world or situations within it in order for sentences of natural languages to be true. Moreover, truth-conditional semantics scales reasonably well to discourse and may also be supplemented by alternative analytical frameworks, with a focus on dynamic interpretation processes and belief revision, for example. Still, the generalization to discourse is not without challenges, as discourse inevitably creates more pressure to take into account the context and purpose of the discourse in assessing the meaning of constituent sentences: the texts of narrative histories, legal statutes, fiction and so on make distinct commitments to the relationship between their constituent sentences and their truth in the world, for example. All the same, people happily read paragraphs and, making adjustments for their purpose, feel they understand what the paragraphs mean.

Even though dialogue semantics appears to be simply further along a spectrum than discourse in requiring attention to context and purpose, dialogue is different in kind, as well as degree. If asked what a conversation means, we think that most people would sooner comment on the relationships that may be guessed to hold among the interlocutors than the set of propositions each in the end believes or the set of questions each entertained. We think that people are able to infer relationships reasonably well when they observe conversations in languages they do not understand, when they cannot grasp the content of utterances but can discern interaction patterns (see, for example, [64]). An interactional semantics for dialogue must certainly attend to collaboration among interlocutors. We return to the analogy of fire. People readily accept the natural regularity that "smoke means fire" but require deeper contextual reflection to understand what fire means. Certainly, it means that the fuel was in a combustible state, and the determination of that configuration and relevant conditions (including intentionality of ignition) forms part of the official investigation of fires that cause injury to life or property. In understanding what dialogue means, the "configuration and relevant conditions" include the determinants of the degree of collaboration.

We have shown that determination of collaboration in dialogue need not heed linguistic interpretation of content, although it likely does rely on the quantity of content produced by each interlocutor and how that content distributes over time within the conversation.

8. Conclusions

This study has directly addressed only one category of dialogue in which participants positively engage in collaboration and, in fact, only one multimodal dialogue corpus. Although there are 23 dialogues within the corpus, this is still a relatively small study. Appropriate to the circumstances of the study, only a small range of positive values for collaboration fits; even in this context, it has proven interesting to seek to identify potential triggers for collaboration assessment disagreement. This work provides a continuation point for addressing additional dialogue corpora in the same spirit. While we would

certainly obtain different coefficients with other corpora, we would expect many of the same quantities to play a role in explaining the variability of collaboration assessments individually and largely in combination. What our work has shown is that making operational intuitive concepts into quantities matters. Thus, we also expect some differences as other datasets are analyzed, even when datasets are pooled. This would not constitute a refutation of our approach but, rather, a refinement.

Within this circumscribed context, we have put forward a theory of collaboration, made specific by the linear regression Equation (5). With respect to a larger set of sessions collected using the experimental paradigm we deployed, other terms that we considered may also play a role (as well as terms that we did not address here). Similarly, if other dialogues designed to trigger collaborative behaviors are studied, we would begin our scrutiny with the measures we described. If we were to apply this work to dialogue types where obstructive acts would be reasonably expected alongside supportive acts (such as a political journalist's interviewing of a politician), it would be necessary to explore a wider range of possible values for collaboration assessments.

Implicit in this work is the position that while it is appropriate to study collaborative dialogues as a genre, it is useful to consider whether they are "collaborative" merely because they were named that way or more because they display properties that constitute collaboration.

Our narrow empirical focus here is something like studying fires, only in fully functioning (i.e., vented, chimney-swept, etc.) wood-burning stoves burning kiln-dried wood and not outdoors at campsites on wet ground in rainy winter weather. However, some knowledge of fires inside wood-burning ovens is relevant to campfires in adverse weather conditions. If we were to cross-generalize what we noticed about collaboration in the MULTISIMO corpus to the study of fire, we would suggest that the duration before the second piece of wood ignites is important, as is the degree to which both bits of wood in a pair equally contribute flame. On the other hand, monotonic increase in flame duration during successive combustion of the same fuel is not possible, since fire consumes its fuel, whereas, in general, dialogue does not.

Author Contributions: Conceptualization, C.V. and M.K.; methodology, C.V. and M.K.; software, C.V.; validation, C.V., M.K. and J.R.; formal analysis, C.V. and M.K.; investigation, C.V., M.K. and J.R.; data curation, M.K.; writing—original draft preparation, C.V.; writing—review and editing, C.V. and M.K.; visualization, C.V. and M.K.; supervision, C.V. and M.K.; project administration, M.K.; funding acquisition, C.V. and M.K. All authors have read and agreed to the published version of the manuscript.

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) and from Science Foundation Ireland (grants 12/CE/I2267 and 13/RC/2106) through the CNGL programme and the ADAPT Centre.

Data Availability Statement: 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 noncommercial license made available by the first two authors at http://multisimo.eu/datasets.html.

Acknowledgments: We are grateful to anonymous reviewers for their constructive responses to an earlier draft of this paper.

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.

References

1. Wood, D.J.; Gray, B. Toward a Comprehensive Theory of Collaboration. *J. Appl. Behav. Sci.* **1991**, *27*, 139–162. [CrossRef]
2. Corbellini, N. Towards Human-Machine Collaboration: Multimodal Group Potency Estimation. In Proceedings of the 2022 International Conference on Multimodal Interaction, London, UK, 21–22 January 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 685–689. [CrossRef]

3. Johnson, D.; Murray, G. Clustering and Multimodal Analysis of Participants in Task-Based Discussions. In *Companion Publication of the 2021 International Conference on Multimodal Interaction, Montreal, QC, Canada, 18–22 October 2021*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 273–277. [CrossRef]
4. Swain, M.; Lapkin, S. Focus on form through collaborative dialogue: Exploring task effects. In *Researching Pedagogic Tasks: Second-Language Learning, Teaching and Testing*; Bygate, M., Skehan, P., Swain, M., Eds.; Routledge: London, UK, 2001; pp. 99–118.
5. Azmitia, M.; Montgomery, R. Friendship, transactive dialogues, and the development of scientific reasoning. *Soc. Dev.* **1993**, *2*, 202–221.
6. Shibani, A.; Koh, E.; Lai, V.; Shim, K.J. Assessing the Language of Chat for Teamwork Dialogue. *J. Educ. Technol. Soc.* **2017**, *20*, 224–237.
7. Olsen, J.K.; Aleven, V.; Rummel, N. Exploring dual eye tracking as a tool to assess collaboration. In *Innovative Assessment of Collaboration*; Springer International Publishing: Cham, Switzerland, 2017; pp 157–172.
8. Larkin, F.; Hobson, J.A.; Hobson, R.P.; Tolmie, A. Collaborative competence in dialogue: Pragmatic language impairment as a window onto the psychopathology of autism. *Res. Autism Spectr. Disord.* **2017**, *43–44*, 27–39. [CrossRef]
9. O'Neil, H.F.; Chuang, S.H.S.; Chung, G.K.W.K. Issues in the Computer-based Assessment of Collaborative Problem Solving. *Assess. Educ. Princ. Policy Pract.* **2003**, *10*, 361–373. [CrossRef]
10. Graesser, A.; Kuo, B.C.; Liao, C.H. Complex Problem Solving in Assessments of Collaborative Problem Solving. *J. Intell.* **2017**, *5*, 10. [CrossRef]
11. Vogel, C.; Koutsombogera, M.; Esposito, A. Aspects of Methodology for Interaction Analysis. In Proceedings of the 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom2020), Mariehamn, Finland, 23–25 September 2020; pp. 141–146. [CrossRef]
12. Bouamrane, M.M.; Luz, S. An analytical evaluation of search by content and interaction patterns on multimodal meeting records. *Multimed. Syst.* **2007**, *13*, 89–102. [CrossRef]
13. Baranyi, P.; Csapo, A. Cognitive infocommunications: CogInfoCom. In Proceedings of the 2010 11th International Symposium on Computational Intelligence and Informatics (CINTI), Budapest, Hungary, 18–20 November 2010; pp. 141–146. [CrossRef]
14. Baranyi, P.; Csapo, A. Definition and Synergies of Cognitive Infocommunications. *Acta Polytech. Hung.* **2012**, *9*, 67–83.
15. Csapo, A.; Baranyi, P. CogInfoCom Channels and Related Definitions Revisited. In Proceedings of the IEEE 10th Jubilee International Symposium on Intelligent Systems and Informatics (SISY), Subotica, Serbia, 15–17 September 2012; pp. 73–78. [CrossRef]
16. Fülöp, I.M.; Csapó, Á.; Baranyi, P. Construction of a CogInfoCom ontology. In Proceedings of the 2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom), Budapest, Hungary, 2–5 December 2013; pp. 811–816.
17. Baranyi, P.; Csapo, A.; Sallai, G. *Cognitive Infocommunications (CogInfoCom)*; Springer: Cham, Switzerland, 2015. [CrossRef]
18. Esposito, A.; Cordasco, G.; Vogel, C.; Baranyi, P. Cognitive Infocommunications. *Front. Comput. Sci.* **2023**, *5*, 1129898. [CrossRef]
19. Esposito, A.; Jain, L.C. Modeling social signals and contexts in robotic socially believable behaving systems. In *Toward Robotic Socially Believable Behaving Systems Volume II—"Modeling Social Signals"*; Esposito, A., Jain, L.C., Eds.; Springer: Cham, Switzerland, 2016; pp. 5–13. [CrossRef]
20. Vogel, C.; Esposito, A. Linguistic and Behavior Interaction Analysis within Cognitive Infocommunications. In Proceedings of the 10th IEEE Conference on Cognitive Infocommunications, Naples, Italy, 23–25 October 2019; pp. 47–52. [CrossRef]
21. Vogel, C.; Esposito, A. Interaction Analysis and Cognitive Infocommunications. *Infocommunications J.* **2020**, *12*, 2–9. [CrossRef]
22. Sudár, A.; Csapó, Á.B. Descriptive Markers for the Cognitive Profiling of Desktop 3D Spaces. *Electronics* **2023**, *12*, 448. [CrossRef]
23. Horváth, I.; Berki, B. Investigating the Operational Complexity of Digital Workflows Based on Human Cognitive Aspects. *Electronics* **2023**, *12*, 528. [CrossRef]
24. Brown, P.; Levinson, S. *Politeness: Some Universals in Language Usage*; Cambridge University Press: Cambridge, UK, 1987.
25. Koutsombogera, M.; Vogel, C. Observing Collaboration in Small-Group Interaction. *Multimodal Technol. Interact.* **2019**, *3*, 45. [CrossRef]
26. Sheridan, S.M. What Do We Mean When We Say "Collaboration"? *J. Educ. Psychol. Consult.* **1992**, *3*, 89–92. [CrossRef]
27. Thomson, A.M.; Perry, J.L. Collaboration Processes: Inside the Black Box. *Public Adm. Rev.* **2006**, 66, 20–32. [CrossRef]
28. Gray, B. *Collaborating: Finding Common Ground for Multiparty Problems*; Jossey-Bass: San Francisco, CA, USA, 1989.
29. Carstairs-McCarthy, A. Synonymy Avoidance, Phonology and the Origin of Syntax. In *Approaches to the Evolution of Language. Social and Cognitive Bases*; Hurford, J.R., Studdert-Kennedy, M., Knight, C., Eds.; Cambridge University Press: Cambridge, UK, 1998; Chapter 17, pp. 279–296.
30. 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.
31. Schütze, C. Thinking About What We Are Asking Speakers to Do. In *Linguistic Evidence: Empirical, Theoretical and Computational Perspectives*; Kepsar, S., Reis, M., Eds.; Studies in Generative Grammar 85; Mouton De Gruyter: Berlin, Germany, 2005; pp. 457–485.
32. Vogel, C.; Koutsombogera, M.; Costello, R. Analyzing Likert Scale Inter-annotator Disagreement. In *Neural Approaches to Dynamics of Signal Exchanges. Smart Innovation, Systems and Technologies*; Esposito, A., Faundez-Zanuy, M., Morabito, F., Pasero, E., Eds.; Springer: Singapore, 2019; pp. 383–393. [CrossRef]

33. Lawani, S.M. Quality, Collaboration and Citations in Cancer Research: A Bibliometric Study [Microform]. Ph.D. Thesis, The Florida State University, Tallahassee, FL, USA, 1980; 412p. Available online: https://nla.gov.au/nla.cat-vn3287291 (accessed on 1 May 2023).
34. Subramanyam, K. Bibliometric studies of research collaboration: A review. *J. Inf. Sci.* **1983**, *6*, 33–38. [CrossRef]
35. Ajiferuke, I.; Burell, Q.; Tague, J. Collaborative coefficient: A single measure of the degree of collaboration in research. *Scientometrics* **1988**, *14*, 421–433. [CrossRef]
36. Staudt, C. Analysis of Scientific Collaboration Networks: Social Factors, Evolution, and Topical Clustering. Ph.D. Thesis. University of the State of Baden-Wuerttemberg, Stuttgart, Germany, 2011.
37. Frey, B.B.; Lohmeier, J.H.; Lee, S.W.; Tollefson, N. Measuring Collaboration Among Grant Partners. *Am. J. Eval.* **2006**, *27*, 383–392. [CrossRef]
38. Elwyn, G.; Frosch, D.; Thomson, R.; Joseph-Williams, N.; Lloyd, A.; Kinnersley, P.; Cording, E.; Tomson, D.; Dodd, C.; Rollnick, S.; et al. Shared Decision Making: A Model for Clinical Practice. *J. Gen. Intern. Med.* **2012**, *27*, 1361–1367. [CrossRef]
39. Dougherty, M.; Larson, E. A Review of Instruments Measuring Nurse-Physician Collaboration. *J. Nurs. Adm.* **2005**, *35*, 244–253. [CrossRef]
40. Weiss, S.; Davis, P. Validity and reliability of the collaborative practice scales. *Nurs. Res.* **1985**, *34*, 299–305. [CrossRef]
41. Gedney, J. Development of an instrument to measure collaboration and satisfaction about care decisions. *J. Adv. Nurs.* **1994**, *20*, 176–182. [CrossRef]
42. Hojat, M.; Fields, S.K.; Veloski, J.J.; Griffiths, M.; Cohen, M.J.M.; Plumb, J.D. Psychometric Properties of an Attitude Scale Measuring Physician-Nurse Collaboration. *Eval. Health Prof.* **1999**, *22*, 208–220. [CrossRef] [PubMed]
43. Scoular, C.; Care, E. A Generalized Scoring Process to Measure Collaborative Problem Solving in Online Environments. *Educ. Assess.* **2019**, *24*, 213–234. [CrossRef]
44. Hesse, F.; Care, E.; Buder, J.; Sassenberg, K.; Griffin, P. A Framework for Teachable Collaborative Problem Solving Skills. In *Assessment and Teaching of 21st Century Skills: Methods and Approach*; Griffin, P., Care, E., Eds.; Springer: Dordrecht, The Netherlands, 2015; pp. 37–56. [CrossRef]
45. Tran, M.Q.; Biddle, R. Collaboration in Serious Game Development: A Case Study. In Proceedings of the 2008 Conference on Future Play: Research, Play, Share, Toronto, ON, Canada, 3–5 November 2008; ACM: New York, NY, USA, 2008; pp. 49–56. [CrossRef]
46. Borden, L.; Perkins, D. Assessing your collaboration: A self evaluation tool. *J. Ext.* **1999**, *37*, 78–83.
47. Healey, P. Communication as a Special Case of Misunderstanding: Semantic Coordination in Dialogue. Ph.D. Thesis, Centre for Cognitive Science, University of Edinburgh, Edinburgh, UK, 1995.
48. Branigan, H.P.; Pickering, M.J.; Cleland, A.A. Syntactic co-ordination in dialogue. *Cognition* **2000**, *75*, B13–B25. [CrossRef]
49. Healey, P.G.T.; Mills, G.J. Participation, Precedence and Co-ordination in Dialogue. In Proceedings of the 28th Annual Conference of the Cognitive Science Society, Vancouver, BC, Canada, 26–29 July 2006; pp. 1470–1475.
50. Reitter, D.; Keller, F.; Moore, J. Computational Modeling of Structural Priming in Dialogue. In Proceedings of the Human Language Technology Conference of the NAACL, New York, NY, USA, 4–9 June 2006; pp. 121–124.
51. Reitter, D.; Moore, J. Predicting Success in Dialogue. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, 23–30 June 2007; pp. 808–815.
52. Howes, C.; Healey, P.G.T.; Purver, M. Tracking Lexical and Syntactic Alignment in Conversation. In Proceedings of the 32nd Annual Conference of the Cognitive Science Society, Portland, OR, USA, 11–14 August 2010; pp. 2004–2009.
53. Colman, M.; Healey, P. The distribution of repair in dialogue. In Proceedings of the Annual Meeting of the Cognitive Science Society, Boston, MA, USA, 20–23 June 2011; Volume 33, pp. 1563–1568.
54. Vogel, C. Attribution of Mutual Understanding. *J. Law Policy* **2013**, *21*, 377–420.
55. Healey, P.G.T.; Purver, M.; Howes, C. Divergence in Dialogue. *PLoS ONE* **2014**, *9*, e98598. [CrossRef]
56. Reverdy, J.; Vogel, C. Measuring Synchrony in Task-Based Dialogues. In Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH2017), Stockholm, Sweden, 20–24 August 2017; pp. 1701–1705. ISSN 2308-457X.
57. Reverdy, J.; Vogel, C. Linguistic Repetitions, Task-based Experience and A Proxy Measure of Mutual Understanding. In Proceedings of the CogInfoCom 2017, Debrecen, Hungary, 11–14 September 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 395–400.
58. Reverdy, J.; Hayakawa, A.; Vogel, C. Alignment in a Multimodal Interlingual Computer-Mediated Map Task Corpus. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018; Koiso, H., Paggio, P., Eds.; European Language Resources Association (ELRA): Paris, France, 2018; pp. 55–59.
59. Kendall, M.G. A new measure of rank correlation. *Biometrika* **1938**, *30*, 81–93. [CrossRef]
60. 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.; IEEE: Piscataway, NJ, USA, 2018; pp. 141–146.
61. Murray, G. Information Processing and Overload in Group Conversation: A Graph-Based Prediction Model. *Multimodal Technol. Interact.* **2019**, *3*, 46. [CrossRef]
62. John, O.P.; Donahue, E.M.; Kentle, R.L. *The Big Five Inventory Versions 4a and 54*; Technical Report; Institute of Personality and Social Research, University of California: Berkeley, CA, USA, 1991.

63. John, O.P.; Naumann, L.P.; Soto, C.J. Paradigm shift to the integrative big five trait taxonomy. In *Handbook of Personality: Theory and Research*; The Guilford Press: New York, NY, USA, 2008; Volume 3, pp. 114–158.
64. Riviello, M.T.; Chetouani, M.; Cohen, D.; Esposito, A. On the Perception of Emotional "Voices": A Cross-Cultural Comparison among American, French and Italian Subjects. In *Analysis of Verbal and Nonverbal Communication and Enactment: The Processing Issues*; Springer: Berlin/Heidelberg, Germany, 2011; Volume LNCS6800, pp. 368–377.