The Group Affect and Performance (GAP) Corpus

McKenzie Braley
University of the Fraser Valley
Abbotsford, Canada
mckenzie.braley@student.ufv.ca

Gabriel Murray
University of the Fraser Valley
Abbotsford, Canada
gabriel.murray@ufv.ca

ABSTRACT

In this paper, we present the Group Affect and Performance (GAP) corpus, a publicly available dataset of thirteen small group meetings. The GAP corpus contains meeting audio, transcriptions, annotations, decision-making performance, as well as group member influence, post-meeting ratings of satisfaction, and demographics. In this paper, we discuss all aspects of data collection and preparation. We also present preliminary analyses and findings concerning decision-making performance, group member influence, group member satisfaction, and additional meeting characteristics. We conclude with future directions. In creating and releasing this corpus, it is our goal to stimulate research on the computational analysis of small group meetings, and to supplement the relatively small amount of currently available group interaction data.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; *Machine learning approaches*; • **Human-centered computing**;

KEYWORDS

group interaction, multimodal corpora, group decision-making

ACM Reference Format:
McKenzie Braley and Gabriel Murray. 2018. The Group Affect and Performance (GAP) Corpus. In *Proceedings of Group Interaction Frontiers in Technology (GIFT'18).* ACM, New York, NY, USA, 9 pages.

1 INTRODUCTION

The social aspects of humanity have traditionally been explored in the social sciences. However, with advances in artificial intelligence and machine learning, computational approaches can now complement and further progress this area of study. Fields such as social signal processing (SSP) [31], affective computing [9, 26], and natural language processing (NLP) [23] apply computational tools to the study of social science topics. Specifically, in these areas, verbal and nonverbal cues are fed into machine learning models which output meaningful predictions about a range of social phenomena. Automated analyses offer quick, efficient, and enticing alternatives to the traditional observational and manual coding techniques [17].

Of specific interest to this paper are the significant advances presently being made in the automated analysis of small group dynamics. Small group dynamics have been mostly studied by SSP researchers, who have successfully used machine learning models to study a variety of groups-related topics, including competitiveness [6], affect [15, 29], agreement [4, 8, 10], leadership [25, 27], and cohesion [11] in small group interactions. In each of these studies, nonverbal cues extracted from small group meetings predict the phenomenon of interest with significant rates of accuracy. While the SSP small groups research includes an explicit focus on nonverbal cues, there has been a smaller amount of work in the field of NLP that focuses on verbal cues in small group meetings (i.e., meetings composed of three or more members) [7, 20] and dyadic meetings (i.e., meetings composed of two members) [24]. Both veins of research are useful for a number of reasons, including understanding group dynamics and facilitating successful group meetings.

The development of machine learning models that accurately assess group dynamics, and thus the continued progression of this area of research, is dependent upon large amounts of available group data. Therefore, large corpora of small group interactions must be published and made available to researchers [31]. However, due to the expensive and time-consuming nature of data collection and preparation, there exists a limited number of small groups corpora. To address this gap and to stimulate small groups research, we have created a corpus of recorded, transcribed, and annotated small group meetings. We also aim to create a corpus that is amenable to both SSP and NLP, so that nonverbal and verbal cues may be assessed in relation to small group dynamics. In this paper, we present the newly created corpus of small group interaction data, the Group Affect and Performance (GAP) corpus [1].

The structure of this paper is as follows. In Section 2, we discuss related work and corpora. In Section 3, we describe the data collection steps of the GAP corpus, while in Section 4 we describe further preparation and annotation phases. In Section 5, we present preliminary analyses and results. Finally, in Section 6, we conclude and discuss future directions.

2 RELATED WORK

Presently, there exist several available small groups corpora. However, most have mere tangential relations to the GAP corpus and were not specifically developed to stimulate automated small group analyses. In this section, we start by focusing on the few directly related copora that were specifically designed to facilitate computational approaches to studying small group dynamics.

First, the Emergent Leadership (ELEA) corpus [25] is a multimodal dataset that consists of 40 audio- and video-recorded group meetings. In the ELEA corpus meetings, group members discuss a hypothetical survival scenario (described in detail in section 3.2). Briefly, group members must imagine that they have been in a plane crash and that they have salvaged a number of items from the plane. Their task is to rank the items in order of importance to their survival in that situation. The corpus consists of individual and group decision-making performance, individual influence on group performance, self-reports of personality, and group members' perceptions of the other group members. Perceptions include perceived

[1] https://sites.google.com/view/gap-corpus/home

leadership, dominance, competence, and likability. Finally, the corpus consists of automatically extracted nonverbal features, such as speaker segmentation, turn-taking features, prosody, body activity, and head activity. The ELEA corpus has facilitated many SSP studies on small group dynamics. For instance, the dataset has been used to successfully detect group decision-making performance [3], leadership, dominance [25], and personality [2], primarily from nonverbal cues. Because the ELEA corpus is quite small, containing a total of 40 meetings, the field would benefit from an increased amount of available data. In creating the GAP corpus, we replicated most aspects of the ELEA corpus, with the aims of supplementing the available ELEA data.

The Mission Survival (MS-2) corpus [19], consisting of 13 4-person meetings, also involves discussions of a hypothetical survival scenario. The corpus consists of audio- and video-recordings of the group meetings, individual and group decision-making performance, self-reports of locus of control, self-reports of extraversion, and group member ratings of group cohesion. Annotations include speech activity (.e.g., voice and pitch) and body activity (e.g., fidgeting and head orientation). This corpus has primarily stimulated research on automated multimodal personality detection in small group meetings [18, 19].

The SSPNet Mobile Corpus [22] is similar to the previously discussed corpora, as group members work together on a survival scenario. However, the main difference pertains to the channel of communication. While the ELEA and MS-2 corpora consist of face-to-face meetings, the SSPNet Mobile corpus is comprised of telephone conversations. Specifically, the SSPNet corpus consists of 60 dyadic telephone conversations concerning the survival scenario task. Following the phone conversation, participants provided self-report ratings of personality, conflict-handling style, and interpersonal attraction to their conversational partner. Annotations of the conversations include the nonverbal cues laughter, fillers, and backchannels. Analyses using the SSPNet Mobile Corpus have primarily focused on how people convey nonverbal information over the telephone, when visual communicative cues, such as gestures and facial expressions, are absent. As an example, in [30], the temporal distribution of nonverbal cues was assessed in relation to gender, topic, functional role, and conflict-handling style. Similar work is conducted in [22].

As previously mentioned, there are several small groups corpora that were designed for purposes extraneous to automated small group analysis. Even so, the automated small groups literature is rife with studies that have used these indirectly related corpora. As an example, although the Augmented Multiparty Interaction (AMI) corpus [5] is used in several studies to assess small group dynamics, it was originally developed for speech recognition and computer vision purposes. It can be inferred that this is a reflection of the unfortunate lack of small groups corpora in the SSP and related fields. Vinciarelli et al., in [31], explain that ecological validity and generalization are hindered when corpora are not used for their original purposes. Vinciarelli et al. further explain the necessity of generating more corpora specifically for SSP and related research purposes. Nonetheless, indirectly related corpora, such as the AMI corpus and International Computer Science Institute (ICSI) corpus [12], have played important roles in stimulating small groups research and our understanding of small group dynamics. We subsequently provide overviews of these two corpora.

The AMI corpus is a multimodal dataset that consists of approximately 100 hours of meeting recordings. The corpus consists of both naturally occurring meetings and meetings involving discussions of hypothetical scenarios. In the scenario meetings, group members enact arbitrarily assigned roles and discuss remote control designs for a fictitious company. Following each scenario meeting, group members provided ratings of the meetings' affective outcomes, which include indices of leadership and cohesiveness. The recordings have been segmented and transcribed, and include verbal and nonverbal annotations, such as dialogue acts, topics, sentiment, body movement, and head movement. The corpus has been successfully utilized by several researchers to study a variety of groups-related topics, such as the associations between turn-taking and affective outcomes [15], as well as nonverbal cues and dominance [13]. As an example, in [13], total speaking length was used to detect evidence of dominance with a classification accuracy of 85%.

The ICSI corpus consists of 75 group meetings of approximately one hour in length. Unlike the AMI corpus, the meetings are composed of the computer scientists of the institute and are entirely naturally occurring, with no hypothetical scenario-based discussions. Because the meetings are naturally occurring, discussions pertain to specialized and technical computing-related topics, such as NLP and neural networks. The corpus consists of meeting audio, transcriptions, as well as group member characteristics such as gender, native language, level of education, and age. Annotations of the meetings include false starts, pauses, and backchannels. In [8] and [10], data from the ICSI corpus were used to detect agreement and disagreement with robust classification accuracies. Specifically, using an unsupervised learning approach, word-based cues such as positive utterances, negative utterances, and backchannels detected evidence of agreement and disagreement with a classification accuracy up to 82% [8]. In [10], using a supervised learning approach based on adjacency pairs incorporating contextual information, a classification accuracy of 86.9% was achieved.

To sum, there has been a substantial increase in automated analyses of small groups dynamics in recent years, in part owing to the published and available copora of small group interactions. It is our hope that the GAP corpus will contribute to these trends of research. In subsequent sections, we present and describe in detail the GAP corpus.

3 THE GAP CORPUS: DATA COLLECTION

The GAP corpus consists of a total of 13 group meetings and 104.45 minutes of meeting recordings. In each meeting, two to four group members sat around a table while performing an audio- and video-recorded group decision-making task. Group members first completed an individual version of the task, then performed the recorded group task, and finally responded to a series of questions related to demographics and the group meeting.

In this section, we describe all aspects of data collection, including the participants, group task, post-task questionnaire, recording set-up, and procedure.

Label	Item
Time Expectation	(1) "This task took longer than expected to complete."
Worked Well Together	(2) "Our group worked well together."
Time Management	(3) "Our group used its time wisely."
Efficiency	(4) "Our group struggled to work together efficiently on this task."
Overall Quality of Work	(5) "Overall, our group did a good job on this task."
Overall Satisfaction	Items one to five combined and averaged.
Leadership	(6) "I helped lead the group during this task."

Table 1: Post-Task Questionnaire Items

3.1 Participants

A total of 37 participants (26 females and 11 males) made up five groups of two, five groups of three, and three groups of four. Participants were recruited through the university's online sign-up system or through class emails sent out by the researchers, and either participated in exchange for course credit or as volunteers. Because participants were recruited through the university, all participants are undergraduate students. The use of undergraduate university students ensures that participants are demographically similar in terms of age, level of education, and socioeconomic status. Most participants were lower-level students (mean year at the university was 1.9). Finally, participants consisted of both native and non-native English speakers, although the majority were native English speakers (31 native and 6 non-native English speakers).

3.2 The Winter Survival Task

The winter survival task (WST) is a group decision-making exercise that consists of a hypothetical plane crash scenario. Participants are presented with 15 items that they have salvaged from the plane. Examples of items include a compress kit, a cigarette lighter without the fluid, a compass, and a family-sized chocolate bar. Participants must rank each item according to its importance to their survival in that situation. As explained by [14], the WST is a commonly used exercise in social psychology and organizational behavior research to measure decision-making, leadership, and social ability. Moreover, the task has also been used in computing research to study group roles, personality [21], and leadership [25]. Although this is primarily a group decision-making exercise, the participants in this study did the task both individually and as a group.

3.3 Post-Task Questionnaire

Group members filled out a post-task questionnaire containing questions related to basic demographics and the group meeting. They first provided their year at the university, gender, and whether English is their native language. They then responded on five-point Likert scales to how strongly they agreed with statements concerning the meeting. As shown in Table 1, the items specifically concerned (1) whether their expectations of the task's temporal length were met, (2) how well they thought they worked together as a group, (3) time management, (4) efficiency, (5) overall quality of group work, (6) and leadership. In addition to examining each item separately, the first five items were averaged to yield an overall satisfaction with group score. Items one to five were also assessed at the individual level (i.e., responses of the individual group members) as well as at the group level (i.e., average of the group member responses in each meeting). Item six on leadership was the only item excluded from the overall satisfaction score and the group averages. The item is a measure of self-perceived leadership and is not a facet of satisfaction per se, as are the other items. Moreover, the response is informative of the individual group member and not the group as a whole, and thus, we did not derive a group leadership score.

3.4 Recording Set-Up

To record audio, we used the Zoom H1 Handy Recorder, a portable and professional quality audio-recording device. One audio-recorder per meeting was placed directly in the center of the group members. To record video, Logitech HD Webcam C270s were used. One video-recording device was placed directly in front of each group member, capturing upper body views. The recording devices were input into an HP Pavilion laptop, which recorded the meetings using Open Broadcaster Software (OBS) Studio.

3.5 Procedure

First, this study and the following procedures were approved by the university's Human Research Ethics Board. A time was arranged to conduct the study on the university campus. In the data collection phase, participants met in groups of two to four. It should be noted that dyads are quite distinct from small groups in a number of ways, as discussed in detail in [17]. A corpus of both dyadic and small group meetings allows further investigation into the different dynamics of interaction. We will therefore explore the differences between the meeting types in our analyses. At the outset, participants were made aware that they were to be recorded during the group task phase and that the recordings were to be made available for research purposes. All participants explicitly gave consent to be recorded and to have the recordings published. Following informed consent procedures, each group member was given five minutes to complete the WST individually. Group members were then given 15 minutes to complete the WST as a group; they specifically collaborated, discussed their answers on the individual rankings, and came up with one group ranking. The group exercise was audio- and video-recorded. Although participants were given 15 minutes maximum, length of recorded group discussion ranges from 2.38 to 12.65 minutes (M=8.2, SD=3.25). Then, participants individually filled out the post-task questionnaires. Finally, they were debriefed and thanked for their participation.

4 THE GAP CORPUS: DATA PREPARATION

In this section, we describe the data preparation efforts. We discuss the specific details of segmentation, transcription, annotation, coding, and scoring.

4.1 Transcription & Annotation

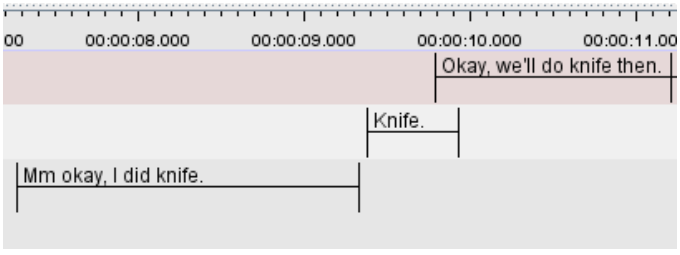
Speech was segmented, transcribed, and annotated using ELAN transcription and annotation software. ELAN is a freely available, professional tool used to annotate audio and/or video recordings. Its multi-tier annotation system allows annotations to be created at multiple levels in an organized hierarchy. This allows the efficient creation of various annotations per speech segment.

Segmentation. Spontaneously produced speech does not divide neatly into individual sentences as does written text. Humans can speak continuously for several minutes without clear boundaries indicating separate sentences. Instead, speech consists of speaker intentions. Each intention represents the intent of a speaker to communicate one piece of information. A stream of speech can therefore be segmented into several separate intentions. Human annotators manually segmented each meeting according to individual speaker intentions. Segments were time-aligned to the recording, meaning that the start and end times of a segment correspond to the temporal beginning and ending of the utterance. The segmentation instructions given to the annotators were based on the AMI corpus dialogue act segmentation instructions[2], though we did not annotate for specific dialogue act types in the GAP corpus.

Transcription. Following segmentation, human annotators manually transcribed each segment verbatim. Speech was transcribed word for word, including grammatical errors, false starts, stutters, and pauses. Symbols are used to represent non-speech information: - - represents false starts, - indicates stuttering, **...** shows pauses, **$** means laughter, **%** represents coughing, and # indicates another noise.

Figure 1 is an excerpt from one group meeting, showing how the files were segmented and transcribed using ELAN.

Figure 1: Segmentation and Transcription in ELAN



Sentiment Annotations. Meetings were annotated for sentiment, which refers to the underlying emotion of the speech segments. To annotate sentiment, we chose a binary annotation scheme. That is, the two sentiment values are positive sentiment and negative sentiment. This is because we are interested in how positive and negative sentiment expressed during the meeting relates to some of the post-task questionnaire ratings, as well as to the decision-making process. A finer-grained annotation, such as a scale of sentimental values, was not necessary for this. Positive sentiment annotations characterize utterances that have an underlying positive emotion, e.g., "That sounds great" or "I am happy that we ranked knife as one". Negative sentiment annotations characterize utterances that have an underlying negative emotion, e.g., "That sucks" or "That's a bad idea". Two independent researchers annotated each meeting. On a subset of five meetings, inter-annotator agreement was calculated using Cohen's kappa. For sentiment annotations, Cohen's kappa is 0.73, a satisfactory reliability value.

We calculated the frequencies of positive and negative annotations to determine how often group members produced affect-related utterances. In order to control for the total number of utterances in each meeting, sentiment annotation frequencies were calculated as a proportion of the number of sentiment annotations out of the total number of utterances in the meeting. Proportions are as follows: positive sentiment (M = 5.45, SD = 5.64), and negative sentiment (M = 2.08, SD = 2.19). These show that on average, five percent of utterances in a meeting have underlying positive emotions, while two percent of utterances in a meeting have underlying negative emotions.

Group Decision-Making Annotations. Meetings were also annotated for group decision-making, which refers to decisions regarding the ranking of an item. The four group decision-making annotation values are proposal, agreement, disagreement, and confirmation. We chose these four values because we are interested in the entire decision-making process, from proposal, to agreement or disagreement, and to the final confirmation. This allows an examination of the different phases of group decision-making and a detailed analysis of how successful decisions form. Proposal refers to statements where a group ranking is proposed, e.g., "I think we should do the knife as one". Agreement refers to statements whereby a group member agrees with a proposal, e.g., "I agree with knife being one". Disagreement refers to statements whereby a group member disagrees with a proposal, e.g., "I do not think knife should be one". Finally, confirmation refers to utterances whereby an already established decision is confirmed, e.g., "So we decided to put knife as one". Again, two independent human annotators were used per meeting. Inter-annotator agreement was calculated based on a subset of five meetings. For group decision-making annotations, Cohen's kappa is 0.7, which is also considered a satisfactory reliability value.

Proportions of each type of decision-making annotation are also calculated. Mean proportions are as follows: proposal (M = 8.36, SD = 3.74), agreement (M = 7.61, SD = 2.53), disagreement (M = 1.17, SD = .89), and confirmation (M = 1.95, SD = 1.68).

4.2 Coding & Scoring

Meeting Codes. To ensure anonymity and to keep track of each group member's data, codes were assigned and used in each stage of data collection and preparation. First, each group meeting is identified with a number based on the order of when the data was collected, i.e., one to thirteen. Meetings are also coded with the date and time of data collection.

[2] http://groups.inf.ed.ac.uk/ami/corpus/Guidelines/dialogue_acts_manual_1.0.pdf

Group Member Codes. Identification codes (IDs), i.e., one of five colors (blue, green, pink, orange, or yellow), were arbitrarily assigned to each group member. The IDs are used to identify the participants in the recordings, their utterances, as well as their task and questionnaire data. A sticker of the corresponding color was thus placed on the group members' shirts in the frame of the camera as well as their individual WST sheets and post-task questionnaires.

Text Codes. In a text file of the transcriptions, each utterance is labeled with the group member's ID, a number that represents the number of utterances produced by that group member, e.g., "Blue.1, Blue.2, Blue.3, Green.1, Green.2, Green.3", and the start and end timestamps of the segment. Sentiment and group decision-making annotations are each contained in separate text files. In the text files, each annotation contains the ID, number, and start and end timestamps that correspond to the annotated utterance. Figure 2 shows the coding of transcripts in the text files. Both Figure 1 and Figure 2 contain data from group meeting 1.

Figure 2: Coding of Transcripts.

```
Pink.1          So, what did everyone do as one?
                00:00:02.014 - 00:00:03.538

-------------------------------------------------

Blue.1          I did, uh, cigarette lighter.
                00:00:04.000 - 00:00:05.667

-------------------------------------------------

Blue.2          For one.
                00:00:06.420 - 00:00:07.168

-------------------------------------------------

Pink.2          Mm okay, I did knife.
                00:00:07.275 - 00:00:09.333
```

WST Scoring. Following the procedures used by [25], we derived three scores from the WST: absolute individual score (AIS), absolute group score (AGS), and absolute individual influence (AII). AIS was calculated by summing the differences between the group member ranking and the expert ranking for each item. Lower AIS reveals greater similarity to the expert ranking and thus greater decision-making performance. AGS was calculated by summing the differences between the group ranking and the expert ranking for each item. Again, lower AGS reveals greater similarity to the expert ranking and thus greater decision-making performance. Finally, AII was calculated by summing the differences between the group member ranking and the group ranking. Lower AII reveals greater similarity to the group ranking and therefore greater influence on the group. Figures 3 and 4 show the distributions of AIS and AGS scores, respectively.

Post-Task Questionnaire Scoring. From the six post-task questionnaire responses regarding the group meeting, seven scores are derived: items one to six as well as the overall satisfaction score. It should also be noted that some items were reverse scored (items one and four) so that, for all items, greater scores reflect positive views of the meeting. The first five items and the overall satisfaction score were calculated both at the individual- and group-level. At the individual-level, scores are derived for each group member. At the group-level, the average of the group member responses for that item is assigned to each group. As stated previously, the item on leadership was excluded from the overall satisfaction score and group averages. The resulting scores are: Time Expectation (GroupTE & IndTE), Worked Well Together (GroupWW & IndWW), Time Management (GroupTM & IndTM); Efficiency (GroupEf & IndEf), Overall Quality of Work (GroupQW & IndQW), Overall Satisfaction (GroupSat & IndSat), and Leadership (IndLead).

Figure 3: Histogram of Absolute Individual Scores

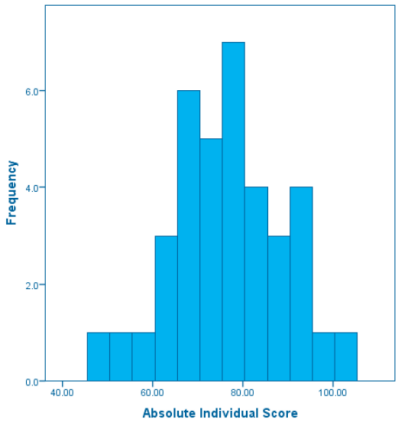
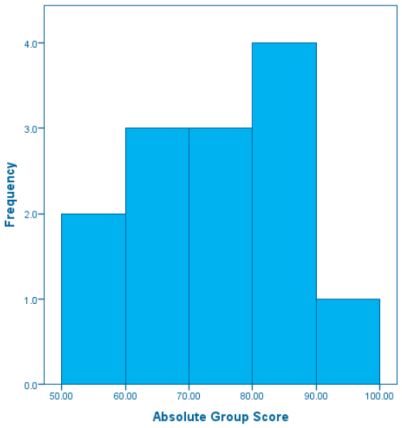


Figure 4: Histogram of Absolute Group Scores



5 PRELIMINARY ANALYSES

Although machine learning analyses will be performed in subsequent papers using verbal and nonverbal cues, we start by discussing some preliminary analyses performed with the WST and

post-task questionnaire data, as well as participant and meeting characteristics.

5.1 Descriptive Statistics

In Table 2 and 3, descriptive statistics for the WST data, post-task questionnaire data, and meeting length (denoted as MLength) are provided. Table 2 shows the individual-level data (with 37 data points for each participant) and table 3 shows the group-level data (with 13 data points for each group).

Variable	M	SD	Min	Max
AIS	76.84	12.15	48	102
AII	37.22	15.57	12	76
IndTE	4.11	.84	2	5
IndWW	4.49	.61	3	5
IndTM	4.59	.55	3	5
IndEf	4.49	.9	1	5
IndQW	4.38	.68	3	5
IndSat	4.42	.51	2.8	5
IndLead	3.43	.93	2	5

Note. N=37.

Table 2: Descriptive Statistics for Individual-Level Winter Survival Task Data and Post-Task Questionnaire Responses

Variable	M	SD	Min	Max
AGS	72.85	11.55	50	90
GroupTE	4.11	.75	2.5	5
GroupWW	4.46	.32	4	5
GroupTM	4.58	.31	4	5
GroupEf	4.43	.95	1.5	5
GroupQW	4.35	.38	3.75	5
GroupSat	4.38	.43	3.3	4.87
MLength	8.2	3.25	2.38	12.65

Note. N=13.

Table 3: Descriptive Statistics for Group-Level Winter Survival Task Data and Post-Task Questionnaire Responses

5.2 WST Data & Post-Task Responses

First, Spearman's rank correlation coefficients were used to examine the possible associations between the WST data and the post-task questionnaire responses. Table 3 shows the inter-item correlations for all variables. Among the correlations, there are a few worth noting. We first report the group-level correlations. AGS was negatively correlated with GroupTE (r_s = .34, p = .04) and GroupEf (r_s = -.44, p = .01), showing that greater group decision-making performance is associated with post-task responses that the task did not take longer than expected and that the group worked efficiently together. Our next set of correlations concerns the individual group member responses. IndLead was positively correlated with IndWW (r_s = .4, p = .02), IndTM (r_s = .4, p < .02), and IndSat (r_s = .33, p < .05). These correlations show that group members who perceived themselves as leaders thought that the group worked better together, used its time more wisely, and had greater levels of overall satisfaction with the meeting. Interestingly, IndLead was not significantly correlated with AAI (p > .05), showing that group members who perceived themselves as leaders did not necessarily convince the group to adopt their individual ratings.

5.3 Participant & Group Characteristics

In addition to correlations between between WST data and post-task responses, we also performed analyses to examine any associations with participant characteristics (i.e., gender) and group characteristics (i.e., individual vs. group performance, meeting length, and meeting type).

To examine the effects of gender and meeting type on WST data and post-task responses, a 2 (gender: male & female) x 2 (meeting type: dyadic & small groups) multivariate Analysis of Variance (ANOVA) was used. First, we eliminated some dependent variables from the analysis due to high multicollinearity. We specifically eliminated GroupTE, GroupQW, and GroupSat because of the substantial correlations with other variables (i.e., r_s > .80). Bonferroni corrections were also applied. Results indicated that there was no main effect of gender (F (13, 21) = 1.26, p = .31, η^2 = .44), no main effect of meeting type (F (13, 21) = 1.07, p = .43, η^2 = .4), and no gender by meeting type interaction (F (13, 21) = 1.26, p = .12, η^2 = .52). The strong effect sizes, as indicated by the partial eta squared values, reveal that the non-significant effects are likely due to low statistical power. Thus, it is possible that significant differences between genders and meeting types will be found with a larger sample size.

We also sought to examine whether performance is best when participants worked individually or as a team. A paired samples t-test was used to examine the differences between AGS and AIS values. Results revealed no significant difference between the two values (p > 0.05), showing that participants performed to similar degrees when they worked on the WST individually as they did as a group.

We also examine correlations between meeting length, WST data, and post-task questionnaire responses. MLength formed negative correlations with group satisfaction (r_s = -.34, p = .04), showing that as meeting length increases, satisfaction with the meeting decreases. Meeting length was also negatively correlated with time expectations, both at the individual level (r_s = -.56, p < .001) and the group level (r_s = -.55, p < .001). Unsurprisingly, as meeting length increased, participants became more likely to endorse the item that the task took longer than expected.

6 CONCLUSIONS & FUTURE DIRECTIONS

In this paper, we present a newly created corpus of small group interaction data, the GAP corpus. We first collected recordings of group meetings and then prepared the data for publication and availability. As described in detail in this paper, we have included meeting audio, meeting transcriptions, sentiment annotations, group decision-making annotations, individual decision-making performance, group decision-making performance, individual influence on group, as well as post-task questionnaire data related to demographics and satisfaction with the group meeting. It should also

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. AIS	-																
2. AGS	.37*	-															
3. AII	-.06	-.2	-														
4. GroupTE	-.25	-.34*	.08	-													
5. IndTE	-.23	-.26	.15	.79**	-												
6. GroupWW	.2	-.15	.07	.28	.29	-											
7. IndWW	- .01	-.07	.02	.09	.17	.5**	-										
8. GroupTM	.16	-.15	.11	.62**	.51**	.69**	.33*	-									
9. IndTM	.16	-.13	.24	.34*	.27	.38*	.54**	.53**	-								
10. GroupEf	-.13	-.44**	.07	.43**	.36*	.49*	.22	.51**	.3	-							
11. IndEf	-.23	-.35*	.12	.35*	.36*	.03	.38	.33*	.37*	.74**	-						
12. GroupQW	.24	.03	.03	.55*	.49**	.86**	.42**	.88**	.47**	.44**	.29	-					
13. IndQW	.25	.02	.19	.29	.31	.48**	.45**	.49**	.7**	.2	.29	.57**	-				
14. GroupSat	.05	-.29	.1	.73**	.62**	.78**	.37*	.89**	.48**	.47**	.47**	.91**	.5**	-			
15. IndSat	-.1	-.28	.23	.59**	.66**	57**	.63**	.63**	.76**	.48**	.65**	.64**	.7**	.73**	-		
16. IndLead	.04	-.07	-.06	.01	.02	.02	.4*	.11	.4*	-.12	.07	.06	.22	.04	.33*	-	
17. MLength	.01	-.03	.12	-.55**	-.56**	-.02	.04	-.29	-.14	-.24	-.22	-.25	-.13	-.3	-.34*	.04	-

Note. N=37. $^{*}p < .05$. $^{**}p < .01$.

Table 4: Inter-item Correlations for Winter Survival Task Data and Post-Task Questionnaire Responses

be noted that while meeting audio will be released, due to ethical considerations, video will not be published. Because there is currently a paucity of available small groups corpora, it is our hope that the GAP corpus will address this gap and stimulate research on the computational analysis of small groups. It will also supplement valuable existing resources such as the ELEA corpus.

Preliminary analyses showed important relationships between WST performance, post-task ratings, and group characteristics. A few findings are worth expanding upon. First, we found that greater WST performance is associated with post-task responses that the task did not take longer than expected and that the group worked efficiently together. This finding is likely due to the associations between group efficiency and group decision-making performance. When groups work efficiently together, their ability to make successful group decisions improves. It should be noted that we assessed group member perceptions of group efficiency; therefore, based on our findings, the key factor is how the group members perceive the efficiency of their group. It is possible that objective measures of efficiency, such as researcher-ratings, do not coincide with group member self-reported efficiency. Regardless, we show that how a group member perceives their group's efficiency is informative of the group's ability to make successful decisions.

We also show that self-perceptions of leadership are associated with greater satisfaction with the group meeting. It is possible that people enjoy taking the lead and are thus more satisfied with meetings when they have perceived themselves as leaders. Another possible explanation stems from the basic human egocentric bias. As humans, we are intrinsically motivated to cast ourselves in a positive light and to take credit for our successes [16, 28]. Because self-perceptions of leadership also correlated with responses that the group worked well together and used its time wisely, it is possible that group members rated themselves as leaders as a means to take credit for the perceived group efficiency and success.

We also found a non-significant correlation between self-perceived leadership and AII on the group WST. In order words, those who perceived themselves as leaders did not necessarily convince the group to adopt their individual WST ratings. Sanchez-Cortes et al. similarly found a non-significant association between leadership and influence on the group WST ratings [25] . However, they did find that dominance was positively associated with influence on the group WST ratings. This first suggests that group members use additional factors, beyond whether they convinced the group to adopt their individual WST ratings, to determine whether they led a task. It additionally suggests that leadership and dominance are separate entities, with leaders accepting the ideas of their fellow group members and dominant group members focusing solely on their own ideas.

A final finding to note is the lack of differences between meeting types in WST performance, post-task responses, and meeting lengths. This finding is perhaps counter-intuitive and diverges from our previous claim that dyads and small groups are indeed distinct. However, we cannot conclude from this evidence alone that the dyads and small groups experienced equivalent social dynamics. First, the lack of statistical power likely prevented significant results from being found. Moving forward, we plan to collect more meetings to increase our sample size and power. We specifically plan to record and release an additional seven meetings, to reach a total of 20 meetings. It is also likely that upon examination of micro-level communicative cues, various differences will be found. As explained in [17], a major difference between the meeting types pertains to trends in communication. In dyads, conversational partners are focused on each other, with communication directly channeled back and forth. However, in small groups, communication takes many different forms, with different trends of back-channeling and turn-taking. As an example, comments may be directed toward all group members, one group member, or another group member. Eye

gazes may also be directed at one person to a greater extent than the other group members. Put simply, small groups are much more complex than dyads [17]. Thus, it is possible that any differences between dyads and small groups did not manifest in our presently analyzed data, but will manifest in future analyses on extracted verbal and nonverbal cues. Therefore, in future work, after extracting verbal and nonverbal cues, we will also examine meeting type differences.

In future work, we will use NLP for understanding and predicting small groups-related phenomena, such as decision-making performance and group member satisfaction. To date, automated research on small groups has been mostly conducted in the SSP realm. SSP has an explicit focus on nonverbal cues, such as turn-taking features and movement, which are undoubtedly vital for understanding and predicting small groups-phenomena. However, the result is a sparsity of research that uses verbal cues for understanding and predicting small group dynamics. In order words, NLP and small group dynamics is a relatively neglected area. That being said, from results of previous studies, it can be inferred that NLP can be extremely useful in this area and thus merits an increased focus in the small groups literature. For example, Murray and Oertel [20] show that linguistic features can be very useful for predicting group performance with the ELEA corpus. Reitter and Moore [24] have found a relationship between linguistic alignment and task success using the MapTask corpus [1]. Therefore, in future work, we will extract linguistic features from the transcripts, such as lexical, sentiment, part-of-speech, and syntactic features. We will explore the use of various machine learning models to predict decision-making performance, the annotated phases group decision-making, group member reported satisfaction, and annotated group member sentiment using the extracted linguistic features.

Moreover, although outside the scope of NLP, we will also extract and exploit the use of additional nonverbal features, such as acoustics (e.g., pitch, energy, and loudness), turn-taking, and movement features.

To conclude, the GAP corpus will contribute to the literature on the automated and computational analysis of small group dynamics, progressing our knowledge of small groups and successful group meetings. We specifically aimed to replicate an existing corpus, the ELEA corpus, in order to supplement the available meeting data. Our corpus additionally contains novel meeting aspects, such as annotated group decision-making phases. This will allow a finer grained analysis of how successful, or non-successful, decisions form. Further, future analyses using the corpus data will also demonstrate the role of NLP for understanding and predicting small groups-related phenomena. Finally, we demonstrate the vital role of computational techniques for the social sciences generally and for small group dynamics in particular.

Acknowledgement Both authors were supported by an NSERC Discovery Grant.

REFERENCES

[1] Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The HCRC map task corpus. *Language and speech* 34, 4 (1991), 351–366.

[2] Oya Aran and Daniel Gatica-Perez. 2013. One of a kind: Inferring personality impressions in meetings. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 11–18.

[3] Umut Avci and Oya Aran. 2016. Predicting the performance in decision-making tasks: From individual cues to group interaction. *IEEE Transactions on Multimedia* 18, 4 (2016), 643–658.

[4] Konstantinos Bousmalis, Marc Mehu, and Maja Pantic. 2013. Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools. *Image and Vision Computing* 31, 2 (2013), 203–221.

[5] Jean Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation* 41, 2 (2007), 181–190.

[6] Shammur Absar Chowdhury and Giuseppe Riccardi. 2017. A Deep Learning approach to modeling competitiveness in spoken conversations. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 5680–5684.

[7] Arash Eshghi and Patrick GT Healey. 2016. Collective contexts in conversation: Grounding by proxy. *Cognitive science* 40, 2 (2016), 299–324.

[8] Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. 2004. Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 669.

[9] Hatice Gunes and Björn Schuller. 2013. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing* 31, 2 (2013), 120–136.

[10] Dustin Hillard, Mari Ostendorf, and Elizabeth Shriberg. 2003. Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2*. Association for Computational Linguistics, 34–36.

[11] Hayley Hung and Daniel Gatica-Perez. 2010. Estimating cohesion in small groups using audio-visual nonverbal behavior. *IEEE Transactions on Multimedia* 12, 6 (2010), 563–575.

[12] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The ICSI meeting corpus. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, Vol. 1. IEEE, I–I.

[13] Dinesh Babu Jayagopi, Hayley Hung, Chuohao Yeo, and Daniel Gatica-Perez. 2008. *Modeling dominance in group conversations using nonverbal activity cues*. Technical Report.

[14] Jill Kickul and George Neuman. 2000. Emergent leadership behaviors: The function of personality and cognitive ability in determining teamwork performance and KSAs. *Journal of Business and Psychology* 15, 1 (2000), 27–51.

[15] Catherine Lai, Jean Carletta, and Steve Renals. 2013. Modelling participant affect in meetings with turn-taking features. In *Proc. Workshop of Affective Social Speech Signals*.

[16] Mark R Leary. 2007. Motivational and emotional aspects of the self. *Annu. Rev. Psychol.* 58 (2007), 317–344.

[17] Nale Lehmann-Willenbrock, Hayley Hung, and Joann Keyton. 2017. New frontiers in analyzing dynamic group interactions: Bridging social and computer science. *Small group research* 48, 5 (2017), 519–531.

[18] Bruno Lepri, Nadia Mana, Alessandro Cappelletti, Fabio Pianesi, and Massimo Zancanaro. 2009. Modeling the personality of participants during group interactions. In *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 114–125.

[19] Nadia Mana, Bruno Lepri, Paul Chippendale, Alessandro Cappelletti, Fabio Pianesi, Piergiorgio Svaizer, and Massimo Zancanaro. 2007. 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*. ACM, 9–14.

[20] Gabriel Murray and Catharine Oertel. 2018. Predicting Group Performance in Task-Based Interaction. In *Proceedings of the 20th ACM on International conference on multimodal interaction*. ACM.

[21] Fabio Pianesi, Nadia Mana, Alessandro Cappelletti, Bruno Lepri, and Massimo Zancanaro. 2008. Multimodal recognition of personality traits in social interactions. In *Proceedings of the 10th international conference on Multimodal interfaces*. ACM, 53–60.

[22] Anna Polychroniou. 2014. *The SSPNet-Mobile Corpus: from the detection of non-verbal cues to the inference of social behaviour during mobile phone conversations*. Ph.D. Dissertation. University of Glasgow.

[23] Nihar Ranjan, Kaushal Mundada, Kunal Phaltane, and Saim Ahmad. 2016. A Survey on Techniques in NLP. *International Journal of Computer Applications* 134, 8 (2016), 6–9.

[24] David Reitter and Johanna D Moore. 2014. Alignment and task success in spoken dialogue. *Journal of Memory and Language* 76 (2014), 29–46.

[25] Dairazalia Sanchez-Cortes, Oya Aran, Marianne Schmid Mast, and Daniel Gatica-Perez. 2012. A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Transactions on Multimedia* 14, 3 (2012), 816–832.

[26] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. 2015. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE transactions on pattern analysis and machine intelligence* 37, 6 (2015), 1113–1133.

[27] Stefan Scherer, Nadir Weibel, Louis-Philippe Morency, and Sharon Oviatt. 2012. Multimodal prediction of expertise and leadership in learning groups. In *Proceedings of the 1st International Workshop on Multimodal Learning Analytics*. ACM, 1.

[28] C Sedikides. 86. Gregg. AP (2003). Portraits of the self. *Sage handbook of social psychology* 1 (86).

[29] Leimin Tian, Johanna D Moore, and Catherine Lai. 2017. 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*. ACM, 45–46.

[30] Alessandro Vinciarelli, Paraskevi Chatziioannou, and Anna Esposito. 2015. When the words are not everything: the use of laughter, fillers, back-channel, silence, and overlapping speech in phone calls. *Frontiers in ICT* 2 (2015), 4.

[31] Alessandro Vinciarelli, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D'Errico, and Marc Schroeder. 2012. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing* 3, 1 (2012), 69–87.