Computational Representation of Discourse Practices Across Populations in Task-based Dialogue

Elijah Mayfield, David Adamson, Alexander I. Rudnicky, and Carolyn Penstein Rosé
Carnegie Mellon University
Language Technologies Institute
5000 Forbes Avenue, Pittsburgh, PA 15213
{emayfiel, dadamson, air, cprose}@cs.cmu.edu

ABSTRACT
In this work, we employ quantitative methods to describe the discourse practices observed in a direction giving task. We place a special emphasis on comparing differences in strategies between two separate populations and between successful and unsuccessful groups. We isolate differences in these strategies through several novel representations of discourse practices. We find that information sharing, instruction giving, and social feedback strategies are distinct between subpopulations in empirically identifiable ways.

Author Keywords
task-based dialogue, discourse practices, information sharing, intercultural collaboration

ACM Classification Keywords
I.2.7 Artificial Intelligence: Natural Language Processing—*Discourse*;

INTRODUCTION
A critical task in natural language processing is understanding how dialogue is structured and information is shared between speakers. This structure is not captured well with surface-level features such as word distributions, and varies drastically based on who is speaking to whom. In the systemic functional linguistics literature, it has frequently been argued that making discourse practices associated with social interpretations explicit is an important step towards resolution of social problems related to positioning within an interaction or within a community more broadly [24].

In this paper we illustrate the usage of a machine learning methodology for identifying cultural differences in discourse practices between communities. We demonstrate through corpus based experimentation the important connection between the representation of the data and the nature of the differences that can be identified using such a methodology. We present multiple novel representations of discourse practices, all driven by structured dialogue annotation based on sociolinguistic theory. We also present analysis of the impact of these strategies on task success, and discuss implications for design of culturally aware interactive systems.

Advances in internet based collaboration technology have enabled industry and academia to create distributed, multidisciplinary teams that can address complex problems on an unprecedented scale. The Boeing Dreamliner and the Airbus A380, each a project on the order of 10 billion USD, involved tens of thousands of workers in hundreds of companies around the world [19]. Such global teams have the advantage of providing a diverse set of disciplinary and cultural perspectives on a topic, but at the same time mismatches in disciplinary and cultural conventions, work styles, power relationships and conversational norms. Such mismatches can lead to misunderstandings that negatively affect the interaction, relationships among team members, and ultimately, the quality of group work.

Building on our previous work, in this paper we explore the use of a different conversational construct that has both task and social relevance. Specifically, we use the Negotiation coding scheme, which operationalizes the authoritativeness of stance taken by participants within an interaction in relation to one another. This work has its roots in the systemic functional linguistics literature [24] and was first formalized in our prior work, where we not only define the coding scheme but show that it can be automatically applied in real time with high accuracy [26]. The conversational moves within this framework become the building blocks with which we represent differences in communication practices between cultures.

The result of our analysis in this paper is a description the relationship between speakers in terms of authority over information; an empirical model of the trajectory of a conversation as speakers take on more or less authoritative roles; and an understanding of the way that these behaviors are impacted by the influence of culture and group success.

This paper is divided into three overarching sections:

1. We give a brief overview of related work, our data, and a qualitative and quantitative overview of observations motivating our exploratory analysis.
2. We introduce three detailed methodologies for describing interactions between speakers:

ICIC'12, March 21–23, 2012, Bengaluru, India.

(a) Extracting frequent patterns of interaction over multiple turns in sequence.

(b) Building transition matrices of adjacent turns and examining differences in edge weights.

(c) Plotting authority over time, given a definition of authority based in our coding schemes.

3. We then use these frameworks to describe empirical findings on two problems:

(a) Identifying differences in discourse practices between cultures.

(b) Identifying behaviors that lead to task success.

RELATED WORK

This work is certainly not the first attempt to represent information sharing in interaction. In addition to the approaches we describe below, work in collaborative learning especially has studied the transfer of information in groups, through the use of statistical discourse analysis [10] or uptake graph analysis [35]. This prior work is more focused specifically on the process of group problem-solving in collaborative learning and is less generalizable to other domains, but has been shown to be particularly useful for the study of intercultural collaboration [36], a key goal of our work.

Studies suggest that problems do indeed arise when people from different cultural backgrounds converse face to face or via the Internet [13]. For example, an individual from a task-oriented culture such as the United States may focus exclusively on achieving an external goal, overlooking the social niceties expected by a teammate from a relationship-focused, high power distance culture such as China or Japan. Similarly, an individual from a culture that relies primarily on verbal language may miss subtleties of facial expressions or tone of voice that can modify or contribute to the meaning of the verbal language in other cultures. Gao (2000) describes differences in communication styles of Chinese students in Australia and of native English-speaking Australians (e.g., fewer politeness markers, more indirectness, and different uses of nonverbal behaviors on the part of the Chinese) that can lead to erroneous inferences about a speaker (e.g., an English-speaking Australians may perceive a Chinese speaker as rude) [15]. Recent work on large scale machine learning analysis of regional dialect differences in Twitter has revealed differences in term distributions associated with specific regions in the US [14]. However, interpretation of the implications of some of these linguistic differences in light of issues like trust and effective communication is unclear. Many of the differences between cultural groups that have been measured have failed to reliably predict the lower levels of trust and understanding that have been measured in inter-cultural groups in comparison with homogeneous groups [29, 33].

On the other hand, some prior work in the intercultural communication literature suggests that stylistic differences in communication that have tangible implications for collaboration may exist at a deeper level [13, 33]. For example, Chen (1995) compared dyadic conversations between Americans vs. Americans and Americans vs. East Asians and found that the topics of messages in the American dyads were more likely to overlap [9], suggesting that members of culturally homogeneous pairs were more likely to engage in what is referred to as transactive conversational behavior, where interlocutors orient their contributions towards the contributions of their partners [6]. Similarly, Li (1999) found more problems in information exchange when a nonnative speaker was talking to a native English speaker than vice versa [22]. She suggests that nonnative speakers may not realize they do not understand and thus fail to ask for needed clarification, which may also be indicative of low transactivity. Transactivity has been noted to be associated with trust and intimacy between conversational partners [4]. These differences reflect aspects of conversations that have both task relevance and social relevance, since rich constructs like transactivity represent the process of building consensus within groups and also reflect a level of multual respect between group members. Machine learning work measuring transactivity from text [30, 2], as well as speech [17] illustrate the importance of rich representations of text and speech in addition to powerful machine learning algorithms.

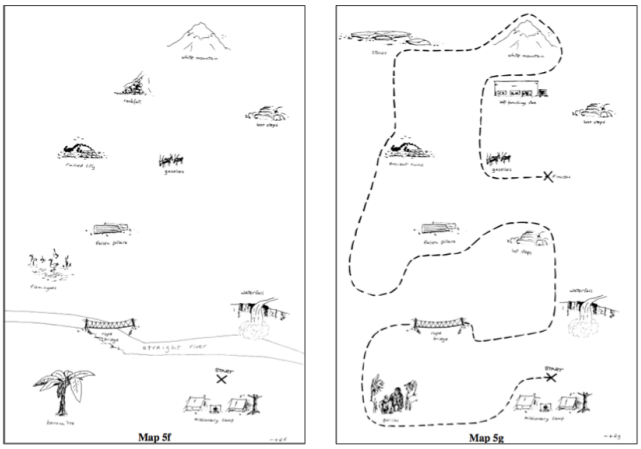

Figure 1. Example pair of maps from the MapTask corpus.

THE COLLABORATIVE TASK

In this work, we analyze the MapTask direction-giving setting. In this task, pairs of participants are each given a map. Each map has approximately twelve landmarks distributed across the map. One participant, hereafter referred to as the "Instruction Giver", has a start and finish point, with a path drawn between them, on their map. The other participant, who we refer to as the "Instruction Follower", has only the start point marked. The task is for the Giver to instruct the Follower to reproduce the path to the finish point as closely as possible, navigating the landmarks over the course of the dialogue.

A complication for the participants is that the maps are similar but not identical. Half of the landmarks on each pair of maps are identical; however, the other half of the landmarks are altered. These alterations include different names (e.g. *Ancient Ruins* instead of *Ruined City*), swapped landmarks (different landmarks occupying the same space on each map), invisible

landmarks (only marked on one participant's map), or duplicated landmarks (appearing once in the same place on each participant's map, and a second time in a different location on one of the two maps).

Our data for this study comes from two previously collected corpora. The original HCRC MapTask corpus [3] was collected from 64 participants in Scotland, totalling 128 dialogues. The DCIEM MapTask corpus [5], collected later in Canada, reproduced the same study to examine the effects of sleep deprivation on military personnel. 66 dialogues were completed under normal conditions, and 150 additional dialogues were recorded in various impaired conditions.

In this work we sample 28 dialogues from the HCRC corpus and 33 conversations from the control condition of the DCIEM corpus (where there is no sleep deprivation or drug use). This corpus is a superset of the 20 used for analysis in [26]. In total, these 61 conversations make up 14,720 lines of dialogue. From this point forward, we consistently refer to the two subpopulations as the Scottish Civilian or Canadian Military groups.

BASELINE ANALYSIS WITH WORD DISTRIBUTIONS

The state-of-the-art in modeling regional dialect variation within the language technologies community starts with a representation of text referred to as a "bag of words" model [14]. In this paradigm, a text is represented as a binary vector with n dimensions, where n is the vocabulary size, and each dimension maps to a single word. For each instance in a data set, a given dimension receives a value of 1 if its word appears at any point in the text, and value 0 if it does not appear. While this representation is simple, it has proven to be surprisingly robust as in a wide range of language modeling tasks where content is the focus. However, it poses challenges with respect to generalizability of trained models where style rather than content is the focus.

Here we illustrate a different issue, namely that the regional dialect differences that are discovered may not be task relevant. To illustrate this, we begin by building a linear support vector machine (SVM) classifier that takes this "bag of words" representation of dialogues, and predicts the subpopulation, namely Scottish versus Canadian. Such a model is 100% accurate in cross validation - given an entire dialogue worth of text, it can predict with virtual certainty which corpus a dialogue has come from.

By standards of success typically adopted within the language technologies community for classification tasks, this would be considered a success. However, by observing the weights assigned to each feature, we are left with quite a different impression. The most predictive pair of words is "round" (weighted heavily towards Scottish civilians) and "around" (weighted heavily towards Canadian military). In context, these words are used for direction giving - speakers of Scottish dialects will instruct the follower to *"go round... [the landmark]*, while speakers of Canadian dialects will say *"go around..."*. Similar examples exist for "little" and "wee", "anyway" and "anyways", region-specific measurements such as "inches" and "centimetres", or region-specific terms like "aye". While these differences are arguably indexical of culture, and lexical differences such as these can be important playing cards in interactions where cultural dialect differences are associated with differences in status, we will see that the differences we are able to detect using richer representations of the language operate at a level that is more task relevant, whereas these word level differences are not.

Some distinctions across word distributions would arguably be task relevant. For example, we find when we examine the corpora a difference in use of cardinal directions, with 71.8% of words such as "north", "east", "southeast", etc. occurring in Canadian dialogues, and relative terms occurring more frequently in Scottish dialogues ("over" and "under", for instance, occur 63.4% of the time in Scottish conversations). While these differences are potentially interesting, they are given very little weight by any machine learning model as they are not as predictive as dialect specific but task-irrelevant keywords.

Instead of modelling conversations based on word choices, then, we wish to study how a discourse is structured, and the high-level ways in which groups go about completing a task.

ANNOTATION WITH THE NEGOTIATION FRAMEWORK

The relationship between sharing information and task success has been repeatedly examined in a variety of contexts. Issues such as shared visibility and common grounding of objects [18], recognizing understanding or comprehension in the partner you are speaking to [11], and the use of physical action to communicate intent or understanding [16] have all been shown to have a major impact on the efficiency of communication. The issue of how communication differs between groups is particularly important to any research on interactive systems where cultural styles of interaction are relevant. Prior work has shown that subpopulations react to and interact with systems very differently [1]. Other work has also shown the utility of dialogue systems in military applications, such as soldier training [28]. Designing systems for specific groups of users should take into account the specific interaction patterns common among those users.

To represent information about communication and information sharing at a level near that described above, we base our analysis on the Negotiation framework, developed in the systemic functional linguistics (SFL) community [24]. Annotation with the Negotiation framework gives us the building blocks with which to study interaction. Annotation is only an initial step, however. Counts or even distributions of these annotations are not informative enough from our perspective. What lends the most insight into styles of interaction and footing between participants are sequences of annotations.

The Negotiation framework attempts to describe how speakers use their role as a source of knowledge or action to position themselves relative to others in a discourse [24]. The Negotiation framework is primarily made up of four main codes, K1, K2, A1, and A2. The four main codes are divided on two axes, illustrated in Figure 2. First, is the utterance related to exchanging information, or to exchanging services and actions? If the former, then it is a K move (knowledge); if the

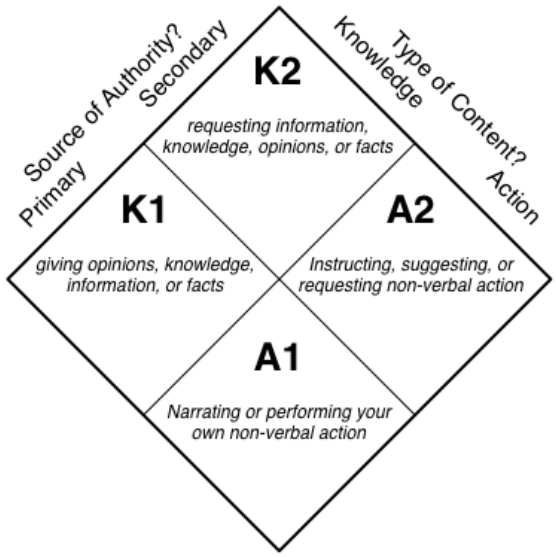

Figure 2. The main codes of the Negotiation framework.

latter, then an A move (action). Second, is the speaker acting as a primary or secondary source of action or knowledge? In the case of knowledge, this often corresponds to the difference between assertions (K1) and queries (K2). For instance, a statement of fact or opinion is a K1:

g	*K1*	*well i've got a great viewpoint here just below the east lake*

By contrast, asking for someone else's knowledge or opinion is coded as a K2:

g	*K2*	*what have you got underneath the east lake*
f	*K1*	*a tourist attraction*

In the case of action, the codes usually correspond to narrating action (A1) and giving instructions (A2), as below:

g	*A2*	*go almost to the edge of the lake*
f	*A1*	*yeah okay*

Four additional categories are used for all other moves. These categories were selected from a larger set available in the systemic functional linguistics literature, and are the most common codes to appear repeatedly in different authors' analyses.

- Followup (f) moves are marked when a K1 or A1 move is being directly acknowledged as understood, without contributing additional new content to the discourse (such as backchanneling).

- Challenge (ch) moves are marked when a move directly undermines some assumption of the previous line, as in the example below:

g	*A2*	*come directly down below the stone circle and we come up*
f	*ch*	*I don't have a stone circle*
g	*o*	*you don't have a stone circle*

- Tracking moves (tr) indicate a restatement or request for restatement, and are simply marking a failure to hear or understand what was said because of poor emphasis or pronunciation.

- Finally, all other moves are classified as o. This includes floor-grabbing moves, false starts, preparatory moves, and any other non-contentful contributions.

Agreement between annotators for this scheme is high both for distinguishing our four core codes and a collapsed "other" category (κ = .74) and for labelling all codes (κ = .66).

A final layer of complexity in the Negotiation annotation scheme is that labels are not assigned independently. Instead, there is a notion of a sequence - a connected series of moves, similar to the concept of adjacency pairs in conversation analysis. This notion states that a series of turns must follow some consistent structure, based around a primary move. A sequence is defined as a primary move (K1 or A1) and the context around that move, be it a secondary move (K2 or A2) that requested the move, o moves representing false starts or floor grabbers, and responses to the primary move in the form of followup or challenge moves. This structure was defined as a set of formal constraints in the first publication of this framework [26].

Analysis Details

Throughout our analysis, we use a consistent set of labels. We mark, for each turn, which of the eight possible codes from the Negotiation framework the line was annotated with, crossed with the speaker of the label. Therefore, a turn from the instruction giver which was labelled as a K2 receives the label "gK2". This gives sixteen possible labels for our analysis, and allows us to study the interacting effects of both speaker roles and turn-by-turn behavior. In this work, these labels were applied by hand; however, they have been shown to be reproducible with high reliability [26], meaning that these analyses could be incorporated into real-time systems.

In the next section we describe multiple representations of conversation which circumvent these dialect-specific keywords by relying solely on coded labels from the Negotiation framework. We show that coding schemes can be represented in meaningful ways to describe sequential events, without overfitting to dialect or topic, and instead highlighting the ways in which information sharing and instruction giving differ across subpopulations. This analysis allows us access to the next level of analysis, determining which practices have an impact on what results are achieved, and which are culturally specific but do not impact a group's ability to collaborate.

REPRESENTATING INTERACTION IN DIALOGUE

Numerous disparate communities have informative and distinct methodologies for studying sequential conversational data. We believe that these approaches are complementary, and as such we draw inspiration from three distinct fields for our data analysis: transition graph analysis; stretchy interaction patterns; and trajectories of authority.

Transition Matrix Analysis

Exploratory sequential data analysis (ESDA) has been used as a phrase to describe a large number of different approaches to data mining [32]. Here, we use the term to describe a specific style of analysis where sequences of moves are analyzed for frequently co-occurring activities. This analysis usually involves deriving a transition probability matrix based on observed moves and interpreting the resulting transition graphs. This has been successfully applied to infer subtasks based on closely co-occurring moves [37], the impact of gender or argument style on group interactions in message boards [21, 20] and studying collaborative problem solving in student groups [34, 8].

The primary unit of analysis for ESDA is an interaction graph. These graphs represent transitions between adjacent labels in some sequential data, and are in theory a complete graph. First, a transition matrix T is built, with cell $t_{x,y}$ counting the number of times label x occurred immediately before label y. Then, a complete graph can be built giving the probability that for any utterance u, label y will occur after label x for any possible combinations of x and y.

These graphs are difficult to interpret in their complete form, as the number of edges grows polynomially. However, comparisons between two subsets of a data set can be made by building two graphs, one from each subset, and comparing the differences in probabilities between graphs. In our visualizations, only the transitions that we are interested in describing qualitatively will be displayed, to make graphs readable; other connections were not significant or are not relevant to our conclusions.

Stretchy Interaction Patterns

Prior work has introduced the notion of a "stretchy pattern" [25]. These patterns display the sequence of categories of behavior that occur, based on a coding scheme. A limitation of transition matrices is that they limit the observations of interactions between annotations to adjacent turns. Stretchy patterns overcome this by allowing longer interactions to be captured, if they are frequent and informative enough. The resulting patterns cannot be comprehensively listed in a matrix; hwoever, the advantage they give in expressive power allows them to be analyzed in more detail individually.

A pattern is comprised of a series of tokens which can be drawn from a small number of classes. These tokens also encode the speaker of an utterance. A token may also be a gap, which is allowed to consume up to some number of concrete tokens; or a shift in sequence, either to the next sequence (marked by $\rightarrow$) or shifting back to a previous sequence that was unfinished (marked by $\leftarrow$). In our case, we set the range of allowed pattern sizes to be $3-6$ tokens, with gaps (marked by $\square$) allowed to consume from $1-3$ tokens. Location of gaps is fixed at particular points and thus mirrors and extends the concept of lags in sequential data analysis [31].

Authority Trajectories

We may also consider a conversation based on two constantly adjusting metrics: the flow of information between speakers (Information authority), and the flow of directions for action between speakers (Action authority). At each utterance, we define these measures based on the Negotiation codes that have occurred up to that point. In our representation, a K1 move from the instruction giver represents a shift of +1, while a K2 move from the instruction giver represents a shift of -1; similarly, a K1 move from the instruction follower is a shift of -1 and a K2 move from the follower is a shift of +1. The same formula can be used for Action authority, but the polarity is reversed; an A2 move represents an authoritative move, thus an instruction giver A2 move is marked as +1, and so on.

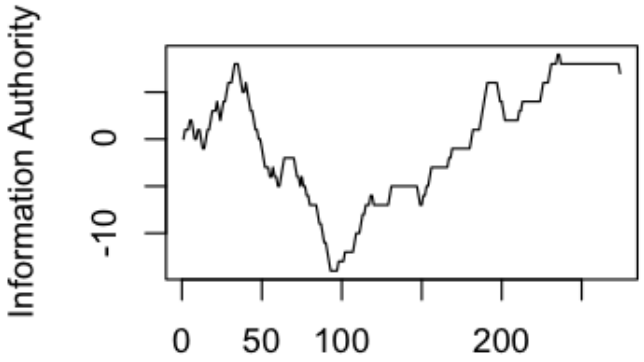

Figure 3. Example K-authority trajectory over a single conversation.

An example of the resulting trajectory is given in Figure 3, showing only information authority. In that dialogue, authority shifts towards the instruction giver early, before drastically shifting towards the instruction follower as a series of questions are asked by the giver about the follower's map, followed by a gradual shift back to the instruction giver for the rest of the dialogue.

Trajectories as described above give a quick visual depiction of the flow of information and instructions between speakers over the course of a conversation. To understand similarities between these trajectories across multiple conversations, we want to empirically group similar conversations together. To do this, we use time series clustering based on dynamic time warping [7], a standard method for measuring similarity between time series data with different lengths. From this, we can build a similarity matrix between each conversation. Conversations are then grouped together through hierarchical agglomerative clustering [23], a standard clustering algorithm. The resulting output is a progressively more refined taxonomy of conversations, which can be simplified to arbitrary levels of granularity.

An advantage of the trajectory formulation is that it can be used for any span of utterances, not just the spans from first to last utterance of a conversation. In our experiments, we consider both this whole-conversation case, and a second case which parallels the segmentations from the other representations of our data - only utterances within sequences for a given landmark are analyzed to produce the trajectory for the mentions of a given landmark can be considered.

The final stage of this trajectory analysis is to cluster trajectories. Using hierarchical agglomerative clustering, we group

similar trajectories. We cluster whole conversations twice, once by Action authority and once by Information authority. This results in a dendrogram of relatedness which can be grouped at arbitrarily refined subsets. We produce clusters such that each cluster contains at least four conversations. This results (coincidentally) in eight clusters for both Action authority and Information authority.

TASK: IDENTIFYING A DYAD'S SUBPOPULATION

We first consider the problem of identifying, based on a transcribed and annotated conversation, whether that interaction comes from the Canadian Military subpopulation, or the Scottish Civilian subpopulation. Later, we will come back to these same patterns and identify which are relevant to success at completing a task, and which may cause misunderstandings between culture but do not have an impact on performance.

We divide our data based on landmarks, from the first time a landmark is mentioned to the last. The notion of sequences, introduced when defining our framework, is key to this analysis - if any utterance in a sequence contains a reference to a given landmark, the whole sequence of turns is included in the interaction on that landmark. Because landmarks often overlap, the same sequence of turns may occur multiple times in our data set, once in the context of each landmark that sequence references.

Transition Matrix Analysis

The single most common interaction in the MapTask domain is instruction giving. Therefore, the way in which this interaction is structured is worth especially detailed study.

Figure 4 gives a detailed breakdown of the instruction-giving process from our ESDA analysis, and the variations in how instruction is given and received between subpopulations. We see that the standard interaction is as expected - the instruction giver begins with an A2 move, followed by an A1 move from the instruction follower, and a followup move from the giver to show that the narration was noticed.

A pattern for A2 instructions and responses is clear in both subpopulations. Differences exist between subpopulations. Most notably, the feedback from the instruction giver is more common in the Canadian military population; over half of all actions from the instruction follower are met with an acknowledgement, compared to roughly one third among Scottish civilians. We also see more non-contentful o moves from the instruction follower at the end of an instruction in the Scottish civilian population. The Canadian military population, by contrast, is much more likely to shift immediately to the next instruction.

An explanation of this that emerges from examining the graph in Figure 4 is based on the formulaic nature that seems to emerge most strongly in Canadian military dialogues. The **gA2-fA1-gf** pattern is very strongly emphasized, and deviations from that cycle are unusual. In the Scottish civilian cycle of instructions, however, there is more noise. Fewer moves from the instruction follower are explicitly acknowledged, more time is filled with non-contentful moves, and a

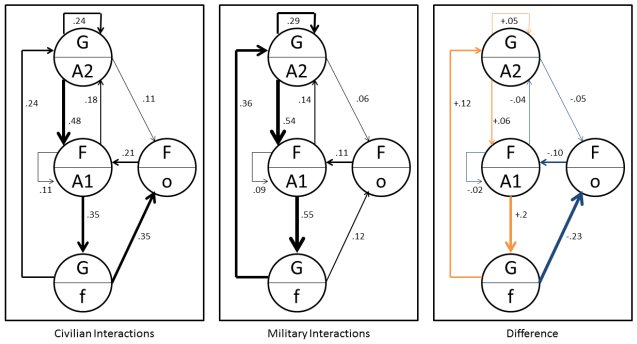

Figure 4. Variation in standard instruction sequences between subpopulations. In our diagrams, line thickness corresponds to probability, and color in the final graph denotes direction of difference - blue lines represent a transition more common in the left-hand graph, orange lines are more common in the right-hand graph.

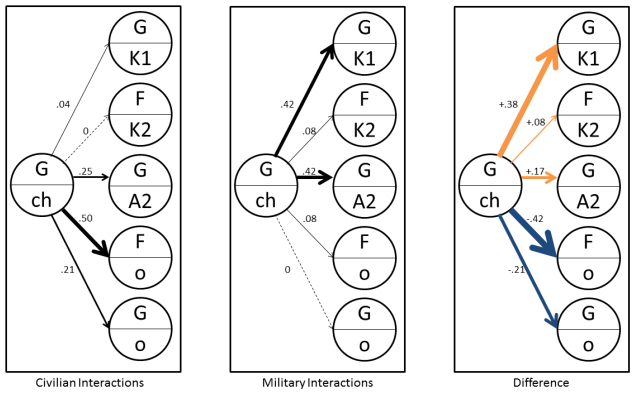

Figure 5. Differences in responses elicited from a challenge move from the giver between subpopulations.

return to the gA2 label for the next instruction is much weaker (indicating divergences to other parts of the graph not shown).

Challenge moves and the breakdown of common ground

Challenge moves are relatively rare in our data set (1.1% of utterances in total), but they are in fact critical points in a discourse. From qualitative analysis, we know that challenges are made when assumptions about shared information clash. An A2 move about a landmark you cannot see may prompt a challenge, for instance, as the instruction is undermined.

Responses to challenge moves are distinctly different between conditions in our data set. Figure 5 shows the difference in subpopulations when an instruction giver makes a challenge. The Scottish civilian response is usually contentless - a backchannel or acknowledgement before any response is made. In the Canadian military subpopulation, however, the follower very rarely responds directly to a challenge move; instead, the instruction giver follows immediately with a contentful move (A2 or K1). This suggests that, as we have observed a much more formulaic structure to the interactions in the Canadian Military subpopulation, a challenge move may prompt the instruction follower to stay silent until the dialogue resumes its formulaic structure.

Stretchy Patterns Analysis

Examples from within the top 25 stretchy patterns for each population are shown in Table 1. While the ESDA analysis highlighted the tightly-knit structure of instruction based interactions, the stretchy pattern analysis finds a stronger signal in the information based sequences, where interruptions, false starts, and other more minor moves are more common and transition matrices (which observe only immediate transitions) are less likely to find a signal.

In the Canadian military population, many of the dominant patterns highlight part or all of an information exchange. **gf** moves are shown to frequently follow immediately after or within a few tokens of **fK1**. Among the most distinctively Scottish-civilian patterns, most deal with action exchanges. The top Scottish pattern **fA1** $\rightarrow \square \rightarrow \square \rightarrow$ also illustrates that a succession of short turns, starting with an **fA1** action-completed move, is indicative of this population. We also see a number of strong Scottish patterns that begin with moves labelled **go**. In particular, we see patterns along the lines of **go gA2** - this represents the instruction-giver grabbing the floor before requesting an action.

Pattern	Predicts	Kappa
fK1 $\square$ gf $\rightarrow$	Canadian	0.424
gK2 $\square$ fK2 $\square$ gf $\rightarrow$	Canadian	0.415
fK1 gf $\square \rightarrow \square$ gA2	Canadian	0.310
fA1 $\rightarrow \square \rightarrow \square \rightarrow$	Scottish	0.179
go $\square$ gA2 $\square \rightarrow$	Scottish	0.143
go $\square$ fK2	Scottish	0.134

Table 1. Highlighted patterns predictive of the source subpopulation of an interaction.

Authority Trajectory Analysis

Trajectories of Action authority highlight differences in the ordering of information sharing. Figure 6 shows information authority clusters, for a visual understanding of the difference in trajectories; each graph represents the average trajectory of the dialogues in that cluster, normalized for length. We also show the distribution of subpopulations in each cluster. In some conversations, exclusively Canadian, the Action authority trajectory is flat for the first third or half of the conversation. These groups were clustered together in our unsupervised algorithm. This pattern highlights a strategy of delayed instruction giving; initial communication is almost entirely building a shared knowledge base, going over the entire map, and clarifying differences.

We see that this two-phase process is not always well-explained even within a pair. For instance, the instruction giver sometimes is forced to clarify that their explanations are not instructions, but attempts to build common ground:

g	*K1*	*bottom line of the green bay would be between the word haystack and the actual haystack*
g	*K1*	*and it starts about uh an inch and a half to the right of the haystack*
f	*K2*	*so i go over the top of the haystack or underneath*
g	*A2*	*you're not drawing any line yet*

We find that Action authority almost uniformly climbs towards the instruction giver, as expected. The remaining difference in Action authority clusters seems highly dependent on the slope; that is, the total number of instructions given. These slope differences do not discriminate between subpopulations.

Differences in Information authority are more starkly different, and show patterns of information sharing that are specific to subpopulations. The most starkly different is cluster 3. Conversations grouped in this cluster gave almost no information about the instruction giver's map, and instead focused entirely on building common ground from the instruction follower's perspective, describing locations on their map. This behavior is exclusively existent in the Canadian military subpopulation.

Other clusters showed clearer splits between authority of the two speakers. Clusters 4 and 7 are related in that for the first half of the conversation, information is given mostly by a single speaker, and incremental information is then fixed by the opposite speaker in the second half of the dialogue; however, the roles are reversed between clusters, and these clusters correspond to subpopulations. Groups giving the instruction follower precedence are exclusively Canadian military, while groups beginning from the instruction giver's perspective are more likely to be Scottish civilian.

The remaining clusters show a more balanced approach to information giving, with both instruction giver and instruction follower trading roles of information authority as needed. In some cases, information sharing stops almost entirely at a certain point, shifting entirely to instruction giving (potentially when a group feels a sense of "hitting a stride").

Synthesis of findings

Interactions in the Canadian Military subpopulation are distinctly more orderly and predictable, compared to the Scottish civilian subpopulation. This emerges in all phases of interaction, from question-answer pairs, to instruction giving, to the response to challenges and problematic points within an interaction. Also, two distinct strategies for building common ground emerge - one where participants spend a large amount of time at the beginning of an interaction to collaboratively build a shared representation of the map, and a second where information is shared on an as-needed basis, immediately jumping to instruction giving. The first strategy appears only in Canadian military interactions, while the second occurs in both populations.

TASK: PREDICTING GROUP SUCCESS

Differences in discourse practices between subpopulations can be identified with our analysis techniques. We now shift to a related question: are these practices related to the end success of a group at completing a task?

Work on task success in the MapTask corpus has usually been based on absolute error between drawn paths and source paths

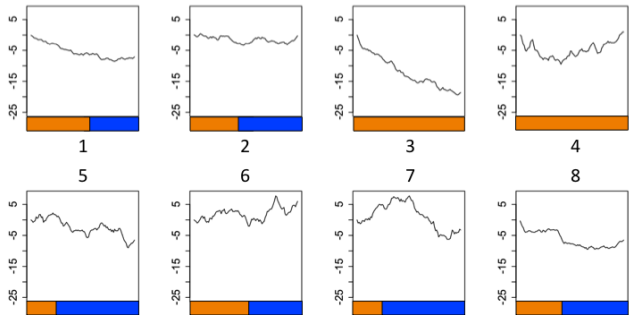

Figure 6. Averaged trajectories of Information Authority across clusters, normalized for time. For each cluster, the distribution is shown for Canadian military (orange) and Scottish civilian (blue) conversations.

[25, 27]. This error, measured in cm^2, is useful as an aggregate comparison between groups. However, we wish to understand successful interactions at a much more fine grained level, specifically at the level of an individual landmark. In order to achieve this, we first divide our conversations based on the spans of utterance that a landmark is referenced in. These segments will serve as the unit of analysis for the following section.

Aggregate analysis

Our first analysis is to test the predictive values of participant metadata on success. In our analysis, we use Visibility as a moderating variable. Since some landmarks are more challenging than others, we also use Landmark as a moderating variable.

We associate each landmark segment with a success metric, which we refer to as the Grade, using the Incorrect Entity Score metric first used in [12]. This measure marks, for each landmark, whether the drawn path reproduces the source path for that particular landmark. In each conversation, each landmark is marked as perfect, "good" miss (for a path that is too close or too far from a landmark, but passes it on the correct side or corner), "bad" miss (for a path that passes a landmark on the wrong side or goes through the landmark), or no attempt (for a path that never comes close to a landmark). The HCRC MapTask corpus (Scottish Civilian) was already annotated with these landmark references; we reproduce that annotation for the DCIEM MapTask corpus.

The first question we must address is whether there is any main effect of cultural subpopulation on task success. We do this using a χ^2 analysis where Subpopulation is the Independent variable, Grade is the Dependent variable, and Landmark and Visibility are moderating variables. When we examine the Likelihood ratio tests we find that Landmark ($p < .0001$) and Visibility ($p < .05$) are significant predictors of Grade. Subpopulation is not significant, nor is there an interaction between Subpopulation and Visibility.

Transition Matrix Analysis

An ESDA analysis of task success highlights the importance of the **ch** move. Differences in responses to challenges exist for both giver and follower challenges, though the response is distinctly different. Giver challenges in successful interactions are usually responded to with a backchannel from the follower; however, in unsuccessful interactions, those challenges are immediately followed up with an A2 move from the instruction giver.

For follower challenges, a similar pattern emerges; in successful interactions, the next turn is usually an acknowledgement from the instruction giver, while in unsuccessful interactions, the follower is much more likely to continue the conversation, whether with a K1 or K2 move. An acknowledging o move from the other speaker, in both cases, is more likely in successful interactions. On the other hand, when no vocal acknowledgement is made in response to the challenged assumption, future moves may not be as likely to shift to compensate.

Authority Trajectory Analysis

Earlier in this work we clustered conversations based on their authority trajectories. We did this for both action-based authority trajectories (A-clustering) and knowledge authority trajectories (K-clusters). We now test to see whether these same clusters are predictive of task success.

Using a χ^2 test, we confirmed that these clusters represent significant distributions of the Canadian and Scottish subpopulations, however most are not solely dominated by one subpopulation or another. Task relevance to the behavior patterns represented by the clusters then are consistent with the nature of cultural differences between populations. These clusters do not predict cultural subpopulation, but some are more strongly associated with particular subpopulations.

To assess the task relevance of these differences, we again conduct a χ^2 analysis with Grade as the Dependent variable and Landmark and Visibility as moderating variables. This time the Independent variable is alternately the K-clustering or the A-clustering.

In the A-clustering analysis, we find a marginal main effect of A-clustering ($p = .1$), and a significant interaction between A-clustering and Visibility such that Visibility only has a significant effect for certain patterns of behavior. The effect of K-clustering is only a trend ($p = .16$) and there is no significant interaction with Visibility.

The result trends in favor of the marked pattern of delayed instruction giving in the Canadian subpopulation, until common ground is achieved. It is notable that the evidence of task relevance is relatively weak; in particular, weaker than one would expect given how strongly individuals may cling to cultural practices. Despite the extensive differences in trajectories of authority and the rate and distribution of moves for reaching common ground, there was no main effect on task success related to these differences. Patterns extracted from just the successful interactions (and likewise from just the unsuccessful instances) for each cultural group were nearly identical to those extracted from the whole dataset. We do not consider these results to indicate a failure on the part of the analysis technique. Rather, they suggest that cultural differences are not the driver of task success.

Synthesis of findings

Our findings highlight the difficulty of attributing success to any one indicator of group composition. We highlight the importance of challenges indicates a breakdown of communication, where shifts in shared common ground may depend on appropriate acknowledgement of disagreements. However, our main finding is that those differences which separate out cultures are only weak predictors of group success. This affirms the findings of prior work, which has focused on improving communication between diverse groups, rather than coercing participants into culture-specific behaviors stereotyped as more or less effective.

CONCLUSIONS

This work presents a thorough discussion of the issue of information sharing and instruction giving in dialogue. We began by showing that a simple surface representation of dialogue is insufficient for answering questions about information sharing and building of common ground. Three more complex representations for sequential data analysis were highlighted: transition matrices, interaction patterns, and authority trajectories.

A general factor which repeatedly arises in our results is the relative orderliness of the Canadian military interactions. The structure of instruction giving, followed by narration, and then acknowledgement (**gA2**→**fA1**→**gf**) is much more consistent in that subpopulation. Similarly, challenge moves are responded to with contentful moves G far more often than acknowledgements, which are more common in the Scottish civilian subpopulation. We also observe a difference in information sharing strategies over the course of an entire conversation. Scottish civilians gave information almost entirely on an as-needed basis, not planning past the next landmark. Canadian military pairs split between an this as-needed strategy and a strategy of building common ground to start a conversation, before giving any instructions.

These findings show that our methodology for describing interactions is effective both at the turn-by-turn level, describing frequent sequences of interaction multiple turns long, and over the course of a dialogue, describing the shift in roles over time. This means that similar techniques can be used both in real-time systems, which rely on a short window of previous turns, and in post hoc analyses of interactions, which can take advantage of full transcripts.

Most importantly, though, our work suggests that the most significant barrier to effective intercultural collaboration is not the adoption of strategies specific to one culture or another. The relatively stronger impact of subpopulation on behavior, rather than task success, suggests that the development of methodology and technology that improves on mutual understanding of cultural practices is needed, and is likely to make the largest contribution to intercultural collaboration moving forward. Incorporating these elements in a real-time system is a non-trivial task, but is promising given that prior work has already shown that automatic annotation of these labels is highly accurate [26].

This automated process will motivate continued research on this topic. One issue that we have not yet studied is the problem of threading. In a single dialogue, multiple issues may be relevant at any given moment. Speakers may refer to both on-task and off-task information, and even as information relates to a task, speakers may be attempting to resolve multiple issues concurrently. This problem is only exacerbated as we extend our analysis to domains with more than two speakers. Resolving these issues of threading and proper attribution of authority to topics is a major thrust of our continuing work.

Acknowledgements

The research reported here was supported by National Science Foundation grant IIS-0968485, and in part by the Pittsburgh Science of Learning Center, which is funded by the National Science Foundation grant SBE-0836012.

REFERENCES

1. H. Ai, A. Raux, D. Bohus, M. Eskenazi, and D. Litman. Comparing spoken dialog corpora collected with recruited subjects versus real users. In *Proceedings of the ACM SIG on Discourse and Dialogue*, 2007.
2. H. Ai, M. Sionti, Y.-C. Wang, and C. P. Rosé. Finding transactive contributions in whole group classroom discussions. In *Proceedings of the International Conference of the Learning Sciences*, 2010.
3. A. Anderson, M. Bader, E. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. Thompson, and R. Weinert. The hcrc map task corpus. In *Language and Speech*, 1991.
4. M. Azimitia and R. Montgomery. Friendship, transactive dialogues, and the development of scientific reasoning. In *Social Development*, 1993.
5. E. Bard, C. Sotillo, A. Anderson, and M. M. Taylor. The dciem map task corpus: Spontaneous dialogue under sleep deprivation and drug treatment. In *ESCA-NATO Tutorial and Workshop on Speech under Stress*, pages 25–28, 1995.
6. M. Berkowitz and J. Gibbs. Measuring the developmental features of moral discussion. In *Merrill-Palmer Quarterly*, 1983.
7. D. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *AAAI Workshop on Knowledge Discovery in Databases*, 1994.
8. M. Cakir, F. Xhafa, N. Zou, and G. Stahl. Thread-based analysis of patterns of collaborative interaction in chat. In *Proceedings of AI in Education*, 2005.
9. L. Chen. Interaction involvement and patterns of topical talk: A comparison of intercultural and intracultural dyads. In *International journal of Intercultural Relations*, 1995.
10. M. M. Chiu. Group problem-solving processes: Social interactions and individual actions. In *Theory of Social Behavior*, 2000.

11. H. Clark and M. Krych. Speaking while monitoring addressees for understanding. In *Memory and Language*, 2004.

12. B. Davies. Principles we talk by: Testing dialogue principles in task-oriented dialogues. In *Pragmatics*, 2010.

13. E. Diamant, S. Fussell, and F.-L. Lo. Collaborating across cultural and technological boundaries: Team culture and information use in a map navigation task. In *International Workshop on Intercultural Collaboration*, 2009.

14. J. Eisenstein, N. A. Smith, and E. P. Xing. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2011.

15. M. Gao. Influence of native culture and language on intercultural communication: The case of prc student immigrants in australia. 2000.

16. D. Gergle. *The Value of Shared Visual Information for Task-Oriented Collaboration*. PhD thesis, 2006.

17. G. Gweon, P. Agarwal, M. Udani, B. Raj, and C. P. Rosé. The automatic assessmnet of knowledge integration processes in project teams. In *Proceedings of Computer Supported Collaborative Learning*, 2011.

18. J. E. Hanna, M. K. Tanenhaus, and J. C. Trueswell. The effects of common ground and perspective on domains of referential interpretation. In *Memory and Language*, 2003.

19. A. Hellemans. Manufacturing mayday. In *IEEE Spectrum*, 2007.

20. A. Jeong. The sequential analysis of group interaction and critical thinking in online threaded discussions. In *American Journal of Distance Education*, 2003.

21. A. Jeong and G. Davidson-Shivers. The effects of gender interaction patterns on student participation in computer-supported collaborative argumentation. In *Educational Technology, Research, and Development*, 2006.

22. H. Li. Grounding and information communication in intercultural and intracultural dyadic discourse. In *Discourse Processes*, 1999.

23. C. Manning, P. Raghavan, and H. Schutze. *Introduction to Information Retrieval*. 2008.

24. J. Martin and D. Rose. *Working with Discourse: Meaning Beyond the Clause*. 2003.

25. E. Mayfield, M. Garbus, D. Adamson, and C. P. Rosé. Data-driven interaction patterns: Authority and information sharing in dialogue. In *Proceedings of AAAI Fall Symposium on Building Common Ground with Intelligent Agents*, 2011.

26. E. Mayfield and C. P. Rosé. Recognizing authority in dialogue with an integer linear programming constrained model. In *Proceedings of Association for Computational Linguistics*, 2011.

27. D. Reitter and J. Moore. Predicting success in dialogue. In *Proceedings of ACL*, 2007.

28. A. Roque, A. Leuski, V. Rangajaran, S. Robinson, A. Vaswani, S. Narayanan, and D. Traum. Radiobot-cff: A spoken dialogue system for military training. In *Proceedings of Interspeech*, 2006.

29. C. P. Rosé and S. Fussell. Towards measuring group affect in computer-mediated communication. In *ACM SIG-CHI Workshop on Measuring Affect in HCI: Going Beyond the Individual*, 2008.

30. C. P. Rosé, Y.-C. Wang, Y. Cui, J. Arguello, K. Stegmann, A. Weinberger, and F. Fischer. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. In *International Journal of Computer Supported Collaborative Learning*, 2008.

31. G. Sackett. The lag sequential analysis of contingency and cyclicity in behavioral interaction research. In *Handbook of Infant Development*. 1979.

32. P. M. Sanderson and C. Fisher. Exploratory sequential data analysis: foundations. In *Human-Computer Interaction*, 1994.

33. L. Setlock, S. Fussell, and C. Neuwirth. Taking it out of context: Collaborating within and across cultures in face-to-face settings and via instant messaging. In *Proceedings of the Conference on Computer-Supported Cooperative Work*, 2004.

34. G. Stahl. *Group Cognition: Computer Support for Collaborative Knowledge Building*. 2005.

35. D. Suthers, N. Dwyer, R. Medina, and R. Vatrapu. A framework for eclectic analysis of collaborative interaction. In *International Conference on Computer-Supported Collaborative Learning*, 2007.

36. R. Vatrapu and D. D. Suthers. Cultural influences in collaborative information sharing and organization. In *International Conference on Intercultural Collaboration*, 2010.

37. O. Vortac and M. B. Edwards. Sequences of actions for individual and teams of air traffic controllers. In *Human-Computer Interaction*, 1994.