Natural Language Engineering: page 1 of 45. © Cambridge University Press 2012
doi:10.1017/S1351324911000386

Modeling Sociocultural phenomena in discourse

GEORGE AARON BROADWELL[1],
JENNIFER STROMER-GALLEY[1],
TOMEK STRZALKOWSKI[1], SAMIRA SHAIKH[1],
SARAH TAYLOR[2], TING LIU[1], UMIT BOZ[1],
ALANA ELIA[1], LAURA JIAO[1] and NICK WEBB[1]

[1]*ILS Institute, University at Albany, SUNY, NY, USA*
e-mails: g.broadwell@albany.edu, jstromer@albany.edu, tomek@albany.edu, samirashaikh@gmail.com, ting.tim.liu@gmail.com, umitboz@yahoo.com, aelia@nycap.rr.com, jiao_gh@msn.com, webbn@union.edu
[2] *Lockheed Martin Corporation, Bethesda, MD, USA*
e-mail: Taylmail59@gmail.com

(*Received 16 September 2010; revised 15 September 2011; accepted 14 November 2011*)

Abstract

In this paper, we describe a novel approach to computational modeling and understanding of social and cultural phenomena in multi-party dialogues. We developed a two-tier approach in which we first detect and classify certain sociolinguistic behaviors, including topic control, disagreement, and involvement, that serve as first-order models from which presence the higher level social roles, such as leadership, may be inferred.

1 Introduction

We investigate the language dynamics in small group interactions across various settings. Our focus in this paper is on English online chat conversations; however, the models we are developing are intended to be universal and applicable to other conversational situations: informal face-to-face interactions, formal meetings, moderated discussions, as well as interactions conducted in languages other than English, e.g., Urdu and Mandarin.

Multi-party online conversations are particularly interesting to examine not only because they are a relatively common means of communication through the Internet, but also because the reduced cue environment implies that the only ways for group dynamics to unfold is through discourse. Although the use of chat rooms is relatively light when compared with other online communication mechanisms, particularly e-mail, there is an increase in the use of social media websites, including blogs and social networks such as Facebook. These websites present opportunities for commenting on others' posts and responding to those comments, making small group communication a common feature of online interactions. The channel characteristics of online group communication minimize the social cues that commonly help structure interaction (e.g., use of eye gaze to manage turn-taking), and that typically structure influence in the group (based on proxemics, height, vocal tone, and the

like). As such, studying online chat relies on the more explicit linguistic devices necessary to convey social and cultural nuances than is typical in face-to-face or even telephonic conversations.

Our objective is to develop computational models of how certain social phenomena, such as leadership, conflict, and group cohesion, are signaled and reflected in language through the choice of lexical, syntactic, semantic, and conversational forms by discourse participants. In this paper we report the results of an initial phase of our work during which we constructed a prototype system called Detecting Social Actions and Roles in Multi-party Dialogue-1 (DSARMD-1). Given a representative segment of multi-party task-oriented dialogue, DSARMD-1 automatically classifies all discourse participants by the degree to which they engage in selected sociolinguistic behaviors (SLB), such as topic control, task control, involvement, and disagreement. These are the mid-level social phenomena that are deployed by discourse participants in order to achieve or assert higher level social roles, including leadership. In this work we adopted a two-tier empirical approach where sociolinguistic behaviors are modeled through *observable* linguistic features that can be automatically extracted from dialogue. The high-level social roles are then inferred from a combination of sociolinguistic behaviors attributed to each discourse participant; for example, a high degree of influence and a high degree of involvement by the same person may indicate a leadership role. In this paper we limit our discussion to the first tier only: How to effectively model and classify selected sociolinguistic behaviors in multi-party dialogue.

2 Related research

Issues related to linguistic manifestation of social phenomena have not been systematically researched before in computational linguistics; indeed, most of the effort thus far has been directed toward the communicative dimension of discourse. While the Speech Act theory (Austin 1962; Searle 1969) provides a generalized framework for multiple levels of discourse analysis (locution, illocution, and perlocution), most current approaches to dialogue focus on information content and structural components in dialogue (Carberry and Lambert 1999; Stolcke *et al.* 2000; Blaylock 2002); few take into account the effects that speech acts may have upon the social roles of discourse participants. Somewhat more relevant to social roles is the research that models sequences of dialogue acts (Bunt, 1994) in order to predict the next dialogue act (Samuel, Carberry and Vijay-Shanker 1998; Stolcke *et al.* 2000; Ji and Bilmes 2006, inter alia) or to map them onto subsequences or 'dialogue games' (Carlson 1983; Levin *et al.* 1998) from which participants' functional roles in conversation (though not social roles) may be extrapolated (e.g., Linell 1990; Poesio and Mikheev 1998; Field *et al.* 2008).

Another issue related to the social aspects of dialogue studied previously is that of initiative in human–computer interaction, where the objective has been to make the interaction more 'natural', although the underlying sociolinguistic phenomena were not directly modeled. Some prior works in this area (cf. Chu-Carroll and Brown 1998; Core, Moore and Zinn 2003) separated the concepts of task initiative

and dialogue initiative in a manner that is similar to our distinction between task and topic control. However, most prior works concentrated on specific elements and instances of initiative control necessary for an artificial agent functioning in a collaborative dialogue setting (e.g., Whittaker and Stenton 1988), but were not directly concerned with the associated social phenomena, other than making the overall experience more acceptable to the human user. Similar studies of human–human conversational behavior related to group participation level can also be mentioned here (e.g., DiMicco, Pandolfo and Bender 2004). In contrast with the above, our interest in tracking the control of dialogue is focused on modeling sustained sociolinguistic behaviors by human speakers.

There is a body of literature in anthropology, linguistics, social psychology, and communication concerning the relationship between language and power, as well as other social phenomena, for example., conflict, leadership; however, the most existing approaches typically look at language use in situations where the social relationships are already established (e.g., doctor–patient) rather than using language predictively. For example, conversational analysis (Sacks, Schegloff and Jefferson 1974) is concerned with the structure of interaction: turn-taking, when interruptions occur, how repairs are signaled in specific social situations, while research in anthropology and communication has concentrated on how certain social norms and behaviors may be reflected in language (e.g., Agar 1994; Scollon and Scollon 2001). These paradigms provide us with valuable insights. However, there are few systematic studies in the current literature that explore the way in which language may be used to make *predictions* of social roles in groups where (a) these roles are not known *a priori*, or (b) these roles do not exist prior to the beginning of the discourse and only emerge through interaction. This is the case with the participants in most of our experiments with online discussion described in this paper.

Theories of communication have noted the function of language to exert force on others. Austin's (1962) Speech Act theory is a useful case in point. His theory advanced a performative view of language, noting that language acts on or has a force on others in communication. Searle's (1969) theory of speech acts articulates categories of speech with distinct forces on interlocutors, creating a power differential between them. For example, asking a question puts social pressure or expectations on the question–recipient to respond, often in preferred ways. This notion of force is important because it establishes the possibility that language in interaction constructs or allows for the negotiation of social roles.

A major focus of our research is the social role of leadership in multi-party discussion. Most research on leadership focuses on the personality characteristics of leaders. Some research has examined the ways that leaders *communicate* leadership. Hogg and Reid (2006) in their examination of the relationship between self-categorization and social norms within groups identify an effective leader as one who can 'transform individual action into group action by influencing others to embrace as their own and exert effort on behalf of, and in pursuit of, new group normative values, attitudes, goals, and behaviors' (p. 19). Although they do not specify how this manifests in practice, directives, introducing and controlling the

topic, and possibly expressing disagreement or rejecting others' ideas may all be ways to effectively communicate leadership. Searle (1969), in his theory of speech acts, notes that directives are a kind of act specifically done by leaders. Reid and Ng (1999) note that effective leaders control the conversation, and they do so in several ways, including adjacency pairs (question–response; offer–reply), or by directing comments at particular members. Another way a group is led is through initiating the topics of discussion (Pavitt *et al.* 2007; Pomerantz and Denvir 2007), and by managing the agenda or task process (Bonner 1959; Phillips 1973; Ketrow 1991; Ellis and Fisher 1994; Pavitt *et al.* 2007). Research has also shown that the volume of talk, that is, those who contribute the most, is a strong indication of leadership (Morris and Richard 1969; Stein and Heller 1983).

As the Internet-based communication channels have diffused, research has investigated the text-based communication affordances and limitations of these online platforms, attending especially to decision groups and the collaborative work of geographically distributed teams. Yoo and Alavi (2004) found that leaders in virtual teams send more and longer e-mail messages, produce more task-related messages, and enact the roles of initiating topics or tasks, schedule future meetings, and integrate and synthesize discussions. Similarly, Misiolek and Heckman (2005) found that leaders were more likely to initiate discussion in general, produce task or process-related messages, and be the targets of messages from others.

Other research has examined online discussion groups, such as Usenet forums, for evidence of leadership. Huffaker (2010) analyzed over 600,000 messages posted asynchronously to Google Groups on topics ranging from health to politics to science and technology. One of the challenges with investigating leadership is determining a 'ground truth' or establishing dependent variables, given that in spontaneously arising groups online, there is little opportunity to survey participants to establish perceived leadership within discussion groups. Huffaker's approach identified three dependent variables that are considered key to leadership practices: (a) the likelihood of creating messages that others respond to; (b) the likelihood of creating an initial post that sparks a long conversation thread; and (c) the degree to which the terms used are then repeated by others. His independent measures include (a) judging the rate of communication activity (e.g., number of posts); (b) identifying social networks of communication (e.g., expansiveness, reciprocity); and (c) characterizing language use (e.g., affect, linguistic diversity, and assertiveness). His results suggest that leaders post more messages, reply to more messages, are around longer in the groups, and have larger social networks (meaning that they reply to a broader set of people). He also found that leaders post longer messages, are more assertive (as measured by the use of certainty words), have more affect (as measured by the use of words with emotional valence), and greater linguistic diversity (as measured by the type/token ratio).

3 Sociolinguistic phenomena in discourse

We are interested in modeling a variety of sociolinguistic phenomena that occur in multi-party task-oriented discourse. These include social roles, such as leadership, as

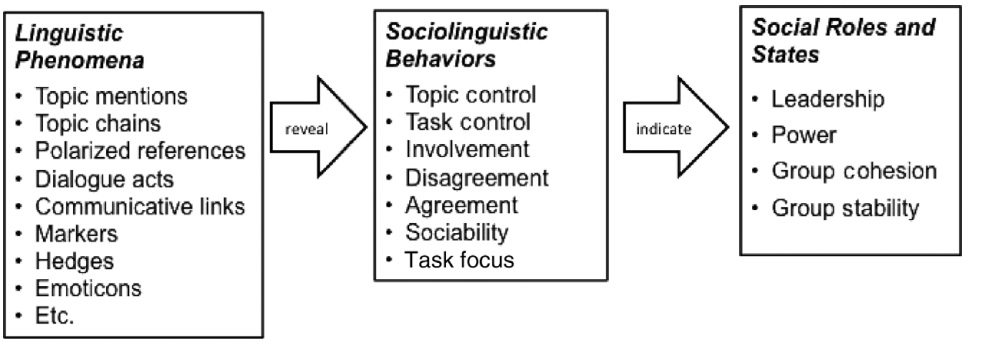

Fig. 1. A graphical hierarchy of sociolinguistic phenomena under investigation in the DSARMD project. Only a subset is discussed in this paper.

well as group states, such as group cohesion and stability. We describe intermediate measures needed to obtain reliable computational models of leadership; modeling of other sociolinguistic phenomena will be the subject of a future paper. Leadership can be defined as the ability to influence others to help accomplish group or organizational objectives (Chhokar, Brodbeck and House 2007); such influence is achieved through talk-in-interaction (Fairhurst 2007). A wide range of criteria have been used to describe leadership, from early theories focusing on a leader's traits and specific behaviors, to more recent constructs that recognize the importance of follower's perceptions and expectations, a leader's cognitive strengths and flexibility, charisma and values (Lowe, Kroeck and Sivasubramaniam 1996; Avolio, Bass and Jung 1999; Fairhurst 2007). Power, which is the ability to exert control over others (Ellis and Fisher 1994), is often vested in speaker's prominence, status, or position in relation to other members of a group (Bales 2001).

The research question is: can a high-level social phenomenon, such as leadership, be detected through analysis of the language of group interactions? In order to answer this question, we need to extend the current reach of discourse processing to encompass the analysis of *linguistic* elements that (a) reflect *sociolinguistic behaviors* of discourse participants, and in turn (b) reflect their status and social roles in the group. In other words, the sociolinguistic behaviors[1] link linguistic elements deployed in discourse (from lexical to pragmatic) to high-level social constructs (Social Roles and States) obtaining for and between the participants. Our hypothesis is that the Social Roles (e.g., leadership) and States (e.g., group cohesion) can be detected and attributed to discourse participants based on their sociolinguistic behavior in multi-party dialogue. The sociolinguistic behaviors, or language uses that we are currently studying are *agenda control*, *disagreement*, and *involvement* (Broadwell *et al.* 2010).[2]

The above chart (Figure 1) shows a partial hierarchy of sociolinguistic phenomena under investigation in the DSARMD project, a subset of which is discussed in

[1] An alternative term 'Language Use' (LU) has been adopted by the SCIL program (McCallum-Bayliss, 2010).

[2] Other sociolinguistic behaviors are also considered in our research, including collective behaviors such as 'sociability' but these are not directly relevant to the contents of this paper.

Table 1. *A quick reference guide to the sociolinguistic behaviors discussed in the paper and their measures*

Sociolinguistic behavior	Applicable measures	Acronyms used in paper	Commentary
Topic control	Local Topic Introduction	LTI	Scalar measures per speaker.
	Subsequent Mentions of Local Topics	SMT	
	Cite Score Index	CSI	
	Turn Length Index	TL	
Task control	Directive Index	DI	Scalar measure per speaker.
	Directive (followed by) Topic Shift Index	DTSI	Currently not implemented.
	Process Management Index	PMI	Scalar measure per speaker.
	Process Management Success Index	PMSI	Currently not implemented.
Involvement	Noun Phrase Index	NPI	Scalar measures per speaker.
	Turn Index	TI	
	Topic Chain Index	TCI	
	Allotopicality Index	ATP	
Disagreement	Disagree–Reject Index	DRX	Scalar measure pairs of speakers.
	Cumulative Disagreement Index	CDX	Scalar measure per speaker.
	Topical Disagreement Measure	TDM	Scalar measure pairs of speakers.[3]

this paper. These are classified into three levels. The linguistic phenomena include elements such as topic structures and topic chains in discourse, dialogue acts, communicative acts, and a variety of common linguistic devices such as markers and hedges. The mid-level sociolinguistic behaviors include both individual behaviors (topic control etc.) as well as collective behaviors (sociability and task focus). Finally, the high-level social constructs include roles (e.g., leadership) as well as states (group cohesion etc.), this last not directly discussed in this paper. The accompanying table (Table 1) provides a quick reference to the selected sociolinguistic behaviors that are defined in detail in the balance of this section along with the measures defined for assessment of their presence.

Our research so far has focused on the analysis of English language synchronous chat. Based on the literature cited in Section 2, we have defined a range of measures to capture the sociolinguistic behaviors that in turn define leadership roles. At the level of defining and detecting linguistic elements, multiple measures point to the same sociolinguistic behaviors; these are intentionally somewhat redundant and should correlate in their predictions of the presence of a sociolinguistic behavior. This multiplicity of measures addressing the same sociolinguistic behavior is deliberate:

[3] This measure is defined but currently not used in our system.

We are not sure yet which measures are most easily computed with sufficient accuracy or which are the most predictive ones. All are justified by the literature, however, and so their correlation is itself an evidence of their validity. One element of our research is to test our accuracy in computing them, and their predictive utility. In addition, we expect that some of these correlations may be culturally specific or language-specific, as we move into the analysis of Urdu and Mandarin discourses in the next phase of this project. Note that the relationship between sociolinguistic behaviors is different. Sociolinguistic behaviors do not necessarily point in the same direction with respect to the larger social phenomenon being explored, for instance, leadership. They are not redundant but intend to define different aspects of leadership. However, the balance between different sociolinguistic behaviors is also expected to change when we move to the investigation of leadership in groups using Urdu and Mandarin discourses.

3.1 Agenda control in dialogue

Agenda control is defined as efforts by a member or members of the group to advance the group's task or goal (e.g., Barnes 2005), thus an indicator of leadership. This is a complex sociolinguistic behavior that we will model along two dimensions: (1) *Topic Control*, and (2) *Task Control*. Topic control refers to attempts by any discourse participants to impose the topic of conversation. Task control, on the other hand, is an effort by some members of the group to define the group's project or goal and/or steer the group toward that goal. We believe that both behaviors can be detected using scalar measures per participant based on certain linguistic features of their utterances. In the following sections we discuss a number of specific hypotheses, illustrated by examples that form the basis for computational rules in our system.

3.1.1 Topic Control

Our first hypothesis is that topic control in dialogue is indicated by the rate of *local topic introduction* (LTI), where our notion of local topic follows that of Givon (1983). Local topics may be defined quite simply as noun phrases introduced into discourse, which are subsequently mentioned again via repetition, synonym, pronoun, or other forms of co-reference. We note that our definition of *local topic* is intended to look at topics at a 'micro' level. A local topic is anything that the participants discuss for some portion of the discourse. We use the term *meso-topic* (Section 3.3.2) when discussing topics at the level of the discourse as a whole. Our hypothesis is that one measure of topic control is the number of local topics introduced by each participant as a percentage of all local topics in a discourse.

In order to illustrate the concept of a local topic as used to construct the LTI metric, let us consider the following fragment of a multi-party online discussion, which we shall refer to as *YMCA-1*. This fragment comes from a much larger corpus of conversational data that we collected from online chat rooms as well as face-to-face discussions among multiple participants. The corpus currently comprises over 90 hours of conversations in multiple languages, including conversations in

English, Urdu, and Mandarin Chinese. The participants, all native speakers of their languages, were selected from a population varied in age and other social demographics. Additional face-to-face meetings were recorded and their transcripts added to our corpus. More details about our data collection process and resulting corpus can be found in Section 4 of this paper. YMCA-1, from which we draw most of the examples in this paper, is a 90-minute chat conversation among seven participants, covering approximately 700 turns. The dialogue centers on the task of hiring a counselor for a local YMCA from a set of candidates.

Example 1. *A fragment of online chat with underlines showing selected local topic references.*

1. LE: *I guess we should just start, not wait for CS and JY?*
2. JR: *sure*
3. KN: *ok*
4. LE: *Fundraising was Mark, Nanny was Carla, I think, if you were talking about my comment.*
5. JR: *gotcha- so that is not he most important to get this job….*
6. JR: *sorry about my typos- not used to this laptop yet*
7. JR: *wanna go thru carlas resume first ?*
8. KN: *sure*
9. LE: *Sure.*
10. KN: *i wonder how old carla is*
11. LE: *Ha, yeah, when I hear nanny I think someone older.*
12. KN: *she's got a perfect driving record and rides horses! coincidence?*
13. JR: *'06 high school grad*
14. KN: *i think she rides a horse and not a car!*

In this fragment, LE, JR, KN, etc. are the discourse participants (speakers) and each numbered line represents a separate turn in an online chat (i.e., the speaker submits the entire line to the chat room by hitting the RETURN key on the computer keyboard). Turns are listed in the order they arrive in the chat room, which occurs at a fairly robust pace (usually a few seconds apart).[4] In this small excerpt, a few local topics are introduced, including *nanny*, *Carla*, and *horses*, as well as possibly others. These local topics are underlined in different ways, with the first mention set in boldface. For example, *Carla* is introduced by LE in turn 4, and is subsequently mentioned by JR (turn 7), KN (turn 10), and again by KN (via *she*) in turns 12 and 14. Similarly, KN introduces horses in turn 12, and then self-mentions it again in turn 14.

We note that this discourse is not about horses, but *horses* is still a local topic in the discourse, by our definition, because it is introduced by one of the speakers and subsequently mentioned again by repetition. Our LTI metric is intended to count all such local topics.

[4] Due to the asynchronous nature of the online chat, some turns may occur near simultaneously.

Table 2. *Topic control distribution among seven speakers in the YMCA-1 dialogue based on the LTI metric*

	LTI	
Speakers	Index value (%)	Degree of topic control
JR	24	4
LE	26	5
KN	16	3
KI	16	3
CS	11	2
KA	7	2
JY	1	1
	Distribution statistics	
Mean		14%
80 percentile		25%
Std. Dev.		9%

Using an LTI index we can construct assertions about topic control in a discourse. For example, suppose the following information is discovered about the speaker LE in the YMCA-1 discourse, where ninety local topics are identified:

(1) LE introduces 23/90 (25.6%) of local topics in this dialogue.
(2) The mean rate of LTI in this dialogue is 14.29%, and standard deviation is 9.01.
(3) LE is in the top quintile of participants for introducing new local topics.

Based on the above information, we can now claim the following, with a degree of confidence[5]:

$TopicControl_{LTI}$ (*LE, 5, YMCA-1*)

We read this formula as follows: Speaker LE exerts the highest degree (5 on scale 1–5) of topic control in the YMCA-1 dialogue.

Analogous claims may now be made about each speaker in the dialogue based on their rates of LTI. As a result, we arrive at the following tentative assessment of topic control in this group, as shown in Table 2.

The use of a numerical scale (apart from the relative ranking of the participants by index values) allows for making informative claims about individual participants without displaying the entire group ranking. In a quintile-based scale used here, the degree of 5 is equivalent to being in the 5th (highest) quintile, that is to say, displaying more measurable behavior features than 80% of other members of the group. Arguably, it is a fairly simplistic metric; however, development of any more advanced mapping requires further research into the meaning of relative index values, that is to say, how a 2% change in LTI translates into an influence differential in discourse. Here we should note that the quintile scale was not used in evaluating

[5] The degree of confidence will be determined experimentally in the future research.

the system performance, but it was found highly intuitive by potential users of such a metric.

Of course, LTI is just one source of evidence, and we developed other metrics to complement it. We mention four of them here:

- *SMT Index*: This is another measure of topic control that we developed and is based on subsequent mentions of already introduced local topics. Speakers who introduce local topics that are discussed at length by the group tend to control the topic of the discussion. In order to capture this phenomenon, we create the *subsequent mentions of local topics* (SMT) index, which calculates the percentage of second and subsequent references to the local topics, by repetition, synonym, or pronoun, relative to the speakers who introduced them.
- *Cite Score*: This index measures the extent to which other participants discuss topics introduced by each speaker. The difference between SMT and Cite score is that the latter reflects the degree to which other participants in a conversation assent to a speaker's efforts to control the topic, as shown by their utterances. In other words, while SMT includes self-citations that reflect the speaker's effort to sustain the topic of conversation, Cite score is a more objective measure of the successful of these efforts.
- *TL Index* (TL): This index stipulates that more influential speakers take longer turns than those who are less influential. The TL index is defined as the average number of words per turn for each speaker. Our hypothesis is that TL reflects the extent to which other participants are willing to 'yield the floor' in conversation, which is a fairly straightforward measure in spoken conversation but less obvious in synchronous chat.

Like LTI, all the above indices are mapped onto a 5-point scale representing varying degrees of topic control for discourse participants. Table 2 shows the distribution of index values for all four indices associated with the topic control sociolinguistic behavior derived from the YMCA-1 dialogue. Other index values are calculated analogously to LTI, as shown previously. It is interesting to note that the indices appear to capture a sense of the internal dynamics within this group of speakers, where LE, KN, and JR are clearly the more influential participants, with LE standing out as having the highest degree of topic control by all measures. As noted earlier, we currently map degrees of topic control to quintiles of index value distribution; this simple method may be refined in the future by considering the distance from the mean instead (e.g., at least $\alpha\times$ Std. Dev, etc.).

As already described, we have deliberately defined multiple indices that point to the same sociolinguistic behaviors. Ideally, we would like all the indices shown in Table 3 (and others yet to be defined) to predict the same outcome, that is to say, for each dialogue participant each index should assign the same degree of topic control, relative to other speakers. In this dialogue at least, the indicators are not completely aligned but show a fairly good degree of correlation, which is further illustrated by the chart in Figure 2 that plots topic control predictions for each speaker using all four metrics (correlation tables are found in Section 3.4). This correlation, which has

Table 3. *Topic control distribution in the YMCA-1 dialogue. Each row represents a speaker in the group (LE, JR, etc.). Columns show indices used, with their values (%) and the assigned degrees of topic control per speaker on the 5-point scale based on quintiles in value distributions*

	LTI		SMT		CSI		TL	
Speakers	(%)	(deg.)	(%)	(deg.)	(%)	(deg.)	(%)	(deg.)
JR	24	4	21	4	19	4	15	3
LE	26	5	35	5	35	5	25	5
KN	16	3	22	4	24	4	15	3
KI	16	3	8	2	7	3	8	1
CS	11	2	11	3	12	3	14	2
KA	7	2	3	2	3	2	15	3
JY	1	1	0	1	0	1	9	2
Distribution statistics								
Mean (%)	14		14		14		14	
80 percentile (%)	25		27		28		19	
Std. Dev. (%)	9		12		12		6	

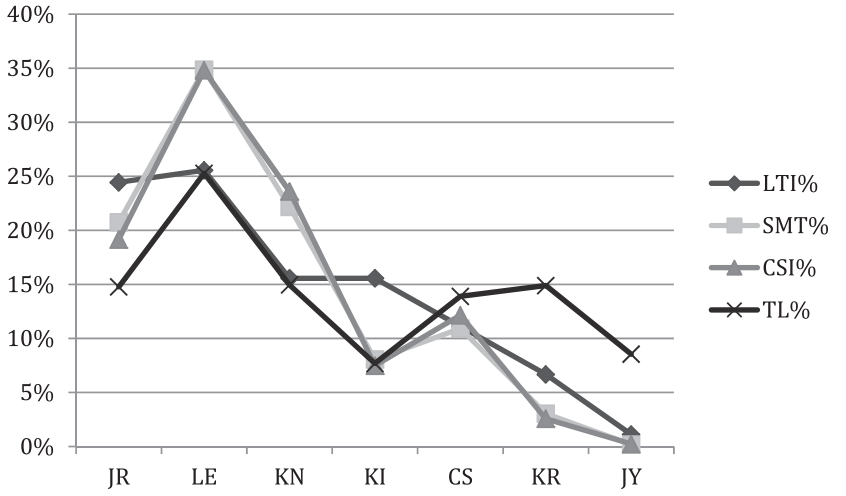

Fig. 2. Correlation of topic control metrics for YMCA-1 dialogue. Similar strong correlations were also noted for other datasets (see also Tables 9A and 9B in Section 3.4).

been noted across nearly all datasets that we examined, supports our claim that all five indices are measuring the same phenomenon, namely topic control in discourse. Moreover, as we will see in a further section of this paper, this correlation is also supported by the human assessment of speaker influence.

Nonetheless, we do not expect these correlations will always hold, and where individual indices divert in their predictions, our level of confidence in the generated social behavior claims decreases. For example, while SMT, CSI, and TCI align very well, LTI and TL show more variability. At least among the social group represented

in our experiments thus far (English speaking college students and recent graduates, to the extent they may be considered a social group), participants who have the highest degree of topic control also have longer turns. While turn length (TL) index correlates with other measures of topic control in our studies, we also recognize the possibility that there may be cultural or sub-cultural variation on this point. We are currently conducting experiments to determine an optimal way of combining values of individual metrics in order to maximize the accuracy of topic control claims. Later in this paper, we discuss a baseline predictor for topic control based on an unweighted linear combination of a subset of these metrics.

3.1.2 Task Control

The other aspect of the Agenda control phenomenon is task control. It is defined as an effort to determine the group's goal and/or steer the group toward that goal. Unlike topic control, which is imposed by influencing the subject of conversation, task control is gained by directing other participants to perform certain tasks or accept certain opinions. Consequently, task control is detected by observing the usage of certain dialogue acts, including action-directive, agree–accept, disagree–reject, in addition to any process management statements that may take the form of directives, assertion–opinions, or questions. The dialogue fragments shown in Example 2 contain several instances of such utterances. In this example, KI and JD may be seen as attempting to control the task.

Example 2. *A fragment of online chat with dialogue act annotation. The PRCS–MGMT label indicates utterances classified as process management.*

1.	KI: (PRCS–MGMT: Confirmation–Request) *So if we were to weed them out than it seems as though the 3 we would interview would be Emily, Jason and Carla, right?*
2.	LE: (Agree–Accept) *true*
3.	JE: (PRCS–MGMT: Assertion–Opinion) *we only interview those that have potential*
4.	JR: (PRCS-MGMT: Assertion–Opinion) *Or we can interview everyone*
5.	LE: (PRCS–MGMT: Disagree–Reject) *we can't interview everyone ...*
6.	JR: (Action–Directive) *Why don't we figure out what we'd ask Carla*
7.	LE: (Agree–Accept) *good question*
8.	JR: (Action–Directive) *Then pick our next choice*
9.	KA: (Agree–Accept) *ok*
10.	JR: (Action–Directive) *And figure out what to ask them*

We define several indices that allow us to compute a degree of task control in dialogues for each participant.

- *Directive Index* (DI): The participant who directs others is attempting to control the course of the task that the group is performing. We count the number of directives, that is to say, utterances classified as action–directive, made by each participant as a percentage of all directives in discourse.

- *Directed Topic Shift Index* (DTSI): When a participant who controls the task offers a directive on the task, the topic of conversation shifts. In order to detect this condition, we calculate the ratio of coincidence of directive dialogue acts by each participant with topic shifts following these directives.
- *Process Management Index* (PMI): Another measure of task control is the proportion of turns from each participant that explicitly address the problem-solving process. This includes utterances that involve coordinating the activities of the participants, planning the order of activities, etc. These fall into the category of Process (or Task) management in most DA tagging systems.
- *Process Management Success Index* (PMSI): This index measures the degree of success by each speaker at controlling the task. A credit is given to the speaker, whose suggested course of action is supported by other speakers, for each response that supports the suggestion. Conversely, a credit is taken away for each response that rejects or qualifies the suggestion. PMSI is computed as a distribution of task management credits among the participants over all dialogue utterances classified as process management.[6]

As an example, let us consider the following information computed for the PMI index over the YMCA-1 dialogue:

(1) YMCA-1 contains 246 utterances classified as process management rather than doing the task.
(2) Speaker KI makes sixty-five of these utterances for a PMI of 26.4%.
(3) Mean PMI for participants is 14.3%; 80th percentile is >21.2%. PMI for KI is in the top quintile for all participants.

Based on this evidence, we may claim (with yet to be determined confidence) that

$$TaskControl_{PMI}(KI,\ 5,\ YMCA\text{-}1)$$

This may be read as follows: *Speaker KI exerts the highest degree of task control in the YMCA-1 dialogue*. We note that task control and topic control do not coincide in this discourse, at least based on the PMI index.

The other index values for task control may be computed and tabulated in a similar way. This is captured in Table 4 that shows the values of task control indicators for the YMCA-1 dialogue.

From this analysis a different picture of the YMCA-1 chat group emerges, with speakers KI, KA, and JR coming across as more effective at steering the group's task, and LE being less effective here. Taking together the analyses given in Tables 3 and 4 we further note that only JR is both highly influential and highly effective (if not necessarily the highest on either scale), which may point to this speaker as a possible leader of the group. But before we attempt such a claim, let us look at other social behaviors first.

[6] The exact structure of the credit function is still being determined experimentally. For example, more credit may be given to the first supporting response and less for subsequent responses; more credit may be given for unprompted suggestions than for those that were responding to questions from others.

Table 4. *Task control distribution in the YMCA-1 dialogue. Each row represents a speaker in the group (LE, JR, etc.). Columns show indices used, with their values (%) and the assigned degrees of topic control per speaker on the 5-point scale based on quintiles in values distribution*

	DI		DTSI		PMI		PMSI	
Speakers	(%)	(Deg.)	(%)	(Deg.)	(%)	(Deg.)	(%)	(Deg.)
JR	44	5	29	4	23	4	19	4
LE	<1	1	<1	1	10	2	9	2
KN	<1	1	<1	1	15	3	12	3
KI	22	4	29	4	27	5	24	5
CS	11	2	14	2	4	1	4	1
KA	11	2	14	2	16	4	17	4
JY	11	2	14	2	5	2	6	2
Distribution statistics								
Mean (%)	11		14		15		12	
80 percentile (%)	31		29		24		21	
Std. Dev. (%)	15		12		9		7	

The task control metrics represented in Table 4 are still under development as we continue to research their operational definitions; nonetheless, they already show a reasonable correlation with each other. We also note that while task control and topic control may correlate for some speakers, it shows significant variance for others, for instance, LE. We shall return to this point later on.

3.2 Involvement in dialogue

The next type of social behavior that we discuss here is Involvement. It is defined as the degree of engagement or participation in the discussion of a group. In our experiments with American participants, high involvement and influence (topic control) often correlate with group leadership, as shown in Section 6. While involvement is an important element of leadership, we also recognize that its importance may differ between cultures. In our ongoing research on Urdu and Mandarin Chinese, we are investigating cultural variability in this area. Involvement in a conversation may be manifested through substantive contributions to its content or simply by taking frequent turns, even though these do not appear to add much to the discussion. Furthermore, participation in discussion of other participants' topics as well as more popular topics is a strong indicator of one's involvement. Involvement is harder to illustrate using a brief example; however, in the fragments shown earlier in Examples 1 and 2, speaker JR may be seen as more involved by taking frequent turns and responding to topics and issues that other speakers raise.

In order to measure involvement, we designed several indices based on turn characteristics for each speaker. The four indices we currently use are briefly explained below:

- *NP Index* (NPI): This index is a measure of gross informational content contributed by each speaker in discourse. NPI counts the ratio of nouns and third-person pronouns used by a speaker to the total number of nouns and pronouns in the discourse. We note that NPI makes no distinction between a local topic or its subsequent mention and any other noun phrase and thus is a much cruder measure than LTI and SMT indices used in topic control.
- *Turn Index* (TI): This index is a simple measure of *interactional frequency*; it counts the ratio of turns per participant to the total number of turns in the discourse. This measure reflects the hypothesis that more frequent speakers come across as more involved.
- *Topic Chain Index* (TCI): This index measures the degree to which participants discuss the most persistent topics, which allows us to capture selective involvement on the part of some speakers. In order to calculate TCI values, we define a concept of *topic chain*, which comprises the first and all SMTs. The longest of these chains indicate the most persistent topics in discourse. TCI computes frequency of mentions in these longest chain topics for each participant.
- *Allotopicality Index* (ATP). This index measures speaker involvement by counting the number of a speaker's mentions of local topics that were introduced by other participants. We call such mentions allotopical. ATP reflects the willingness of a speaker to actively consider other people's contributions to the discussion. An ATP value is the proportion of a speaker's allotopical mentions, that is to say, excluding 'self-citations', to all allotopical mentions in a discourse.

As an example, we may consider the following situation in the YMCA-1 dialogue:

(1) YMCA-1 contains 796 third-person nouns and pronouns, excluding mentions of participants' names.
(2) Speaker JR uses 180 nouns and pronouns for an NPI of 22.6%.
(3) The median NPI is 14.3%; JR is in the top quintile of participants (>19.9%).

From the above evidence we can draw the following claim:

$Involvement_{NPI}(JR,\ 5,\ YMCA\text{-}1)$

This may be read as: *Speaker JR is the most involved participant in the YMCA-1 dialogue*. Using the NPI index we can calculate a degree of involvement for each participant in the YMCA-1 dialogue. This is shown in Table 5.

Other measures of involvement can be combined into a two-dimensional map capturing the group internal dynamics as shown in Table 6.

From Table 6 yet another picture of the YMCA-1 chat group emerges: Speakers KI, and JR are seen as the most involved, while LE and KA are somewhat less involved. We note that the predictions of all four indices are well correlated in this dataset, which further strengthens our claim (Figure 3).

We may now combine, at first informally, the assessments of the three social behaviors discussed thus far in an attempt to obtain a clearer picture of the group dynamics within the YMCA discussion. In particular, we make an initial projection about leadership among the speakers in this discourse. This is shown

Table 5. *Involvement distribution among the group of seven speakers in the YMCA-1 dialogue based on the NPI metric*

	NPI	
Speakers	Index value (%)	Degree of topic control
JR	23	5
LE	20	4
KN	14	3
KI	19	4
CS	7	2
KA	15	3
JY	2	1
	Distribution statistics	
Mean		17%
80 percentile		21%
Std. Dev.		7%

Table 6. *Involvement distribution in the YMCA-1 dialogue. Each row represents a speaker in the group (LE, JR, etc.). Columns show indices used, with their values (%) and the assigned degrees of topic control per speaker on the 5-point scale based on quintiles in value distributions*

	NPI		TI		TCI		ATP	
Speakers	(%)	(Deg.)	(%)	(Deg.)	(%)	(Deg.)	(%)	(Deg.)
JR	23	5	21	4	25	5	23	5
LE	20	4	11	2	17	3	13	3
KN	14	3	13	3	13	3	11	2
KI	19	4	31	5	21	4	24	5
CS	7	2	5	2	6	2	5	2
KA	15	3	15	4	16	2	20	4
JY	2	1	4	1	3	1	4	1
	Distribution statistics							
Mean (%)	17		13		15		13	
80 percentile (%)	21		25		22		23	
Std. Dev.	7		9		7		7	

in Table 7, where we combined predictions for each sociolinguistic behavior by averaging the component index values for each participant into three categories: high (corresponding to average degree of between 4 and 5), medium (corresponding to average degree of less than 4 but more than 2), and low (for average degree of 2 or less).

We can see from Table 7 that JR is the only speaker in the group who is highly ranked on all three sociolinguistic behaviors, a likely leader and definitely a powerful participant. Other speakers seem to also manifest some leadership qualities in this conversation as a result of the combination or high degree of topic control

Table 7. *Leadership matrix for the YMCA-1 dialogue based on three languages uses*

Speakers	Topic control	Task control	Involvement	Leader?	Power?
LE	High	Low	High	No	Yes
JR	High	High	High	Yes	Yes
KI	Medium	High	Low	No	Some
KN	High	Low	Medium	No	Yes
KA	Low	Medium	Medium	No	Some
CS	Low	Low	Low	No	No
JY	Low	Low	Low	No	No

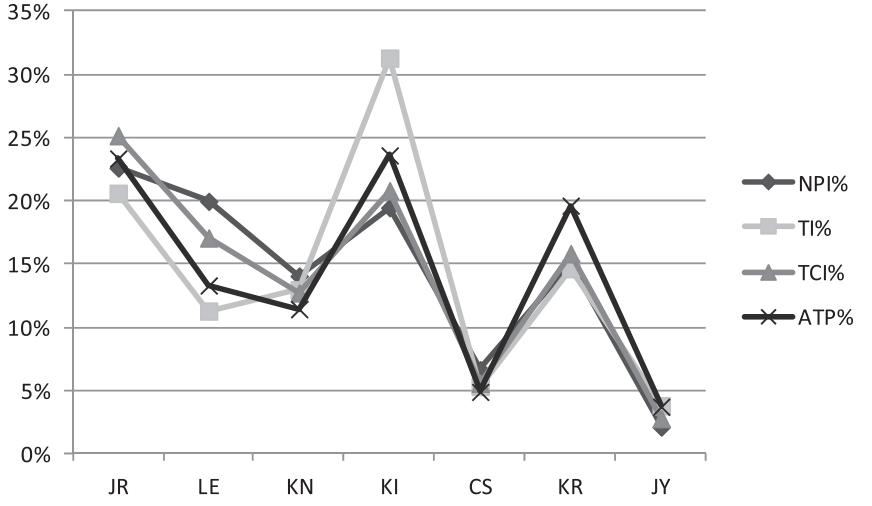

Fig. 3. Correlation of involvement metrics for YMCA-1 dialogue. Similar strong correlations were noted for all other datasets (see also Table 9B in Section 3.4).

(influence) and Involvement. Interestingly, speakers KA and KI are both effective (Task Control) and sometimes involved (KA) but lack influence (topic control), which is necessary for leadership. One possibility is that they may be in conflict with JR, perhaps representing opposing or dissenting positions in the discussion. In order to verify this possibility, we need to consider additional sociolinguistic behaviors that capture disagreement. The above analysis, while informal and preliminary, is based on the strong correlation between the sociolinguistic behaviors discussed thus far and the group leadership; this correlation is reported later in this paper (Tables 13 and 13A in Section 6.2.1). Furthermore, in Section 7, we briefly mention preliminary results of using combinations of sociolinguistic behaviors for automated detection of leadership based on the same analysis. We note, however, that a complete discussion of the leadership phenomenon in group dialogues is outside of the scope of the present paper.

3.3 Disagreement in dialogue

Disagreement is another sociolinguistic behavior that correlates with a speaker's leadership and is commonly deployed by leaders to discourage what may be perceived as attempts to challenge the leader role by other group members. There are two ways in which this behavior is realized: *expressive disagreement* and *topical disagreement* (Stromer-Galley 2007). Both can be detected using scalar measures applied to subsets of participants, typically any two participants. In addition, we can also measure for each participant the rate at which he or she generates disagreement with any and all other speakers. We shall discuss both types of disagreements below; however, in the current system we only implement a single combined Disagreement SLB.

3.3.1 Expressive disagreement

Expressive disagreement is normally understood at the level of dialogue acts, that is, when discourse participants make explicit utterances of disagreement, disapproval, or rejection in response to a prior speaker's utterance. Here is an example taken from the YMCA-1 dialogue where one of the candidates for the youth counselor job is a young woman named Carla.

KA: *CARLA… women are always better with kids*
KI: *That's not true!*
KI: *Men can be good with kids too*

While such exchanges are vivid examples of expressive disagreement, we are interested in a more sustained phenomenon where two speakers repeatedly disagree, thus revealing a social relationship between them. Therefore, one measure of expressive disagreement that we consider is the number of disagree–reject dialogue acts between any two speakers as a percentage of all utterances exchanged between these two speakers. This becomes the basis for the *disagree–reject index* (DRX). In the YMCA-1 dialogue we have the following:

(1) Speakers KI and KA have forty-seven turns between them. Among these there are eight turns classified as disagree–reject, for the DRX index of 15.7%.
(2) (2) The mean DRX for speakers in this dialogue who make any disagree–reject utterances is 9.5%. The pair of speakers KI–KA is in the top quintile (>13.6%).

Based on this evidence, we can conclude the following:

$ExpDisagreement_{DRX}$ (*KI, KA, 5, YMCA-1*)

This expression may be read as follows: *Speakers KI and KA have the highest level of expressive disagreement in the YMCA-1 dialogue.*

The DRX measure ranks *pairs* of speakers by the degree of expressive disagreement between them. This ranking may not include all speaker pairs; for example, speakers who rarely address each other in a particular section of a dialogue may have no disagree–reject utterances between them. Nonetheless, DRX may be helpful

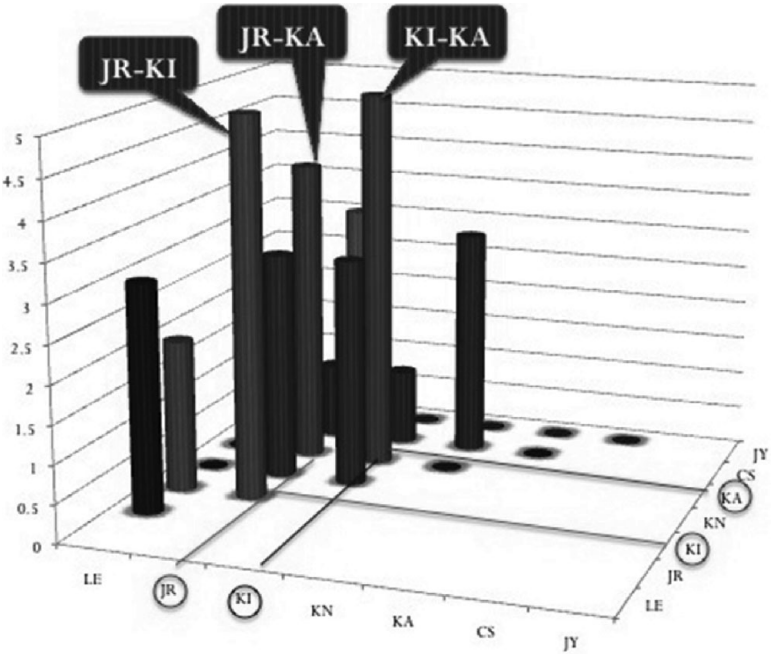

Fig. 4. Distribution of expressive disagreement between pairs of speakers in the YMCA-1 dialogue based on the DRX index.

in assessing some of the internal working structures of the group, including detection of any significant subgroups or factions within. Figure 4 illustrates the distribution of pair-wise expressive disagreement in the YMCA-1 dialogue: We note that most of the expressive disagreement occurs between three pairs of speakers JR, KI, and KA. These are also the three speakers that have the most task control in this dialogue.

Another way to characterize disagreement is as a speaker's propensity to generate expressions of disagreement or rejection toward other speakers in the group. Capturing this behavior may be important in gauging a speaker's standing within the group, including with respect to leadership. In order to measure this aspect of expressive disagreement, we developed the *cumulative disagreement index* (CDX), which is computed for each speaker as the percentage of all disagree–reject utterances in the discourse that are made by this speaker. Unlike DRX, which is computed for pairs of speakers, the CDX values are assigned to each discourse participant and indicate the degree of disagreement that each person generates.

In the YMCA-1 dialogue we find the following:

(1) Speaker JR has 139 turns in discourse. Among these there are fifteen turns classified as disagree–reject directed toward other speakers, for the CDX index of 32%.
(2) The mean CDX for speakers who make any disagree–reject utterances is 9%. The speaker JR is in the top quintile (>31%) for this measure.

Based on this evidence we can conclude the following:

Table 8. *The combined language use matrix for the YMCA-1 dialogue shows JR as a leader*

Speakers	Topic control	Task control	Involvement	Disagreement
LE	High	Low	High	Low
JR	High	High	High	High
KI	Medium	High	Low	High
KN	High	Low	Medium	Medium
KA	Low	Medium	Medium	Medium
CS	Low	Low	Low	Low
JY	Low	Low	Low	Low

$Disagreement_{CDX}$ (*JR, 5, YMCA-1*)

This expression may be read as follows: S*peaker JR generates the highest level of disagreement toward other participants in the YMCA-1 dialogue.* Disagreement levels for other speakers are calculated analogously. The reader may note that we use the *disagreement* predicate here, rather than the *expdisagreement* cited above. It is of course a different predicate (a different number of arguments, since it measures a speaker's disagreements with all participants); furthermore, in the next section, we will expand the scope of the disagreement measure beyond that of expressive disagreement.

We may now add a new column to Table 7 showing the distribution of disagreement computed according to the CDX measure. As before, we map the 5-point scale into a three-way classification for easy comparison. The results are shown in Table 8.

Based on the analysis thus far we may postulate the following about the group involved in the YMCA-1 dialogue[7]:

- JR is the leader: This speaker is influential, effective, involved, and assertive (based on his willingness to argue, generate disagreement for his position); based on the experimental data that we have collected thus far (including human evaluations), topic control (influence) and task control (effectiveness) are both strongly correlated with leadership, while the other two behaviors supply supporting but not decisive evidence. Participants who are high on both topic control and task control are nearly always perceived as leaders by human evaluators.
- LE and KN are influential (topic control) but not leading: Both show low effectiveness (task control); nonetheless, they are definitely influential participants.
- KI and KA are effective (task control) but less involved and less influential: They are in conflict with the leader as well as with each other (Figure 4).

[7] These assessments are confirmed by the participants' post-session questionnaires that will be discussed later in this paper (Section 6.2, especially Tables 13 and 13A).

- CS and JY are marginal participants: They are neither involved nor influential, they are not effective, and avoid conflict.

3.3.2 Topical disagreement

Another form of disagreement found in discourse is disagreement of opinions on a topic. While expressive disagreement is based on the use of more overt linguistic devices, topical disagreement tends to be subtler, and may be defined as a difference in referential *polarity* in utterances (statements, opinions, questions, etc.) made on a topic. Referential polarity (or valence) of an utterance is determined by the type of statement made about the topic in question, which can be positive, negative, or neutral. A positive statement is one in favor of (*express advocacy*) or in support of (*supporting information*) the topic being discussed. A negative statement is one that is against or negative on the topic being discussed. A neutral statement is one that does not indicate the speaker's position on the topic. Below is an example of opposing polarity statements about the same topic in discourse, a job candidate who is identified only as 'he' in the following exchange.

LE: *I like that he mentions* '*Volunteerism and Leadership*'
JR: *but if they're looking for someone who is experienced then I'd cross him off*

The first speaker makes a clearly supportive statement about the candidate, while the second speaker's response is just as clearly unfavorable. The reason for this assessment is not hard to see: both speakers make their positions on the subject quite explicit ('*I like that he…*', '*I'd cross him off*') and moreover the second utterance takes the form of an expressive disagreement ('*but…*') with the first speaker. Nonetheless, detecting topical disagreement in discourse is often quite complex and requires careful judgment because the strength and character (less explicit, more nuanced) of topical disagreement may vary from one topic to another and from one conversation to the next.[8]

While each pair of oppositely polarized utterances, like the two above, may signal topical disagreement, we are interested in a sustained phenomenon where speakers repeatedly (although not exclusively) make opposing polarity (positive–negative) statements on one or more topics. In other words, speakers who have a high proportion of opposing polarity statements about a topic are said to topically disagree to a high degree on this topic. Conversely, speakers who make proportionally few opposing polarity statements about a topic while making more of other sorts of statements on this topic are said to have a low degree of topical disagreement. The same speakers may have varying degrees of disagreement on different topics; therefore, it is important to separately estimate their disagreement for each topic. In the following, we discuss the concept of pair-wise topical disagreement first before

[8] Extra-linguistic cultural and contextual elements may also come into play, for example, when invoking certain concepts (such as 'volunteerism') signals a positive statement even without an explicit attitude marking. In the current work, we limit our analysis to situations where positive and negative valences are linguistically marked.

returning to the notion of the individual sociolinguistic behavior embodied in the $Disagreement_{CDX}$ predicate introduced at the end of the previous section.

Topical disagreement between any two speakers is detected by estimating the extent to which they apply differential polarity opinions to a topic in conversation, as conveyed in these speakers' utterances. In this definition, the topic (or an 'issue') is assumed to include only the most persistent and important local topics in discourse about which polarized statements are made: we shall call these *meso-topics*. For example, in a discussion of job applicants, each of the applicants becomes, potentially, a meso-topic; in a political discourse each key personality, organization, event, or media outlet may become a meso-topic, and there may be additional meso-topics present, such as 'qualifications required' etc.

More formally, a meso-topic is a local topic that satisfies the following two criteria: (a) it has a degree of persistence in the discourse, and (b) the speakers make polarized statements about it. Topic persistence is defined as a length of a continuous chain of utterances mentioning this topic, relative to other chains and to the overall dialogue length. Moreover, short 'gaps' in the chain are permitted (up to ten turns to accommodate digressions, obscure references, noise, etc.). The second criterion narrows the set of persistent topics to only those where a disagreement is likely to occur. For example, in the following fragment 'Mark' is a likely meso-topic (based on the second criterion):

JR: *Ok, resume for Mark*
LE: *I like that he mentions 'Volunteerism and Leadership'*
(Meso-Topic*: Mark;* Polarity*: positive*)
JR: *but if they're looking for someone who is experienced then I'd cross him off*
(Meso-Topic*: Mark;* Polarity*: negative*)

Accordingly, our measure of topical disagreement between any two speakers in a conversation is the differential in net polarity of each speaker's utterances about a meso-topic, even when they are not directly addressing each other.

The resulting *topical disagreement metric* (TDM) captures the degree to which any two speakers advocate the opposite sides of a meso-topic. In order to compute TDM, we first determine each speaker's position on the meso-topic, then calculate the distance between their positions. When two speakers' positions are the farthest apart relative to the distances calculated for other speaker pairs, their degree of topical disagreement is the highest for the group on that meso-topic; conversely, when their positions are the closest to each other, their degree of topical disagreement is the lowest.

Speakers can make utterances of varying polarity about the same meso-topic, and in order to determine their overall position on the topic, we need to estimate the balance, or net polarity of all such utterances. This may be done by simply counting the positive and negative utterances, but complications quickly arise when the discussion of various sides of a topic is not sufficiently balanced. For example, it is often the case that a significant length of discussion focuses on one side of a topic, for instance, lack of experience in a job candidate, leading to a skewed distribution of polarized utterances by some speakers. Therefore, any counts of

polarized utterances attributed to one speaker must always be considered in relation to what other speakers say on the same topic.

We compute TDM values for each pair of speakers, and for each meso-topic in discourse. In order to do so, we count all polarized utterances made by each speaker on the topic – ones addressed to the other speaker as well ones addressed to anyone else – as percentages of the same-polarity utterances on this topic made by all speakers in discourse. This gives us the relative position of each speaker on both positive and negative sides of the topic. In order to make an accurate metric, we need to take into account all utterances made by these speakers that

- make explicit positive or negative statements about the meso-topic, for instance, '*I'm for Carla."* or '*I would not vote for her';*
- offer either supportive or unsupportive information about this meso-topic; for instance, '*She's got experience with youngsters",* or '*If you are looking for computer skills I would cross her off'*.
- respond to other speakers' polarized statements with agreement or disagreement, as appropriate.

Strong prevalence of positive or supportive statements about a meso-topic, as well as agreements with others' positive or supportive statements, relative to other speakers, indicates a strong favorable position of the speaker on this meso-topic. Conversely, strong prevalence of negative or unsupportive statements about a meso-topic, as well as agreements with others' negative or unsupportive statements, relative to other speakers, indicates a strong unfavorable position on this meso-topic. When the imbalance of positive/supportive and negative/unsupportive utterances on a meso-topic is less pronounced, the speaker position becomes less polarized and more neutral. Therefore, a speaker position on a meso-topic can range from strongly favorable to favorable to neutral to unfavorable to strongly unfavorable, based on the balance among his polarized utterances made on the meso-topic.

Using TDM we can construct claims concerning topical disagreement in a given multi-party dialogue, but the dialogue must be of sufficient duration and complexity for this measure to show useful results.[9] Using our running example of the YMCA-1 discourse (which is a 90-minute chat dialogue on job candidates among seven participants, and is sufficiently complex), we find that speaker KA makes 30% of all positive utterances made by anyone about Carla (40), while KI makes 45% of all negative utterances against Carla. This places these two speakers in the top quintiles in the 'for Carla' polarity distribution and 'against Carla' distribution, respectively. If each of these speakers always made statements of opposing polarities about 'Carla', their topical disagreement 'distance' would be the greatest possible, or (KApro-quintile + KIagainst-quintile)/2 = 5. However, in reality, most speakers make both positive and negative remarks about a meso-topic, and as a result their mutual positions are less extreme. For example, while KI is highly negative about Carla, she still allows a few positive statements, which places her in the lowest quintile on the pro-Carla scale. As a result, her position on Carla, while still highly negative,

[9] Exactly what constitutes a sufficient duration or complexity is still being researched.

shifts from degree 5 to degree 4 (KIposition = KIagainst-quintile − KIpro-quintile). Similarly, taking into account any opposing polarity statements made by KA against Carla, we calculate the level of topical disagreement between KA and KI to be 4 (KIposition − KAposition)/2 on the 1–5 scale.

Based on this analysis we can make the following claim:

$$TopDisagreement_{TDM}((KI, KA), \text{`}Carla\text{'}, 4, YMCA\text{-}1)$$

This may be read as follows: *Speakers KI and KA topically disagree to degree 4 on topic (job candidate) 'Carla' in the YMCA-1 dialogue.*

Similar to the DRX measure of pair-wise expressive disagreement, TDM allows for computing topical disagreement between any two speakers in a discourse. The difference here is the scope: While DRX looks only at direct exchanges between the two speakers involved, TDM counts all polarized statements made by these speakers, thus revealing another interesting aspect of internal group dynamics, that is, speakers' positions on issues.

A further generalization we are considering is to determine speakers' relative positions with respect to a set of key topics in conversation, for example, in a political debate. Instead of computing topical disagreement on each topic separately, we build an n-dimensional vector that represents a speaker's position on each of the topics. In other words, for each speaker s, we construct $PV_s = (pv_s(t_1) \ldots pv_s(t_n))$, where $pv_s(t_i)$ is a numerical value representing this speaker's position on topic t_i, which is in turn a relative differential of polarized utterances made on this topic, as described above. The distance between speakers may now be measured using any reasonable vector distance metric, for instance, cosine in the vector space.

Having explored ways of detecting and measuring pair-wise topical disagreement in discourse, we now return to our earlier concept of an individualized measure of disagreement behavior in speakers, that is, a speaker's overall tendency to generate expressions of disagreement or rejection toward other speakers in the group. Putting aside for now any pair-wise relationships or even specific topics, we want to see how topical disagreement can contribute to the overall disagreement SLB for each speaker. As with expressive disagreement, this requires measuring disagreement per speaker rather than for each pair of speakers. The CDX index described in the previous section captures an individual disagreement level based on explicit utterances of expressive disagreement by counting instances of disagree–reject dialogue acts. We now augment this index by also counting the utterances where the speaker topically disagrees with a prior statement by another speaker. To do so, we track how each utterance is addressed when the speakers respond to one another in conversation. Specifically, we note if (a) the speaker's statement appears as a response to a recent statement by another speaker, (b) both statements reference the same meso-topic, and (c) two references have opposite polarities. When all these conditions are met, we count the speaker's polarized statement in the augmented CDX index. Here we should note that many such utterances would also be marked as expressive disagreement, in which case we do not count them twice. Nonetheless, this augmented index allows us to capture unmarked disagreements and thus obtain

Table 9A. *Correlation between selected topic control indices for a typical online chat discourse*

	TL	SMT	LTI	TOCX
TL	1.0			
SMT	0.96	1.0		
LTI	0.78	0.80	1.0	
TOCX (avg.)	0.92	0.95	0.88	1.0
$\alpha = 0.96$				

Table 9B. *Correlation between selected involvement indices for a typical online chat discourse*

	NPI	TI	TCI	ATP	INVX
NPI	1.0				
TI	0.76	1.0			
TCI	0.97	0.83	1.0		
ATP	0.87	0.90	0.95	1.0	
INVX	0.96	0.83	0.98	0.91	1.0
$\alpha = 0.98$					

a more complete metric. The resulting extended CDX measure is the basis of our disagreement SLB that we discuss and evaluate in the remainder of this paper.

3.4 Combining individual indicators into meta-indices

Many of the sociolinguistic behaviors defined thus far are characterized using multiple metrics, each of which may be viewed as an independent indicator that this language use is being deployed by a speaker and to what degree. While we expect the individual indicators to correlate when adequate data are present, we need to account for the error introduced by the computational process: depending upon the type of discourse, some component indices may be less reliably computed than others. As a result, the automated SLB predictions from multiple indicators would not always be consistent; however, we still need a way to form a single best output, even though we may have less than complete confidence in it. While computational reliability is still being improved, our tests on human-annotated data indicate that there is a strong correlation among the currently defined indices for each sociolinguistic behavior, as we have illustrated using the YMCA-1 dialogue. Further experiments confirmed this strong correlation across all datasets for which manual annotation was completed (Figures 5 and 6). Tables 9A and 9B show correlation between selected topic control and involvement indices, respectively: TL, SMT, and LTI for topic control; and NPI, TI, TCI, and ATP for involvement. TOCX and INVX are the combined measures obtained by averaging component percentage scores for each participant. These correlations also hold for face-to-face discussions, as shown in Figure 6 (for topic control). For all other sets that we examined, the correlation remains strong and

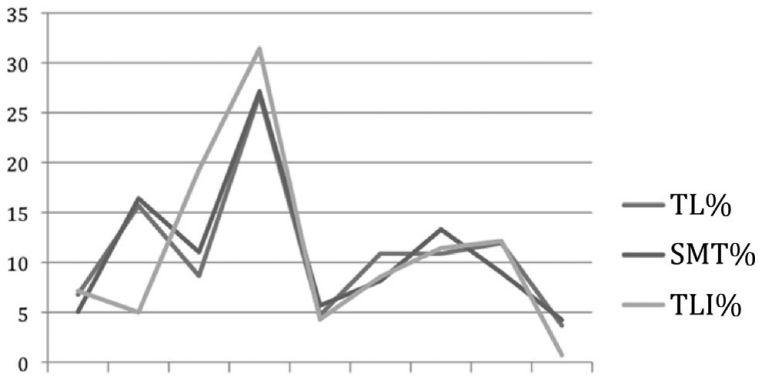

Fig. 5. Index correlation for selected measures of topic control recorded over a nine-person online chat.

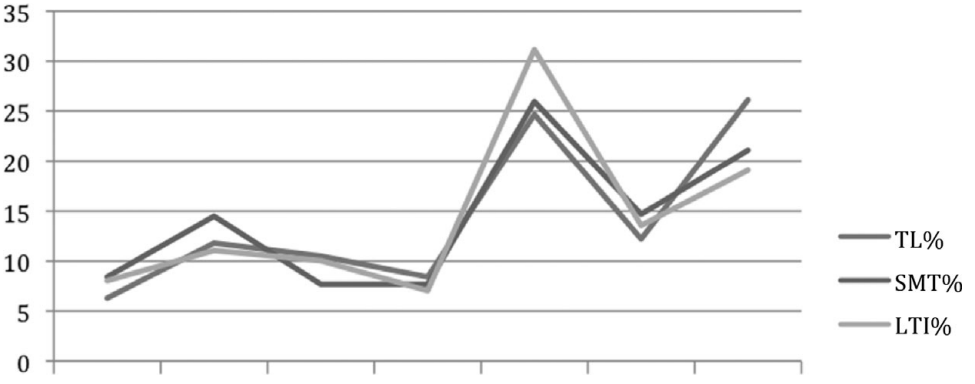

Fig. 6. Correlation of selected measures of topic control for face-to-face discussion between seven participants.

varies between 0.87 and 0.99 using the Cronbach's alpha reliability statistics (generally used to measure reliability of survey items to be combined into a single scale).

For the sociolinguistic behaviors where the component measures appear well correlated, we may combine them into one hybrid index, using an average (weighted or unweighted) of individual indicators. In practice, we may choose only a subset of available indices to include in the combined index, for instance, only those that correlates well with one another, while discarding any outliers. This helps to partially offset inadequacies of automated sub-processes responsible for each index. Table 10 shows the distribution of the combined index TOCX values computed for the YMCA-1 dialogue using all five topic control component indices.

4 Data and annotation

Our initial focus has been online chat dialogues. While chat data are plentiful online, its adaptation for research purposes presents a number of challenges that include users' privacy issues and anonymity. Furthermore, most data that may be obtained from public chat rooms are of limited value for the type of modeling tasks that we

Table 10. *The combined topic control index TOCX computed for the YMCA-1 dialogue*

	TOCX	
Speakers	(%)	(Deg.)
JR	20	4
LE	32	5
KN	21	4
KI	8	2
CS	12	3
KA	5	2
JY	2	1
Distribution statistics		
Mean	14%	
80 percentile	22%	
Std. Dev.	11%	

are interested in due to its high level of noise, lack of focus, and rapidly shifting, chaotic nature, which makes longitudinal studies challenging. To derive complex models of conversational behavior, we need the interaction to be reasonably focused on a task or some social objectives within a fairly stable group.

Few data collections exist covering multi-party dialogue meeting these characteristics, and even fewer with online chat. Moreover, the few collections that exist were primarily built for the purpose of training speech recognition systems and later adapted to modeling classifiers for dialogue acts and similar linguistic phenomena; few if any of these corpora are suitable for deriving pragmatic models of conversation, including sociolinguistic phenomena. Existing resources include a multi-person meeting corpus ICSI-MRDA (Shriberg *et al.* 2004), and the AMI Meeting Corpus (Carletta 2007), the latter containing 100 hours of meetings captured using synchronized recording devices. Still, these resources look at spoken language rather than online chat, which is the focus of the current study. There is a parallel interest in the online chat environment, although the development of useful resources has progressed less rapidly. Some corpora, such as the NPS Internet chat corpus (Forsyth and Martell 2007) exist, which have been anonymized manually and labeled with part-of-speech tags and dialogue act labels. The StrikeCom corpus (Twitchell *et al.* 2004) consists of thirty-two multi-person chat dialogues between players of a strategic game, where in 50% of the dialogues, one participant has been asked to behave 'deceptively'.

A weakness of these existing resources is the limited amount of information available about the dialogue participants, their attitudes, or prior relationships. Such information may be captured through questionnaires or interviews following each data collection experiment, but the questions must be designed to reflect the aims of the study. The resulting participant data, which in our case must include participants' assessment of their behavior and roles in conversation, are critical for model validation. Therefore, given the scarcity of the available data resources, a

new data collection process was required for our study. This is still a fairly typical situation, particularly in the study of the Internet chat, that new corpora are created on an as-needed basis (e.g., Khan *et al.* 2002; Wu *et al.* 2002; Kim *et al.* 2007).

Driven by the need to obtain a suitable dataset, we planned a series of experiments in which recruited subjects were invited to participate in a series of online chat sessions in a specially designed secure chat room. The experiments were carefully designed around topics, tasks, and games to engage the participants so that appropriate types of behavior, for example, disagreement, power play, persuasion, etc., could emerge spontaneously. These experiments and the resulting corpus have been described elsewhere (Shaikh *et al.* 2010b). Ultimately a corpus of 50 hours of English chat dialogue was collected comprising more than 20,000 turns and 120,000 words. In addition, we also assembled a corpus of 20 hours of Urdu chat and 22 hours of Mandarin chat that are used for modeling sociocultural phenomena in these languages. We call this corpus the Multi-Party Chat Corpus (MPC).

A sizeable subset of the English language dataset and a smaller subset of the Urdu collection have been annotated at four levels: communication links, dialogue acts, local topics, and meso-topics (which are essentially the most persistent local topics). Although full details of these annotations are impossible to explain within the scope of this paper, we briefly describe them below.[10] Annotated datasets were used to develop and train automatic modules that detect and classify social uses of language in discourse. It is important to note that the annotation has been developed to support the objectives of our project and does not necessarily conform to other similar annotation systems used in the past.

- *Communicative links*: In a multi-party dialogue an utterance may be directed toward a specific participant, a subgroup of participants or to everyone.
- *Dialogue acts*: We developed a hierarchy of fifteen dialogue acts for annotating the functional aspects of the utterance in discussion. The tag set adopted is based on DAMSL (Allen and Core 1997) and SWBD–DAMSL (Jurafsky *et al.* 1997), but compressed to fifteen tags tuned toward dialogue pragmatics and away from more surface characteristics of utterances (Shaikh *et al.* 2010a).
- *Local topics*: Local topics are defined as nouns or noun phrases introduced into discourse that are subsequently mentioned again via repetition, synonym, or pronoun.
- *Meso-topics*: Some local topics, which we call meso-topics, persist through a number of turns in conversation. A selection of meso-topics is closely associated with the task in which the discourse participants are engaged. In this definition, meso-topics are simply the most persistent local topics and constitute only a relatively small subset of all topics in a typical discourse. The actual length threshold of a meso-topic is a parameter in our system, and 'gaps' are allowed as long as these are brief, no more turns than the number

[10] Interested readers may request a copy of the SARMT annotation manual by contacting the first author by email. It is currently available as a technical report (Shaikh *et al.* 2010a), while a separate publication is under preparation.

Table 11. *Inter-annotator agreement on four annotation levels. N/A indicates that no additional training was performed*

Category	Initial training α	Additional training α
Communicative links	0.70	0.81
Dialogue acts	0.65	0.80
Local topics	0.82	N/A
Topic reference polarity	0.79	N/A

of speakers. In the current implementation, we treat any local topic of length 5 or more as a potential meso-topic.

- *Topic reference polarity*: Meso-topics can also be distinguished from 'ordinary' local topics by noting that the speakers often make polarized statements about them. This is an important property of meso-topics that allows us to measure disagreement in discourse. Meso-topic persistence is required to provide an opportunity for multiple polarized statements to occur.

We used trained annotators to mark up a subset of the collected data. Part of the annotated data was used to derive computational models for language use related to disagreement, involvement, and agenda control that we have described above. We also reserved a portion to serve as a 'ground truth' for the evaluation. In this section we briefly outline only the basic component-level annotation that consists of four interleaved layers: communicative links, dialogue acts, local topic tracking, and meso-topic valences. A more detailed description of the annotation scheme can be found in Shaikh *et al.* (2010a).

About 20 hours of data have been annotated at various levels. Inter-annotator agreement calculated using Krippendorff's alpha (Krippendorff 2004) on each of the annotation schemes is given in Table 11. After an initial round of annotation, some of our annotators were given more extensive training in tagging communicative links and dialogue acts, as well as related tasks, such as utterance splitting, amounting to approximately 60 hours of total training time as compared to approximately 18 hours for the initial training. Subsequently, the inter-annotator agreement on all categories went up to around 0.80 (as noted in the second column of Table 11). We need to note here that the annotation exercises served two purposes in our project. The first purpose was to determine the 'robustness' of the target SLB concept by gauging its sensitivity to the accuracy of the underlying component annotation. For this task, a strict inter-annotator agreement was not required; instead, we were interested whether in each case a valid SLB index could be obtained, that is, whether it would rank the participants in the same way as the index based on high-accuracy annotation. The second purpose was to create training data for improving performance of the linguistic components; this part of the project is currently ongoing and is not reported here.

All annotations were done using Social Actions/Roles Markup Tool (SARMT), a specially designed annotation tool in which each utterance is displayed on a separate line, with its speaker and timestamp information. Annotators go through

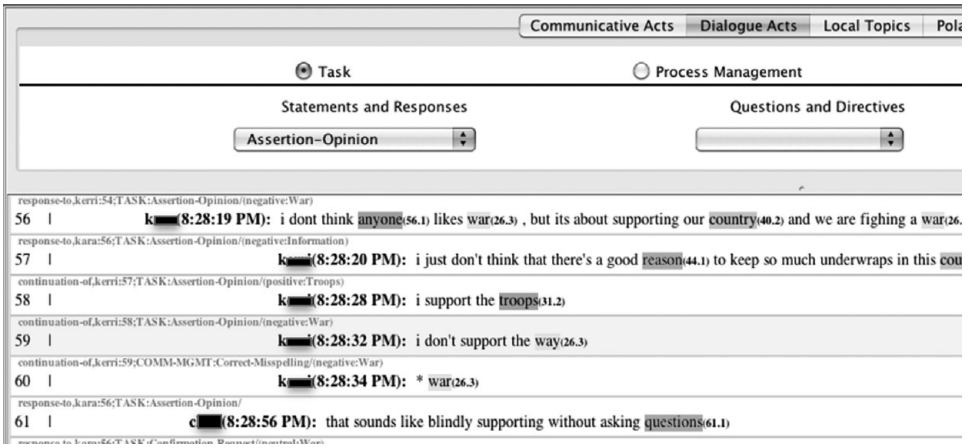

Fig. 7. A portion of the SARMT annotation tool interface (speakers' names obscured except for the first letter).

the dialogue and annotate each line for communicative acts, dialogue acts, local topics, and polarity assignment. Each of these categories is on a separate tab in the annotation tool, allowing annotators to complete one category before moving to the next one. In Figure 7, the dialogue acts tab is highlighted, showing the options of dialogue acts that an annotator can assign. A fully annotated dataset, further validated by measuring inter-annotator agreement, contains explicit markup of all linguistic elements needed to compute required SLB indices from which 'ground truth' SLB assignments could be obtained.

5 Implementing DSARMD prototype

We developed a prototype-automated DSARMD system that comprises a series of modules that create automated annotation of the source dialogue for all the language elements required to support computation of the SLB indices defined in previous sections. These include extraction of mentions of local topics, topic co-references, communicative links between utterances, classification of dialogue acts, identification of meso-topics, and assignment of polarity markers. Automatically annotated dialogue is then used to compute index values from which the SLB degree claims are derived.

Automated discourse processing involves the following modules:

- *Local topics detection* identifies first topic mentions by tracking occurrences of noun phrases. Subsequent mentions are identified using a fairly simple pronoun resolution method based primarily on the presence of specific lexical features as well as temporal distance between utterances. Princeton's Wordnet (Miller *et al.* 1990) is consulted to identify synonyms and other related words commonly used in co-references.
- *Meso-topics* are currently identified as the longest-chain local topics. In addition, we require that there be some differently polarized utterances within these

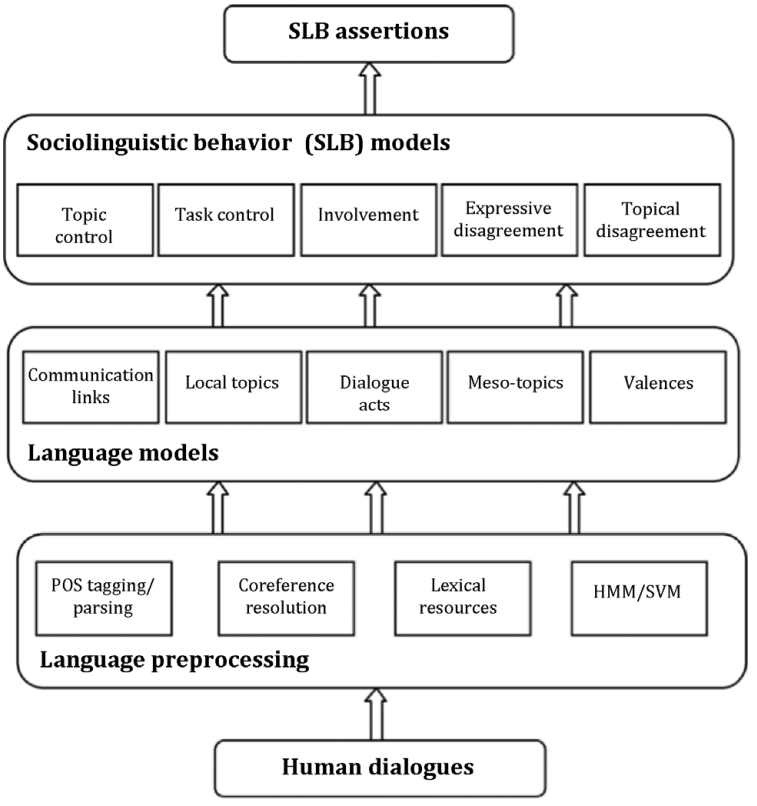

Fig. 8. Schematic architecture of the DSARMD-1 prototype showing main components and processes.

chains. Their *polarity* is assessed by noting the presence of positive or negative cue words and phrases. We utilize several available sources of such cue words, including the subjectivity lexicon described by Wilson, Wiebe and Hoffmann (2005). Machine learning techniques are used to tune up these generic lists to conversational vocabulary of the MPC corpus.

- *Dialogue acts* are tagged based on the presence of certain cue phrases derived from an external training corpus (Webb and Ferguson 2010), tuned against the annotated portion of the MPC corpus. Our dialogue act system is generally compatible with earlier systems as noted previously; however, it uses a greatly reduced tag set (to fifteen tags) and is adapted to better capture significant social nuances within conversation, while eschewing mainly semantic distinctions, for example, *wh-questions* versus *yes–no questions* etc.
- *Communicative links* indicate who is the primary addressee of an utterance and are determined by computing inter-utterance similarity scores based on lexical overlap, n-gram overlap, presence of overt co-references, and other cues such as direct addressing as well as timing.

Figure 8 shows the overall schematic architecture of the DSARMD-1 prototype. The system comprises several basic language-processing components and tools,

including part-of-speech tagging, syntactic parsing, and co-reference resolution, as well as various lexical resources, and basic machine learning tools. These are used to derive language models for computing intermediate-level components, such as local topic chains, dialogue act assignments, and meso-topic valences. Currently these intermediate components perform at 60%–80% accuracy and are subject to further improvement. SLB models are built for individual indices from the intermediate components as discussed earlier. SLB assertions (in XML format) are output for each discourse participant.

6 Evaluation

Up to this point we have discussed the principles and design underlying our approach to computing social language uses in discourse. We showed how the prototype-automated system DSARMD-1 could assign a degree of language use to each discourse participant given a dialogue as input. Now we take a look at how we can validate accuracy of the system output. This is done at two levels: computational process accuracy and theoretical adequacy. First, the system performance is measured against the assessment obtained from manually annotated discourse data; then it is measured against participants' own assessment of sociolinguistic behaviors displayed by the speakers during conversation. In both cases, the system output is based on combined measures for each sociolinguistic behavior involved.

6.1 Evaluating combined SLB predictions against human annotation

In order to evaluate accuracy of the automated process, we compared the SLB claims generated automatically by the system to the claims obtained from human-annotated data. For the purpose of this evaluation, we assume that human annotation, which was performed at the level of linguistic components (as explained in Section 4), embodies a valid operational procedure for detecting the presence of corresponding sociolinguistic behaviors. This assumption is justified by the supporting social science research detailed in references cited in Section 2. Therefore, the objective of this first evaluation was to assess how well our automated algorithms can approximate this process, given the limited accuracy of the system's linguistic components as well as the use of more easily computable proxies for extracting topic chains or assigning dialogue acts. In the next subsection we discuss evaluation of system output against direct human assessment of sociolinguistic behavior in conversation.

We start by briefly outlining the evaluation methodology and metrics used. All sociolinguistic phenomena discussed in this paper lend themselves well to scalar representation, that is, discourse participants, or in some cases pairs of participants, may be ranked by the degree to which they exhibit a specific sociolinguistic behavior, relative to other participants. For example, participants in the YMCA-1 dialogue are ranked by their degree of topic control. This suggests a simple evaluation metric whereby the system generated ranking is compared to some 'gold standard', either produced directly by human judgment, or else derived from human-annotated data. In order to evaluate how well the system performs, we compare all pairs of ranked

Table 12. *Evaluation of the SLB ranking assignments among participants in five dialogues measured against rankings obtained from human-annotated data*

	F19B	M7B	M11B	M14A	M14B	Avg.
Topic control	0.87	0.87	0.67	0.83	0.70	0.79
Task control	0.76	0.66	0.85	0.20	0.73	0.64
Involvement	0.93	0.93	0.81	0.83	0.90	0.88
Disagreement	0.67	0.80	0.90	0.33	0.67	0.67

elements A and B in the machine output and in the gold standard. Thus, if A is ranked ahead of B in the gold standard (A>B), then we also want A>B in the machine output (a match); if this is not the case, we have a fault. In general, for the set of n elements in the system output, we have ${}^1/_2 n(n-1)$ ordered pairs; if m of these pairs are matches, then the system accuracy is $2m/n(n-1)$.

Let us consider a specific example, where four elements are ranked in the gold standard as A>B>C>D while the system output is B>A>C>D. The system output yields the set of ordered pairs: {B>A, A>C, A>D, B>C, B>D, C>D}, which all but one match the gold standard. The resulting accuracy is 5/6 or 0.83. On the other hand, the system output of D>B>C>A would only attain 0.17 accuracy.

Here it may be argued that not all of the ordering pairs are equally significant; specifically, the location of the top-ranked element in the gold standard may be a valid metric of its own. While this may be useful when leadership and other higher level social constructs are assessed (e.g., 'who is the leader?', 'Is Joe the leader?'), at this stage we are interested in an overall accuracy of the automated system. We shall return to this point in the next section when we discuss the preliminary investigation into detecting discourse leaders.

Table 12 summarizes the performance of the automated DSARMD system against a set of five datasets in the set-aside pool of fifteen dialogues. These five sets were manually annotated as described in Section 4. The participants' rankings with respect to each of the four sociolinguistic behaviors (topic control, task control, involvement, and disagreement[11]) were subsequently derived from the annotated dialogues and compared with the rankings obtained from the fully automated system.

What Table 12 shows is the degradation of performance due to errors introduced by the automated process, including local topic detection and tracking, communicative link assignment, dialogue act classification, and so forth. In general, the performance of our automated system on topic control and involvement, which are the two sociolinguistic behaviors that are currently detected using combined multi-index measures, is quite satisfactory across all sets. Computing the remaining language uses, task control and disagreement, is less reliable, and may vary from one dialogue to next (e.g., we note the sharp drop in accuracy computed for M14A dataset). This vulnerability in performance is largely linked to the accuracy of a

[11] Disagreement SLB is based on the CDX that calculates the degree of disagreement generated by each participant toward others. We are not at this time measuring pair-wise disagreement.

single underlying language-processing module, namely the dialogue act classifier on which both measures depend at this point. The reader may recall that both task control and disagreement behaviors are currently assessed based upon participants' use of specific types of dialogue acts, which include directives and disagreements. As a result, the frequency with which these dialogue acts are deployed in a dialogue will affect our system performance. For example, the use of overt directives (action-directive DA) may be quite limited in some round-table style meetings as compared with, for instance, action-oriented quests where other activities (besides discussion) are involved. At very low counts, perhaps just a few directives scattered among the speakers in a several hundred-turn-long discourse, the corresponding indices become unreliable in ranking the relative strength of the corresponding behavior for any participant. This weakness may call for the development of additional measures, especially for task control, which we believe is a key indicator of leadership. We are currently working on implementing additional measures for task control as well as for disagreement that would complement the current measure in situations when it becomes less reliable. For example, including counts for certain forms of questions, specifically what we tag as *confirmation-request*, which are more common in some types of discussion may compensate for relative lack of directives in such contexts.

6.2 Evaluating automated SLB predictions against human assessment

In the preceding section we evaluated the performance of the DSARMD system against the output of trained human annotators who used our specially designed annotation software (Section 4) and followed strict guidelines to accurately mark up the linguistic features that indicate presence of the target sociolinguistic behaviors. The purpose of this evaluation was to assess how well the system could replicate an operational procedure designed for assessing these behaviors, where the procedure itself is based on an established social theory (see Section 2).

Nonetheless, it is even more interesting to see how well such predictions match human intuition, in other words, whether our process captures valid sociolinguistic behaviors. To perform this type of evaluation, we converted the results of post-session surveys into rankings of discourse participants in several categories, corresponding to three of the sociolinguistic behaviors under investigation. The only exception is disagreement, which did not have a directly corresponding survey question: Question 4 in Table 13 asks for assessment of the participant's own disagreeability rather than for ranking others. Nonetheless, in a separate study, we found that these two measures of disagreement are highly correlated.

6.2.1 Post-session questionnaires

In order to evaluate the accuracy of the automated process, we compared the system-generated SLB claims to assessments provided by discourse participants in response to questions regarding language uses. Following each session, participants were instructed to answer a survey aimed at eliciting responses regarding the interaction they had freshly completed. Survey questions were carefully designed by social science

Table 13. *Selected post-session participant survey questions used to assess the degree of sociolinguistic behavior for each participant*

	Survey question	Scale endpoints	Sociolinguistic behavior/role
1.	During the discussion, some of the people talking are more influential or persuasive than others. For the conversation you just took part in, please rate each of the participants in terms of how influential they seemed to you?	Very influential Not influential	Topic control
2.	During the discussion, some of the people have greater effect on the group's decision than others. For the conversation you just took part in, please rate each of the participants in terms of how much they affected the group's decision?	Very effective Not effective	Task control
3.	During the discussion, some of the people talking are very involved in the conversation and talk a lot, while others are fairly quiet. For the conversation you just took part in, please rate each of the participants in terms of how involved they seemed to you?	Very involved Not involved	Involvement
4.	During the discussion, how frequently did you see positions or opinions that differed from your own?	Very frequently Never	Disagreement
5.	Below is a list of participants, including you. Please rank order the participants with 1 being the leader, 2 being a leader but not so much as 1, and so on.		Leadership

standards, requesting participants' reactions without being overtly suggestive. Some questions from the survey are included in Table 13. Participants rated each other, as well as themselves, for these questions on an unnumbered 10-point scale. The ends of the scale were titled for orienting the responses. The responses were then converted into numerical values (1–10) and averaged for each participant over all survey questions to calculate their individual scores.

Using the scores given by participants for each survey question, we can rank order each participant on each sociolinguistic behavior. For example, by taking an average of the scores obtained by each participant in responses to question 1, we can determine their relative ranking with respect to the degree of topic control (or influence). Similarly, for questions 2 and 3, which provide participant assessments of topic control (effectiveness) and involvement. In each case, we obtain a relative ranking of all participants. Question 4, as formulated here, is not directly applicable to evaluating disagreement the way it is currently modeled as a degree of disagreeability by each participant; however, because disagreement is currently based on a single measure (CDX), it can be straightforwardly evaluated against the annotated dialogue corpus, specifically by counting the number of disagree–reject dialogue acts and opposing polarity statements attributed to each speaker. Question

Table 13A. *Correlation between scores for survey questions for selected SLBs and leader role in Table 13 for a chat discourse in our corpus*

	Influential	Effective	Involved	Leader
Influential	1.0			
Effective	0.96	1.0		
Involved	0.95	0.94	1.0	
Leader	0.90	0.88	0.76	1.0
$\alpha = 0.93$				

5 will be used in a future stage of our research when we compute leadership scores for group participants.[12]

We performed a correlation analysis of the sociolinguistic behaviors captured in the above questionnaire and the leadership prediction in order to verify the projections alluded to in Tables 7 and 8. A strong correlation (avg. $\alpha = 0.89$) was found between participants' assessments of speaker's influence (topic control), effectiveness (task control), and involvement, and their assessments of the leader role, as shown in Table 13A, which represents a typical dialogue in our corpus.[13] Additional details on leadership evaluation can be found in Section 7, when we also discuss our future work.

6.2.2 Evaluation results

Evaluation was performed over the set of fifteen English chat dialogue sessions that were set aside for this purpose. The test data encompasses approximately 1,350 minutes of dialogue, fifty participants, and more than 8,600 turns. Table 14 gives more details about each of the fifteen test dialogues relevant to this paper.

Participants' answers to post-session questionnaires were converted into rankings based on average survey scores assigned to each speaker. Table 15 is a cross-evaluation matrix for topic control (Influence) assessment based on responses to question 1 of the survey (shown in Table 13).

Table 16 compares the rankings of participants' topic control behavior in the same test dialogue obtained from the post-session survey above to the rankings produced by the automated DSARMD system. The post-session survey scores are the average scores for each participant out of 10 based on our survey scale, while the system-assigned scores are percentages out of 100 (rounded up to 2 decimal places). The comparison of the two rankings using the pair-wise order metric (explained in Section 6.1) shows the system performance accuracy at 87%. This is equivalent to system misordering two pairs of participants out of fifteen possible pair-wise rankings.

[12] It should be noted here that the five-question survey was originally developed for assessment of various aspects of leadership.

[13] Question 4 (disagreement) is not included because it is not formulated to directly correlate with the way disagreement SLB is operationalized (see Section 3.3).

Table 14. *Selected characteristics of the test collection used in evaluation*

Dialogue Id.	M10A	M10B	F19A	M14B	F26A	F27B	M11B	F19B	M11A	F28B	F28A	M7B	M6A	M14A	M7A
Turns	698	831	882	546	1,107	556	479	590	388	467	521	434	623	283	282
Speakers	6	6	6	4	7	5	5	6	6	5	6	6	8	4	6
Duration (minutes)	95	87	86	90	91	91	86	96	92	87	82	84	90	93	70

Table 15. *Topic control cross-assessment matrix obtained from post-session survey (question 1) following a single six-speaker session*

Topic control assessment	Participants					
Judged by	LA	LI	SA	JF	JE	JN
JF	6.5	6.5	6	6	6	5.5
JN	6	5.5	6.5	6.5	5	5
JE	8.5	8.5	7.5	6.5	8.5	1.5
LA	6	7	7	7	6	6
LI	6.5	6	6	6	6	4.5
SA	8	8	5	5.5	5.5	5.5
Average	6.92	6.92	6.33	6.25	6.17	4.67

Table 16. *An example of measuring system accuracy in ranking topic control behavior in a single 90-minute discourse among six participants*

	Topic control	
Participants	Post-session survey scores and ranks (Q1)	System-assigned scores and ranks
LI	6.92 (1)	0.31 (1)
LA	6.92 (1)	0.27 (2)
SA	6.33 (3)	0.15 (3)
JF	6.25 (4)	0.12 (4)
JE	6.17 (5)	0.12 (4)
JN	4.67 (6)	0.04 (6)
System accuracy	0.87	

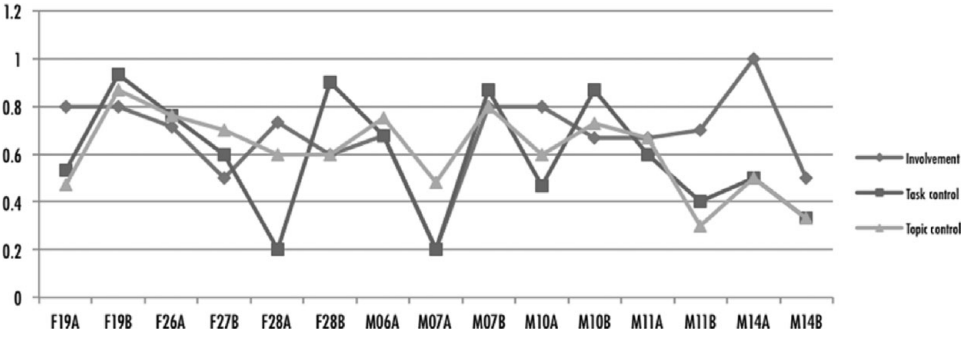

Fig. 9. DSARMD system performance over the fifteen test dialogues.

Table 17 and the accompanying chart in Figure 9 show the system performance across all fifteen test dialogues, as well as the overall averages, for detecting three sociolinguistic behaviors (disagreement is not included, since its performance is currently measured as a rate of detection of disagree–reject dialogue acts only). We note that the performance is good overall, with a few outliers.

Table 17. *Results of evaluation of the SLB assignment rankings showing system accuracy over fifteen dialogue sets against human questionnaires*

Dialogue Id.	F19A	F19B	F26A	F27B	F28A	F28B	M6A	M7A	M7B	M10A	M10B	M11A	M11B	M14A	M14B
INV	0.8	0.8	0.71	0.5	0.73	0.6	0.68	0.2	0.8	0.8	0.67	0.67	0.7	1	0.5
Task	0.53	0.93	0.76	0.6	0.2	0.9	0.68	0.2	0.87	0.47	0.87	0.6	0.4	0.5	0.33
Topic	0.47	0.87	0.76	0.7	0.6	0.6	0.75	0.48	0.8	0.6	0.73	0.67	0.3	0.5	0.33

The chart in Figure 9 presents the same evaluation results graphically, which makes it easier to see the stability of each measure. The performance accuracy generally stays within 60%–80% range, with a few outliers on either side. The averages are 0.68 for involvement, 0.59 for task control, and 0.61 for topic control. We note that task control measures are somewhat less stable than the measures underlying the other two sociolinguistic behaviors, primarily due to relative sparseness of the PMI index, which is based on the count of directives used in conversation. If the two outliers, F28A and M7A, are disregarded, the average performance of task control on the remaining thirteen sets is 0.65.

Error analysis on these and other outlier sets indeed reveals that much of the variation of performance is related to the recall of the underlying linguistic features upon which our measures of sociolinguistic behavior are based. Table 18 shows task control performance sorted by dialogues with varying frequency of directives, starting from the highest frequency to the lowest one. It is quite clear from this chart that the system performance becomes more unpredictable as the frequency of directives decreases. This is because the accuracy of the underlying process, namely, dialogue act classification, impacts the output of the task control measures (in particular the PMI index) more dramatically as the underlying statistics become less reliable. As already mentioned, we are currently exploring additional measures of task control to ease this problem.

7 Conclusion and future work

In this paper we presented a preliminary design for modeling certain types of social phenomena in multi-party online dialogues. Our approach starts with the hypothesis that high-level social constructs (roles and states), such as leadership or group cohesion, can be construed from the degree of participants' engagement in selected sociolinguistic behaviors. These sociolinguistic behaviors can in turn be detected and graded by tracking observable, and thus potentially computable, linguistic features in speakers' utterances. In order to verify this hypothesis, we developed an automated system DSARMD-1 that can compute degrees of several sociolinguistic behaviors in online dialogues, namely, topic control, task control, involvement, and disagreement. Initial evaluation shows promising results in automated detection of these sociolinguistic behaviors, especially where multiple indicator measures are used.

Much work lies ahead, including larger scale evaluation, testing index stability, and resilience to errors in automated language processing, including topic detection, coreference resolution, polarity estimation, and dialogue act classification. Current performance of the system is based on only preliminary versions of these linguistic modules, which perform at only 70%–80% accuracy, so these need to be improved as well. Research on Urdu and Chinese dialogues, to see how these concepts apply to radically different languages and cultures, is just starting.

The next major step in this research is to build computational models for leadership, group cohesion, and other high-level social constructs. The SLB models described in this paper can be used to develop an effective, albeit quite preliminary,

Table 18. *Task control performance stability versus density of linguistic indicators (directives)*

	M10A	M10B	F26A	F19B	F19B	F27B	M11B	M14B	M6A	M11A	F28A	M7B	F28B	M7A	M14A
Task C. accuracy	0.47	0.87	0.76	0.53	0.93	0.6	0.4	0.33	0.68	0.6	0.2	0.87	0.9	0.2	0.5
Directives found	75	74	69	62	50	43	43	40	38	35	32	30	27	26	19
Directives per speaker	12.5	12.3	9.9	10.3	8.3	8.6	8.6	10.0	4.75	5.8	5.3	5.0	5.4	4.6	4.75

method for detecting leaders in group meetings. Specifically, the combined topic control and task control measures are expected to correlate well with the leadership role (Table 13A), while involvement and disagreement may play supporting roles in differentiating types of leaders (involved, assertive, etc.). This assessment is confirmed by a series of preliminary experiments, in which we combined participants' scores for all four sociolinguistic behaviors described here using a weighted average that gave significantly more weight to task control (60%) and topic control (70%), while treating involvement and disagreement as less central behaviors (with 5% and 10% weights, respectively). The results were then compared against the rankings of participants based on question 5 of the survey (c.f. Section 6.2, Table 13) to see if the system was able to correctly pick the leader. It turns out that it can already do so with 87% accuracy; this number increases to over 90% if we allow the choice of the first- or second-ranked speaker in situations where their survey scores are very close. These results are promising but still very preliminary. As our work continues, full results on the classification of leadership and other high-level sociolinguistic phenomena will be reported in a future paper.

Acknowledgments

This research was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the US Army Research Lab. All statements of fact, opinion, or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI, or the US Government.

References

Agar, M. 1994. *Language Shock, Understanding the Culture of Conversation*. New York: Quill, William Morrow.

Allen, J., and Core, M. 1997. Draft of DAMSL: dialog act markup in several layers. www.cs.rochester.edu/research/cisd/resources/damsl/ Accessed 2009.

Austin, J. L. 1962. *How to Do Things with Words*. Oxford, UK: Clarendon Press.

Avolio, B., J., Bass, B. M., and Jung, D. I. 1999. Re-examining the components of transformational and transactional leadership using the multifactor leadership questionnaire. *Journal of Occupational & Organizational Psychology* **72**: 441–62.

Bales, R. F. 2001. *Social Interaction Systems, Theory and Measurement*. Piscataway, NJ: Transaction Publishers.

Barnes, M. S. 2005. Exploring how power is enacted in small groups. In H. L. Chick and J. L. Vincent (eds.), *Proceedings of the 29th Conference of the International Group for the Psychology of Mathematics Education*, vol. 2, pp. 137–44. Melbourne, Australia: PME.

Blaylock, N. 2002. Managing communicative intentions in dialogue using a collaborative problem-solving model. Technical Report 774, CS Department, University of Rochester, Rochester, New York.

Bonner, H. 1959. *Group Dynamics; Principles and Applications*. New York: Ronald Press.

Broadwell, G. A. *et al.* 2010. Social phenomena and language use. ILS Technical Report, SUNY, New York.

Bunt, H. 1994. Context and dialogue control. *THINK* **3**: 19–31.

Carberry, S., and Lambert, L. 1999. A process model for recognizing communicative acts and modeling negotiation dialogue. *Computational Linguistics* **25**(1): 1–53.

Carletta, J. 2007. Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation Journal* **41**(2): 181–90.

Carlson, L. 1983. *Dialogue Games: An Approach to Discourse Analysis*. Dordrecht, Netherlands: D. Reidel.

Chhokar, J. S., Brodbeck, F. C., and House, R. J. 2007. *Culture and Leadership, Across the World: The GLOBE Book of In-Depth Studies of 25 Societies*. Florence, KY: Psychology Press.

Chu-Carroll, J., and Brown, M. K. 1998 (September). An evidential model for tracking initiative in collaborative dialogue interactions. *User Modeling and User-Adapted Interaction* **8**(3–4), 215–53.

Core, M. G., Moore, J. D., and Zinn, C. 2003. The role of initiative in tutorial dialogue. In *EACL'03: Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, pp. 67–74. Association for Computational Linguistics.

DiMicco, J. M., Pandolfo, A., and Bender, W. 2004. Influencing group participation with a shared display. In *Proceedings of the ACM Conference CSCW'04*, Chicago, IL, USA, pp. 614–23.

Ellis, D. G., and Fisher, B. A. 1994. *Small Group Decision Making: Communication and the Group Process*. New York: McGraw-Hill.

Fairhurst, G. 2007. *Discursive Leadership: In Conversation with Leadership Psychology*. Los Angeles, CA: Sage.

Field, D., Worgan, S., Webb, N., Hepple, M., and Wilks, Y. 2008. Automatic induction of dialogue structure from the companions dialogue corpus. In *4th International Workshop on Human-Computer Conversation*, Bellagio, Italy.

Forsyth, E. N., and Martell, C. H.. 2007. Lexical and discourse analysis of online chat dialog. In *First IEEE International Conference on Semantic Computing (ICSC 2007)*, California, USA, pp. 19–26.

Givon, T. 1983. *Topic Continuity in Discourse: A Quantitative Cross-Language Study*. Amsterdam, Netherlands: John Benjamins.

Hogg, M. A., and Reid, S. A. 2006. Social identity, self-categorization, and the communication of group norms. *Communication Theory* **16**: 7–30.

Huffaker, D. 2010. Dimensions of leadership and social influence in online communities. *Human Communication Research* **36**: 596–617.

Ji, G., and Bilmes, J. 2006. Backoff model training using partially observed data: application to dialog act tagging. In *Proceedings of the Human Language Technology/American Chapter of the Association for Computational Linguistics (HLT/NAACL'06)*, New York, USA.

Jurafsky, D., Bates, R., Coccaro, N., Martin, R., Meteer, M., Ries, K., Shriberg, E., Stolcke, A., Taylor, P., and Van Ess-Dykema, C. 1997. Automatic detection of discourse structure for speech recognition and understanding. In *Proceedings of the 1997 IEEE Workshop on Speech Recognition and Understanding*, Santa Barbara, CA, USA.

Ketrow, S. M. 1991. Communication role specializations and perceptions of leadership. *Small Group Research* **22**: 492–514.

Khan, F. M., Fisher, T. A., Shuler, L., Wu, T., and Pottenger, W. M. 2002. Mining chat-room conversations for social and semantic interactions. Technical Report, LU-CSE-02-011, Computer Science and Engineering, Lehigh University, Bethlehem, PA.

Kim, J., Shaw, E., Chern, G., and Feng, D. 2007. An intelligent discussion-bot for guiding student interactions in threaded discussions. In *AAAI Spring Symposium on Interaction Challenges for Intelligent Assistants*, Stanford University, CA, USA.

Krippendorff, K. 2004. *Content Analysis: An Introduction to its Methodology*. Los Angeles: Sage.

Levin, L., Thyḿe-Gobbel, A., Lavie, A., Ries, K., and Zechner, K. 1998. A discourse coding scheme for conversational Spanish. In *Proceedings of the International Conference on Speech*

and Language Processing (ICSLP '98), The Australian Speech Science and Technology Association, Sydney, Australia.

Linell, P. 1990. The power of dialogue dynamics. In I. Markov/a and K. Foppa (eds.), *The Dynamics of Dialogue*, pp. 147–77. Brighton, Sussex: Harvester.

Lowe, K. B., Kroeck, K. G., and Sivasubramaniam, N. 1996. Effectiveness correlates of transformational and transactional leadership: a meta-analytic review of the MLQ literature. *Leadership Quarterly* **7**(3): 385–425.

McCallum-Bayliss, H. 2010. SCIL overview. Socio-cultural content in language, In *PI Meeting*, IARPA 1, College Park, MD, USA.

Miller, G. A., Beckwith, R., Fellbaum, C. D., Gross, D., and Miller, K. 1990. WordNet: an online lexical database. *International Journal of Lexicograph.* **3**(4): 235–44.

Misiolek, N. I., and Heckman, R. 2005. Patterns of emergent leadership in virtual teams. Paper presented at the *38th Hawaii International Conference on System Sciences*, Honolulu, HI, USA.

Morris, C. G. H., and Richard, J. 1969. Behavioral correlates of perceived leadership. *Journal of Personality and Social Psychology* **13**: 350–61.

Pavitt, C., High, A. C., Tressler, K. E., and Winslow, J. K. 2007. Leadership communication during group resource dilemmas. *Small Group Research* **38**(4): 509–31.

Phillips, G. M. 1973. *Communication and the Small Group.* Indianapolis, ID: Bobbs-Merrill.

Poesio, M., and Mikheev, A. 1998. The predictive power of game structure in dialogue act recognition. In *International Conference on Speech and Language Processing (ICSLP-98)*, Sydney, Australia.

Pomerantz, A., and Denvir, P. 2007. Enacting the institutional role of chairperson in upper management meetings: the interactional realization of provisional authority. In F. Cooren (ed.), *Interacting and Organizing: Analyses of a Management Meeting*, pp. 31–51. Mahwah, NJ: Lawrence Erlbaum.

Reid, S. A., and Ng, S. H. 1999. Language, power, and intergroup relations. *Journal of Social Issues* **55**(1): 119–39.

Sacks, H., Schegloff, E., and Jefferson, G. 1974. A simplest systematic for the organization of turn-taking for conversation. *Language* **50**(4): 696–735.

Samuel, K., Carberry, S., and Vijay-Shanker, K. 1998. Dialogue act tagging with transformation-based learning. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Montreal, Germany.

Scollon, R., and Scollon, S. W. 2001. *Intercultural Communication, A Discourse Approach*, 2nd ed. San Francisco, CA: Blackwell.

Searle, J. R. 1969. *Speech Acts: An Essay in the Philosophy of Language.* Cambridge, UK: The University Printing House.

Shaikh, S., Strzalkowski, T., Broadwell, G. A., Stromer-Galley, J., Webb, N., Boz, U., and Elia, A. 2010a. DSARMD annotation guidelines version 2.5. Technical Report 014, ILS, SUNY, Albany, New York.

Shaikh, S., Strzalkowski, T., Taylor, S., and Webb, N. 2010b. MPC: a multi-party chat corpus for modeling social phenomena in discourse. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010)*, Valletta, Malta.

Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., and Carvey, H. 2004. The ICSI Meeting Recorder Dialog Act (MRDA) corpus. In M. Strube and C. Sidner (eds.), *Proceedings of 5th SIGdial Workshop on Discourse and Dialogue*, Cambridge, MA, April 30–May 1, pp. 97–100.

Stein, R. T., and Heller, T. 1983. The relationship of participation rates to leadership status: a meta-analysis. In H. H. Blumberg, A. P. Hare, V. Kent, and M. F. Davies (eds.), *Small Groups and Social Interaction*, vol. 1, pp. 401–6. New York: John Wiley.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C. and Meteer, M. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics* **26**(3): 339–73.

Stromer-Galley, J. 2007. Measuring deliberation's content: a coding scheme. *Journal of Public Deliberation* **3**(1). http://services.bepress.com/jpd/vol3/iss1/art12/

Twitchell, D. P., Nunamaker Jr., J. F., and Burgoon, J. K. 2004. Using speech act profiling for deception detection. In *Intelligence and Security Informatics*, LNCS, vol. 3073. Heidelberg, Germany: Springer.

Webb, N., and Ferguson, M. 2010. Automatic extraction of cue phrases for cross-corpus dialogue act classification. In *The Proceedings of the 23rd International Conference on Computational Linguistics (COLING-2010)*, Beijing, China.

Whittaker, S., and Stenton, P. 1988. Cues and control in expert-client dialogues. In *26th Annual ACL Conference*, Buffalo, NY, USA, pp. 123–30.

Wilson, T., Wiebe, J., and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT-EMNLP-2005*, Vancouver, BC, Canada, October 6–8.

Wu, T., Khan, F. M., Fisher, T. A., Shuler, L. A., and Pottenger, W. M. 2002. Posting act tagging using transformation-based learning. In *Foundations of Data Mining and Discovery, IEEE International Conference on Data Mining*, San Jose, CA, USA.

Yoo, Y., and Alavi, M. 2004. Emergent leadership in virtual teams: what do emergent leaders do? *Information and Organization* **14**: 27–58.

Zimmerman, D. H., and West, C. 1975. Sex roles, interruptions, and silences in conversation. In B. Thorne and N. Henley (eds.), *Language and Sex: Differences and Dominance*. Rowley, MA: Newbury House.