Modeling Leadership and Influence in Multi-party Online Discourse

Tomek STRZALKOWSKI[1,2], Samira SHAIKH[1], Ting LIU[1], George Aaron BROADWELL[1], Jenny STROMER-GALLEY[1], Sarah TAYLOR[3], Veena RAVISHANKAR[1], Umit BOZ[1] and Xiaoai REN[1]

(1) State University of New York – University at Albany, NY 12222 USA
(2) Polish Academy of Sciences
(3) Sarah M. Taylor Consulting, LLC
tomek@albany.edu, samirashaikh@gmail.com

ABSTRACT

In this article, we present a novel approach towards the detection and modeling of complex social phenomena in multi-party discourse, including leadership, influence, pursuit of power and group cohesion. We have developed a two-tier approach that relies on observable and computable linguistic features of conversational text to make predictions about sociolinguistic behaviors such as Topic Control and Disagreement, that speakers deploy in order to achieve and maintain certain positions and roles in a group. These sociolinguistic behaviors are then used to infer higher-level social phenomena such as Leadership and Influence, which is the focus of this paper. We show robust performance results by comparing our automatically computed results to participants' own perceptions and rankings. We use weights learnt from correlations with training examples known leadership and influence rankings of participants to optimize our models and to show performance significantly above baseline for two different languages – English and Mandarin Chinese.

KEYWORDS : computational sociolinguistics, online dialogues, social phenomena, linguistic behavior, influence, multi-disciplinary artificial intelligence, social computing

1 Introduction and Related Work

Our objective is to model high-level sociolinguistic phenomena such as Leadership, Influence, Pursuit of Power and Group Cohesion in discourse. This research project aims to develop a computational approach that uses linguistic features of conversational text to detect and model sociolinguistic behaviors of conversation participants in small group discussions. Given a representative dialogue of multi-party conversation, our prototype system automatically classifies the participants by the degree to which they engage in such sociolinguistic behaviors as Topic Control, Task Control, Disagreement, and several others discussed in this paper. These mid-level sociolinguistic behaviors are deployed by discourse participants in order to assert higher-level social roles such as Leadership. Our approach to this problem combines robust computational linguistics methods and established empirical social science techniques. The focus in this paper is on online multi-party conversations in chat rooms; however, the models we are developing are intended to be universal and are applicable to other conversational situations: informal face-to-face interactions, formal meetings, moderated discussions, asynchronous threaded discussions as well as interactions conducted in languages other than English, e.g., Urdu and Mandarin. We shall discuss the robust detection of Leadership and Influence in discourse in this paper; we defer the discussion of remaining phenomena to a separate, larger publication.

Social science theory indicates that leadership may be manifested in various ways (Bradford, 1978, Huffaker, 2010). We define leadership in the following terms: A leader is someone who guides group toward an outcome, controls group discussion, manages actions of the group and whom members recognize as the task leader. Such a leader is a skilled *task* leader, which corresponds to the social science theory put forth in Beebe and Masterson (2006). On the other hand, a *thought* leader in the group is someone who has credibility in the group and introduces ideas or thoughts that others pick up on or support. Such a person is a participant but need not be active in the portion of discussion where others credit or support him. This definition corresponds to the Initiator-Contributor type of leadership outlined in Bradford (1978). For ease of presentation and understanding – we shall refer to task or skilled leadership as *Leadership* and thought leadership as *Influence*, henceforth in this paper.

Since leadership and influence are manifested differently and may be deployed by distinct participants in a discussion, it is important for an automatic system to recognize the distinction and make a determination of who is deploying such roles. Consider as an example, a debate with panel of experts hosted by a facilitator. Here, the facilitator will exhibit sociolinguistic behavior consistent with being a task leader, by controlling the agenda, putting forth questions to individual panelists, beginning and ending the discussion and so on. However, she will not be a thought leader, or influencer, as she does not contribute much actual content to the discussion apart from asking questions. Any member of the expert panel may exhibit the sociolinguistic behavior consistent with being an influencer. In a peer-oriented group discussion however, it could occur that the task and thought leader (leader and influencer) are the same person.

Human-human interaction affords a rich resource for research. Much prior work has been done in communication that focuses on the communicative dimension of discourse. For example, the Speech Act theory (Austin, 1962; Searle 1969) provides a generalized framework of multiple levels of discourse analysis; work on dialogue analysis (Blaylock, 2002; Carberry and Lambert, 1999; Stolcke et al., 2000) focuses on information content and structure of dialogues. Somewhat more relevant to social roles is research that models sequences of dialogue acts (Bunt, 1994), in order to predict the next dialogue act (Samuel et al. 1998; Stolcke, et al., 2000; Ji & 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). However, the effects of speech acts on social behaviors and roles of conversation participants have not been systematically studied. Research in anthropology and communication has concentrated on how certain social norms and behaviors may be reflected in language (e.g., Scollon and Scollon, 2001; Agar, 1994). But, 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.

Internet-enabled conversation is particularly interesting because in this reduced-cue environment, the only means of engaging in and conveying social behaviors is through written language. 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 telephonic conversations. The use of language by participants as a feature to determine interpersonal relations has been studied by Bracewell et al. (2011) who developed a learning framework to determine collegiality between discourse participants. Their approach, however, looks at singular instances of linguistic markers or single utterances rather than a sustained demonstration of sociolinguistic behavior over the

course of entire discourse. Freedman et al. (2011) have developed an approach that takes into account the entire discourse to detect behaviors such as persuasion; however their analysis is conducted on and models developed upon online discussion threads where the social phenomena of interest may be rare. By contrast, we build our models based on analysis of a data corpus of online chat discourse, where data collection experiments were specifically designed so that the resulting corpus may be rich in sociolinguistic phenomena.

Our research extends the work of Strzalkowski et al. (2010) and Broadwell et al. (2012), who first proposed the two-tiered approach to sociolinguistic modeling and have demonstrated that a subset of mid-level sociolinguistic behaviors may be accurately inferred by a combination of low-level language features. We have adopted their approach and extended it to modeling of leadership and influence. Furthermore, we enhanced the method by adding the evidence learnt from correlations of indices and measures to compute weights through which sociolinguistic behaviors may be combined appropriately to infer higher-level social phenomena.

In this paper, we describe our approach to model Leadership and Influence in online multi-party task-oriented chat dialogues. We show how our models were developed on evidence from online English and Mandarin chat dialogues. Performance on both languages is very encouraging.

2 Sociolinguistic Behaviors to Model Leadership and Influence

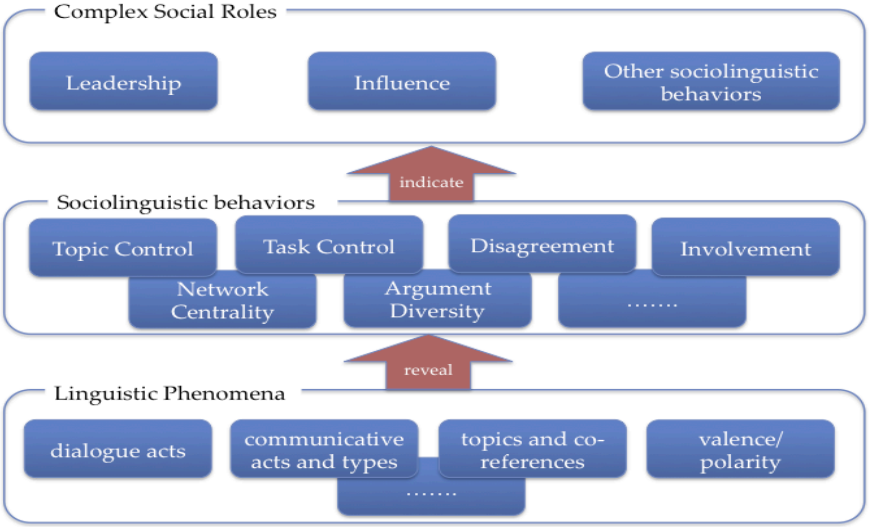

FIGURE 1 – Two-tier approach applied to model social roles in discourse.

In our two-tier approach, we use linguistic elements of discourse to first unravel sociolinguistic behaviors, and then, use the behaviors, in turn, to determine social roles, as shown in Figure 1. It is important to note that, at both levels, our analyses are solidly grounded on sociolinguistic theory. Mid-level behaviors that we shall discuss in this article are Topic Control, Task Control, Disagreement, Involvement, Argument Diversity and Network Centrality that are computed using *indices*. These indices are directly obtained from linguistic elements of discourse, which are described in Section 3. For each participant in the discourse, we compute the degree to which

they engage in sociolinguistic behaviors, using *measures,* which are a linear combination of indices. We describe behaviors, component indices and corresponding measures in this section.

2.1 Topic Control Measure (TCM)

Topic Control is defined as attempts of participants to impose a topic of conversation. This sociolinguistic behaviour is consistent with both Leadership and Influence. In any conversation, whether it is focused on a particular issue or task or is just a social conversation, the participants continuously introduce multiple topics and subtopics. These are called *local topics*. Local topics, following the notion put forth by Givon (1983), may be equated with any substantive noun phrases introduced into discourse that are subsequently mentioned again via repetitions, synonyms, or pronouns. Who introduces local topics, who continues to talk about them, and for how long are some of the indicators of topic control in dialogue. We have developed four indices for Topic Control. Participants who introduce more local topics exert more topic control in dialogue. The first index, called the **Local Topic Introductions Index (LTI)** calculates the proportion of local topics introduced by each participant, by counting the number of first mentions of local topics by each participant as percentage of all local topics in a discourse. The **Subsequent Mentions of Local Topics (SMT)** index calculates the percentage of discourse utterances where the local topics introduced by each participant are being mentioned (by themselves or others) through repetition, synonym, or pronoun. The **Cite Score (CS)** index calculates the percentage of subsequent mentions of local topics first introduced by each participant, but excluding the self-mentions by this participant. The final measure of topic control is the average **Turn Length (TL)** per participant. This index calculates the average utterance length (words) for each participant, relative to other participants.

We shall explain the calculation of one index – Local Topic Introductions (LTI) - in detail. Figure 2 shows a fragment of an actual chat conversation from our corpus with selected local topic references to illustrate how the constructed indices model the concept of Topic Control. In this small excerpt, a few local topics are introduced, including *Carla*, *nanny* and *horses*, as well as possibly others (*record, car,* etc). These local topics are underlined in different ways, with the first mention set in boldface. For example, *Carla* is introduced by speaker JR in turn 1, and is subsequently mentioned by KN (turn 2), LE and KN (via *she*) in turns 3 and 4. Similarly, KN introduces horses in turn 4, and then self-mentions it again in turn 6.

1. **JR**: *wanna go thru **Carlas** resume first ?*
2. **KN**: *i wonder how old carla is*
3. **LE**: *Ha, yeah, when I hear nanny I think she is someone older.*
4. **KN**: *she's got a perfect driving record and rides **horses**! coincidence?*
5. **JR**: *'06 high school grad*
6. **KN**: *i think she rides a horse and not a car!*

FIGURE 2 – Fragment of chat with a few selected local topics highligted.

Once we have computed the scores for each participant on each index, we combine them to compute a single score on the corresponding measure. In Figure 2, we only highlight a few local topics, to illustrate the process. Nouns such as *resume* and *high school*, are not marked for ease of presentation.

In this case, the LTI, SMT, CS and TL indices are combined to get a Topic Control Measure (TCM) for each participant. For instance, suppose we discover the following in a conversation with 7 participants and over 600 total turns:

1. *There are 90 distinct local topics introduced in the conversation.*
2. *The mean rate of local topic introduction is 14.29%.*
3. *Participant LE introduces 23 local topics. Thus, LTI index for LE is 23/90, that is, 23.6%.*
4. *Participant LE scores the highest amongst all on the LTI index in this conversation.*

Using the above information, we can assert that participant LE has the highest topic control in the conversation *based on the LTI index*. This is evidence based on just one index; we compute the index measures for all participants on all component indices to build additional evidence. In our current system prototype, TCM score is computed as the mean of component index scores.

2.2 Task (or Skilled) Control Measure (SCM)

Task Control is an effort by one or more members of a group to define the group's project or goal and/or steer the group towards it. Task Control is gained by telling others to perform certain tasks, or subtasks, or to accept certain decisions about the task. It can also be gained by the speaker offering to perform a task. This sociolinguistic behaviour is primarily consistent with Leadership. One index of Task Control is the number of directives (done as statements or questions) made by each participant as a percentage of all directives in discourse, known as **Directive Index (DI)**. In other words, a participant who tells others what to do (whether overtly or more subtly) is attempting to control the task that the group is performing. Other indices have been developed to support Task Control; these will be explained in future publication.

2.3 Cumulative Disagreement Measure (CDM)

Disagreement has a role to play with regard to leadership and influence in that it is possible that a person in a small group engages in disagreements with others in order to control the topic by way of identifying or correcting what they see as a problem (Ellis and Fisher, 1994; Sanders, Pomerantz and Stromer-Galley, 2010). While each utterance where a participant disagrees with another is a vivid of expression of disagreement, we are interested in a sustained phenomenon where participants repeatedly disagree, thus revealing a social relationship between them. One of the indices we have developed to measure disagreement is the proportion of disagree and/or reject turns produced by a participant that are directed at any other participants in the discourse. This index is called the **Disagree-Reject Index (DRI)**.

2.4 Involvement Measure (INVX)

Involvement is defined as a degree of engagement or participation in the discussion of a group. This behavior is consistent primarily with Leadership. A degree of involvement may be estimated by how much a speaker contributes to the discourse in terms of substantive content. Contributing substantive content to discourse includes introduction of new local topics, taking up the topics introduced by others, as well as taking sides on the topics being discussed. By topics here, we mean the local topics described previously. We have defined five indices in support of Involvement, we shall expand on three of them here. The **Noun Phrase Index (NPI)** is the amount of information content that each speaker contributes to discourse. The NPI measure is calculated by counting the number of content words (e.g., all occurrences of nouns and pronouns referring to people, objects, etc.) in each speaker's utterances as a percentage of all content words

in discourse. The **Turn Index (TI)** is the frequency of turns that different speakers have during a conversation. The **Topic Chain Index (TCI)** is computed by identifying the most frequently mentioned topics in a discourse, i.*e.,* topics chains (i.e., with gaps no longer than 10 turns and then by computing the percentages of mentions of these persistent topics by each participant.

2.5 Network Centrality Measure (NCM)

Another measure is the degree to which a participant is a "center hub" of the communication within the group. In other words, someone whom most others direct their comments to as well as whose topics are most widely cited by others. This behavior is consistent mainly with Influence. Two of the indices used to compute this measure are described. **Communication Links Measure (CLM)** index calculates a degree of Network Centrality for a participant by counting the utterances that are addressed to this participant as a percentage of all utterances in discourse. **Citation Rate Index (CRI)** calculates the degree of Network Centrality for a participant by counting the number of times that the local topics introduced by this participant are cited by other participants. Unlike the Subsequent Mentions measure (SMT) of Topic Control Measure, we calculate CRI by normalizing the citation count by the number of topics introduced by the participant, thus obtaining an average citation rate per topic.

2.6 (Measure of) Argument Diversity (MAD)

Argument Diversity, a behavior consistent with Influence, is displayed by the speakers who deploy a broader range of arguments in conversation. This behavior is signaled by the use of more varied vocabulary, including specialized terms and citations of authoritative sources, among others. A person who uses more varied vocabulary and introduces more unique words into a conversation is considered to have a higher degree of Argument Diversity, which can be measured using the two indices: the **Vocabulary Introduction Measure (VIM)** which is calculated as a proportion of new content words introduced by each participant to all distinct content words in discourse; and the **Vocabulary Range Index (VRI)** which is the number of distinct words used by this participant as a percentage of all distinct words in discourse.

2.7 Combining Indices and Measures

As outlined briefly at the end of Section 2.1, we compute the score of each *measure* by taking linear combination of scores obtained on each *index*. We can thus obtain a full ranking of participants on each sociolinguistic behavior. The measures used to compute Leadership are Topic Control, Task Control, Disagreement, and Involvement. These are indicated in Beebe and Masterson (2006) and borne out in our research. Measures used to compute Influence are Topic Control, Disagreement, Network Centrality and Argument Diversity. Since we have defined Influence as the Initiator-Contributor type of leadership (Bradford, 1978), we shall use those sociolinguistic behaviors that pertain to initiating discussion and contributing substantively in the group. On the other hand, Task Control and Involvement have little or minor role to play in computing Influence, and hence we do not include them while combining behaviors. Similarly, while Task Control and Disagreement are most indicative of Task Leadership, other behaviours such as Network Centrality and Argument Diversity do not correlate with this role. Hence, we do not include them in computation of Leadership. We shall elaborate on this in Section 5, Evaluation and Results.

3 Corpus, Annotation and Computational Modules

The models described in this paper are derived from online chat dialogues. The corpus we use for this analysis is the MPC chat corpus (Shaikh et al., 2010, Liu et al., 2012). This is a corpus of over 90 hours of online chat dialogues in English, Urdu and Mandarin. Participants in these chats are native speakers of these languages. Each chat session is a task-oriented dialogue around 90 minutes in length, with at least 4 participants. This corpus is particularly useful for the type of sociolinguistic analysis we are interested in due to the characteristics of interaction in each chat session – the participants are focused on some task, they form a fairly stable group and the dynamics of conversation unfold naturally through discourse. Other corpora exist such as the ICSI-MRDA corpus (Shriberg et al., 2004) and the AMI meeting corpus (Carletta, 2007), however these are spoken language resources rather than online chat. Where other corpora of online chat do exist, like the NPS Internet chat corpus (Forsyth and Martell, 2007) and StrikeCom corpus (Twitchell et al., 2004), they do not contain any information about the participants themselves or their reactions to the discussion. In order to create a ground truth of assessments of sociolinguistic behavior, we needed certain information to be captured through questionnaires or survey following each data collection session. In the data that comprise MPC corpus, at the conclusion of each chat session, participants were asked to fill out a survey consisting of a series of questions about their perceptions of and reactions to conversation that had freshly participated in. The questions were focused on eliciting responses about sociolinguistic behavior. Questions pertaining to Leadership and Influence are shown in Figure 3. Participants may interpret the notions of socio-linguistic phenomena intuitively, and may rank themselves and other participants accordingly. We refer the reader to the Conclusion section where we address this issue. There are similar questions regarding other behaviors we are interested in modeling and we refer the reader to the cited paper (Shaikh et al., 2010) for a detailed discussion of these.

- *During the discussion, some of the people talking are more influential 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? Scale: Not Influential --- Very Influential.*
- *Below is a list of participants including yourself. Please rank order the participants with regards to leadership.*

FIGURE 3 – Questions regarding sociolinguistic phenomena in post-discussion survey.

We developed a multi-layer annotation process so that automatic modules may be trained to detect and classify social behaviors from discourse. A substantial subset of the MPC corpus was annotated using trained annotators who are native speakers of the respective language. Annotators were trained extensively so that inter-annotator agreement level was sufficiently high (0.8 or higher Krippendorf's alpha). We briefly explain three of the categories that annotation was performed on:

3.1 Communication Links

It is important and very challenging to determine automatically who speaks to whom in multi-party discourse. In our annotation process, we ask annotators to classify each utterance in the chat by marking it as either a) addressed to someone or everyone; b) a response to someone else's specific prior utterance; or c) a continuation of one's own prior utterance. Using annotated data from this layer of annotation; we can train a communication link classification module, which

uses context, inter-utterance similarity and proximity of utterances as some of the features in a Naïve Bayes classifier to automatically classify utterances in one of the above-mentioned three categories. The current performance of this module is 61% accuracy as measured against annotated ground truth data. Indices that are calculated using this automatic module include the Communication Links Measure (CLM).

3.2 Dialogue Acts

We have developed a hierarchy of 15 dialogue acts in order to annotate the functional aspect of an utterance in discourse. The tag set adopted is based on DAMSL (Allen and Core, 1997) and SWBD-DAMSL (Jurafsky et al., 1997), but compressed to 15 tags tuned towards dialogue pragmatics and away from more surface characteristics of utterances. A detailed description of dialogue act tags and annotation procedure has been described in a separate publication. Some dialogue acts that are note-worthy are: Assertion-Opinion, Disagree-Reject, Agree-Accept, Offer-Commit, Information-Request and Action-Directive. Annotated data from this process is used to train a cue-phrase based dialogue act classifier adapted from Webb and Ferguson's (2010) approach, which currently performs at 64% accuracy. Our Cumulative Disagreement Measure (CDM) is calculated using the proportion of disagreement dialogue act utterances detected for each participant by this automatic module. Directive Index (DI) for Task Control is also computed by counting the number of Action-Directive and Offer-Commit types of dialogue acts made by participants.

3.3 Local Topics

Local topics are defined as nouns or noun phrases introduced into discourse that are subsequently mentioned again via repetition, synonym, or pronoun. Annotators were asked to mark all nouns and noun phrases of import from the discussion. We use Stanford part-of-speech tagger (Toutanova et al., 2003) to automatically detect nouns from text. Princeton's Wordnet (Miller et al., 1990) is consulted to identify synonyms commonly used in co-references. Since POS taggers are typically trained on well-formed text, performance of POS tagging on chat text – where grammar may be disorganized, use of abbreviations and symbols etc. may be quite frequent – would affect the accuracy of POS tagging. Our automatic local topic detection module performance is at 70% in the current system prototype. Several indices including Local Topic Introductions (LTI) and Citation Rate Index (CRI) are computed using this module.

We note here that it is not the goal of this research to develop the best POS tagger or the most accurately performing dialogue act classifier. In spite of the shortcomings in the computational modules that support our index calculations, we are able to achieve very robust performance in our intended task of modelling complex social roles. This is because we base our claims of sociolinguistic behaviors on repeated counts of each linguistic phenomenon over the length of entire discourse. When computational modules such as local topic detection fail, such errors are systematic, and would be replicated for each participant in their index scores. If the count for each participant were not fully accurate, nevertheless, the *distribution* of counts for all participants would still hold, thus giving us the desired ranking or the degree of sociolinguistic behaviour for each participant.

Having multiple indices for each behavior helps us account for error introduced from automatic modules. If the predictions on individual indices are not always consistent, we can still combine them into a single output by using different weighting schemes, albeit with lesser confidence. In

order to validate our proposed indices and measures, we analyzed their correlation with each other, both from human annotated data as well as our automatic process, as we shall discuss next.

4 Correlations

4.1 Correlations between proposed indices and behaviors

We compute the scores for each participant of a dialogue for each proposed index. For example, in Table 1, we show the correlation between component indices for Topic Control measure computed for a sample chat session from the MPC corpus – Turn Length (TL), Subsequent Mentions of Local Topics (SMT) and Local Topic Introductions (LTI). If the proposed indices indeed measure the same phenomena, then the correlation between them should be very high. The Cronbach's alpha (1951, 2003) for the scores shown in Table 1 is 0.96, which is extremely high. In Table 2, we show the correlation among selected indices used to compute Involvement measure (INVX). These are Noun Phrase Index (NPI), Turn Index (TI) and Topic Chain Index (TCI). The Cronbach's alpha for this table is also extremely high.

	TL	SMT	LTI	TCM
TL	1.0			
SMT	0.96	1.0		
LTI	0.78	0.80	1.0	
TCM	0.92	0.95	0.88	1.0
α	0.96			

	NPI	TI	TCI	INVX
NPI	1.0			
TI	0.76	1.0		
TCI	0.97	0.83	1.0	
INVX	0.96	0.83	0.98	1.0
α	0.98			

TABLES 1, 2 – Correlation among selected Topic Control and Involvement indices for a sample online chat dialogue

For all sessions we looked at, the correlation among indices of all proposed measures is quite strong, averaging above 0.93 for the Cronbach's alpha reliability statistic. Thus, strong correlations were seen among the index scores both for annotated data as well as automatic computation.

4.2 Correlations between proposed measures and human assessments

In addition to the correlation among indices within a proposed measure, we also computed the correlations among our proposed measures of Leadership and Influence. Figure 4 displays the correlations between the four measures of Leadership – Topic Control Measure (TCM), Skilled Control Measure (SCM), Involvement Measure (INVX) and Cumulative Disagreement Measure (CDM) on an actual 10-participant chat session from MPC corpus. Figure 5 shows the correlations between four measures – Topic Control Measure (TCM), Cumulative Disagreement Measure (CDM), Network Centrality Measure (NCM) and Measure of Argument Diversity (MAD). These measures were calculated for an actual 9-person chat session from the MPC corpus. For both Figures 4 and 5, the x-axes are the anonymized participant names DE, NT and so on and the y-axes are the scores on each of the measures, normalized as percentages.

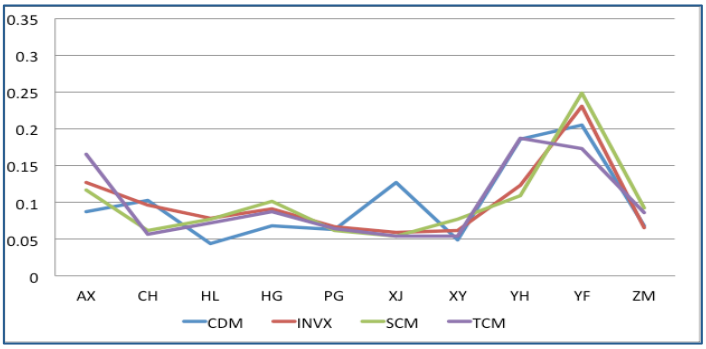

FIGURE 4 – Correlation between selected Leadership measures for a typical chat dialogue

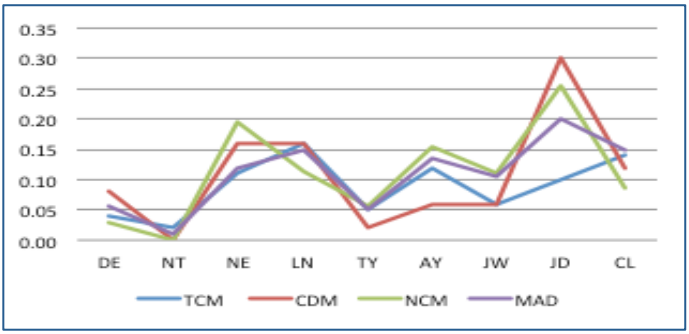

FIGURE 5 – Correlation between selected Influence measures for a typical chat dialogue.

Additionally, we show the correlations between all proposed Leadership and Influencer measures in Tables 3 and 4, to note one important finding. We note that Cumulative Disagreement Measure correlations are lower than the other measures, pointing to evidence of it being the discriminant variable. We have observed similar correlation patterns across the sessions we have looked at.

Leader-ship	SCM	TCM	INVX	CDM
SCM	1.0			
TCM	0.72	1.0		
INVX	0.95	0.77	1.0	
CDM	0.66	0.71	0.76	1.0

Infl-uence	NCM	MAD	TCM	CDM
NCM	1.0			
MAD	0.86	1.0		
TCM	0.98	0.86	1.0	
CDM	0.58	0.59	0.48	1.0

TABLE 3, 4 – Correlation among measures of Leadership and Influence for sample chat dialogue

Computing the correlation against human rankings elicited using survey questionnaire provides us with evidence that indeed the proposed behaviors are measuring the correct phenomena. The correlation between rankings produced by annotated data and ranking induced by participant ratings holds quite strongly across a significant proportion of data sets in our corpus with an average of over 0.80 Cronbach's alpha. Using this evidence of high correlations among indices, among behaviors and their measures, as well as measures against human survey ratings, we can be confident about our approach in measuring and detecting Leadership and Influence.

5 Evaluation and Results

In Table 5, we show how Leadership and Influence are present across different sessions in the MPC corpus. This is determined from participant ratings on post-discussion survey. On average, across both English and Chinese data, in 44.8% of dialogues different participants held the roles of the (task) Leader and the Influencer. Consequently, a significant portion of the MPC corpus has distinct participants exhibiting sociolinguistic behaviors consistent with either Leadership or Influence but not both. In the remaining dialogues, the same person was the leader and the influencer, thus exhibiting both types of behaviors at the same time. An automatic system should be able to distinguish between cases where the leader and influencer are the same person or different; the MPC corpus has a sufficient number of sessions for both cases.

Leader and Influencer are the same participant	**Leader and Influencer are different participants**
55.2%	44.8%

TABLE 5 – Number of sessions in MPC corpus where Leadership and Influence are exhibited by same or different participants (in percentages)

We compute the scores for each participant for all proposed measures. Although we have a full ranking of participants, both from survey ratings as well as system output, we are only interested in participants who have the highest Leadership and Influence. This means, the top-ranking participant on both rankings should match in order evaluate system performance. In cases where the top two individuals are quite close in the survey scores, we may consider top two participants.

We calculate the Leadership and Influencer score for all participants by taking the mean of our measures and deriving a Leadership and Influencer score for each participant. Using this simple weighting scheme of taking an average across all measures of the corresponding behavior, our accuracy is ~51% in predicting the top Leader and ~70% in predicting the top Influencer across our test data set. However, this scheme does not take into account the evidence found using correlations among measures. We have found that, on average, TCM measure correlates higher than other measures for Influence with survey ratings. We also discovered that SCM (Skilled Control Measure) correlates higher in Chinese dialogues than in English for Leadership. Consequently, we devised a weighting scheme that reflects the evidence found from our analysis of correlations against survey ratings.

So, the weighting scheme for English chat dialogues is:

Leader score = $(\alpha_{TCM}$* TCM) + $(\alpha_{SCM}$* SCM) + $(\alpha_{INVX}$* INVX) + $(\alpha_{CDM}$* CDM)

Where $\alpha_{TCM} > \alpha_{SCM} > \alpha_{CDM} > \alpha_{INVX}$

Influencer score = $(\alpha_{TCM}$* TCM) + $(\alpha_{CDM}$* CDM) + $(\alpha_{NCM}$* NCM) + $(\alpha_{MAD}$* MAD)

Where $\alpha_{TCM} > \alpha_{NCM} > \alpha_{MAD} > \alpha_{CDM}$

Similar combinations are derived for Chinese chat dialogues as well. Using this weighting scheme, we compute the Leadership and Influencer scores again. We illustrate this for Leadership in Table 6; the corresponding analysis is also applied for Influence in Table 7. In Table 6, we see that participant CC has the highest score on leadership from survey rating (4.33), followed by AA (score of 4). If we combine the scores computed automatically by our system on Leadership measures, TCM, SCM, INVX and CDM by taking an average, participant AA scores the highest (0.28). However, using the weights learnt from correlations, these scores can be correctly combined to get a score of 0.59 for CC.

Partici -pant	Survey Score	TCM	SCM	INVX	CDM	Leadership without weights	Leadership with weights
AA	4	0.23	0.25	0.31	0.33	0.28	0.48
BB	1.67	0.15	0.13	0.12	0.09	0.12	0.30
CC	4.33	0.27	0.43	0.21	0.14	0.26	0.59
DD	0.67	0.09	0.06	0.09	0.12	0.09	0.20
EE	1	0.10	0.06	0.13	0.09	0.09	0.22
FF	3.33	0.16	0.08	0.15	0.23	0.16	0.29

TABLE 6 – Leadership score computed using linear combination with weights on a sample English chat dialogue

Participant	Survey Score	TCM	CDM	NCM	MAD	Influencer w/ weights
AA	5.5	0.11	0.15	0.06	0.17	0.20
BB	5	0.13	0.15	0.19	0.11	0.28
CC	6.84	0.26	0.25	0.33	0.13	0.49
DD	4.67	0.23	0.25	0.21	0.28	0.44
EE	5	0.15	0.1	0.16	0.14	0.28

TABLE 7 – Influence score computed using linear combination with weights on a sample English chat dialogue

The accuracy of our predictions on our test data set improves to 80% for both Leadership and Influence after factoring in the weights. For different data types and different languages, we have learnt different weighting schemes where the sociolinguistic behaviors may be combined differently to compute scores. In essence, the higher the correlation, the greater the weight given to the measure. As expected and shown in Table 8, different correlations hold across languages – hence possibly cultures, since the participants are native speakers. We can see from the weighting schemes that different behaviors account for Leadership and Influence in different cultures. The Measure of Argument Diversity (MAD) has a higher correlation with Influence in Chinese chat as compared to English chat dialogues. Where the scores are 0, it signifies that the behavior is found to not correlate well with the other measures that comprise the phenomena being modeled; hence we do not include these behaviors while taking linear combinations. They may be either negatively correlated or demonstrate very low correlation; in both cases, the evidence from those behavior should not be included while predicting that sociolinguistic phenomena.

Weights	TCM	SCM	CDM	INVX	NCM	MAD
English Chat Leadership	0.75	0.6	0.3	0.23	0	0
Chinese Chat Leadership	0.68	0.73	0.36	0.45	0	0
English Chat Influence	0.75	0	0.15	0	0.5	0.4
Chinese Chat Influence	0.75	0	0.1	0	0.34	0.75

TABLE 8 – Weighting schemes for combining behaviors learnt from correlation analyses

Illustrated in Table 9, are the correlations between Measure of Argument Diversity (MAD) and Network Centrality Measure (NCM), which are behaviors that are consistent with Influence; and Involvement (INVX) and Skilled Control Measure (SCM) which are the behaviors consistent

with Leadership. We can see that MAD and NCM correlate quite highly with each other; as do INVX and SCM. By contrast, NCM and INVX correlation is very low, as is SCM and MAD. Topic Control Measure (TCM) and Cumulative Disagreement Measure (CDM) are common to both social phenomena; they exhibit high correlations and are not included in Table 9.

Correlation	MAD	NCM	INVX	SCM
MAD	1			
NCM	0.78	1		
INVX	-0.15	0.14	1	
SCM	-0.25	0.11	0.97	1

Table 9. Correlations between Leadership and Influencer measures on sample chat dialogue

In Tables 10 and 11, we show performance of detecting the top leader and influential participant across languages for the two methods we proposed, when compared to baseline. The baseline we have chosen is to pick a participant at random to be top influential person. In a small group discussion found in our corpus, this is a reasonable baseline given the limited number of participants. We could choose another baseline, such as selecting the participant with the most number of turns as the Leader or Influencer. However, we see similar performance for such baselines as the random one.

Performance Leadership	Baseline	Without Weights	With Weights	Including TFM
English Chat	17.85%	37.5%	80%	80%
Chinese Chat	12.5%	64%	72.7%	90.9%
Average	15.65%	50.75%	76.35%	85.45%

TABLE 10 – Performance of system against random baseline, with and without weighting scheme, for Leadership

Performance Influence	Baseline	Without Weights	With Weights
English Chat	17.85%	71.40%	78.50%
Chinese Chat	12.5%	69%	90%
Average	15.65%	70.2%	84.25%

TABLE 11 – Performance of system against random baseline, with and without weighting scheme, for Influence

In a separate publication (Taylor et al., 2012), we have discussed the development of an additional sociolinguistic behavior to predict Leadership in Chinese dialogues – called the **Tension Focus Measure (TFM)**. This behavior is defined as the degree to which a speaker is someone at whom others direct their disagreement, or with whose topics they disagree the most. Using this additional measure, our performance on detecting the top leader goes up to 85% average for English and Mandarin test data set.

Conclusion and perspectives

We have shown a novel, robust method for modeling social phenomena in multi-party discourse. We have combined established social science theories with computational modeling to create a two-tier approach that can detect high-level sociolinguistic phenomena such as Leadership and Influence in language with a high degree of accuracy. In future work, we have planned for a

larger scale evaluation, testing index stability, and resilience to errors in automated language processing, including topic detection, coreference resolution, and dialogue act classification. Current performance of the system is based on versions of these linguistic modules, which perform at about 70% accuracy, so these need to be improved as well. We could also experiment with training a classifier to learn the weights automatically, which we plan to report in a future publication.

The advantage of applying a two-tier approach is that we can add or remove mid-level sociolinguistic behaviors efficiently when applying our models to different data types and languages. As we have noted, we can insert additional sociolinguistic behaviors such as Tension Focus Measure, if our existing models do not completely account for Leadership in certain data sets. Such sociolinguistic analysis is impractical in a straight-forward machine-learning approach where one can add all features to a learning algorithm to decide how features may best be combined. A machine-learning approach modeled directly on linguistic features would not be easily transferable to other data types and could prove brittle. Some measures turn out to be more predictive in a given data genre, and when applied appropriately, perform well at predicting phenomena as rated and understood by human assessors. We note that there may be some variance as to how humans perceive the concept of Leadership and Influence and rate a participant based on their intuitive notion of the concept. The fact that we have multiple indicators in the form of indices and measures helps us overcome the potential variance in this perception. Another advantage of a two-tier approach is that some of the existing measures can be combined differently as we have demonstrated using the CDM and TCM measures. These behaviors are consistent with both Leadership and Influence. When trying to model additional higher-level sociolinguistic phenomena beyond Leadership and Influence, we can use existing measures in a manner that is substantiated by social science theory as well as revealed in our computational analyses.

Acknowledgments

This research was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), and through the U.S. 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 U.S. Government.

References

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

Allen, J. and M. Core. (1997). Draft of DAMSL: Dialog Act Markup in Several Layers. www.cs. rochester.edu/research/cisd/resources/damsl/

Austin, J. L. (1962). *How to do Things with Words*. Clarendon Press, Oxford.

D. B. Bracewell, M. Tomlinson, Y. Shi, J. Bensley, and M. Draper. (2011). Who's playing well with others: Determining collegiality in text. In *Proceedings of the 5th IEEE International Conference on Semantic Computing (ICSC 2011)*.

George Broadwell, Jennifer Stromer-Galley, Tomek Strzalkowski, Samira Shaikh, Sarah Taylor, Umit Boz, Alana Elia, Laura Jiao, Ting Liu and Nick Webb. (2012). Modeling Socio-

Cultural Phenomena in Discourse. *Journal of Natural Language Engineering, Cambridge Press.*

Beebe, S. A., & Masterson, J. T. (2006). *Communicating in small groups: Principles and practices*. Boston, MA: Pearson/Allyn and Bacon

Blaylock, N. (2002). Managing Communicative Intentions in Dialogue Using a Collaborative Problem-Solving Model. Technical Report 774, University of Rochester, CS Dept.

Bradford, L. P. (1978). Group development. La Jolla, Calif., University Associates.

Bunt, H. (1994). Context and Dialogue Control. *Think Quarterly 3(1), 19-31.*

Carberry, S. and L. Lambert. (1999). A Process Model for Recognizing Communicative Acts and Modeling Negotiation Dialogue. *Computational Linguistics*, 25(1), pp. 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): 181-190

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

Cronbach, L. J. Coefficient alpha and the internal structure of tests. (1951). *Psychometrika, 16(3)*, 297-334.

Cronbach, Lee J., and Richard J. Shavelson. (2004). My Current Thoughts on Coefficient Alpha and Successor Procedures. *Educational and Psychological Measurement* 64, no. 3 (June 1): 391-418. doi:10.1177/0013164404266386.

Ellis, D. G., & Fisher, B. A. (1994). *Small group decision making: Communication and the group process*. New York: McGraw-Hill

Forsyth, E. N. and C. H. Martell. (2007). Lexical and Discourse Analysis of Online Chat Dialog. First IEEE International Conference on Semantic Computing (ICSC 2007), pp. 19-26.

Marjorie Freedman, Alex Baron, Vasin Punyakanok and Ralph Weischedel. (2011). "Language Use: What it can tell us?" In *Proceedings of the Association of Computational Linguistics.* Portland, Oregon.

Givon, T. (1983). *Topic continuity in discourse: A quantitative cross-language study*. Amsterdam: John Benjamins.

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

Ji, G. and J. Bilmes. (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)

Jurafsky, D., E. Shriberg and D. Biasca. (1997). Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual. http://stripe.colorado.edu/ ~jurafsky/ manual.august1.html

Levin, L., A. Thyḿe-Gobbel, A. Lavie, K. Ries, and K. Zechner. (1998). A Discourse Coding Scheme for Conversational Spanish. In Proceedings of the International Conference on Speech and Language Processing.

Linell, P. (1990). The power of dialogue dynamics. In Ivana Markov´a and Klaus Foppa,

editors, *The Dynamics of Dialogue*. Harvester, 147–177.

Liu, T., Samira Shaikh, Strzalkowski, T., Broadwell, A., Stromer-Galley, J., Taylor, S., Boz, Umit., Ren, Xiaoai. and Jingsi Wu. (2012) Extending the MPC corpus to Chinese and Urdu - Multi-party Multi-lingual Chat Corpus for Modeling Social Phenomena in Language. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).*

Miller, G. A., Beckwith, R., Fellbaum, C. D., Gross, D., and Miller, K. (1990). WordNet: An online lexical database. *Int. J. Lexicograph*. 3(4): 235--244.

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

Samuel, K., S. Carberry, and K. Vijay-Shanker. (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.

Sanders, R. E., Pomerantz, A., Stromer-Galley, J. (2010). Some ways of taking issue with another participant during a group deliberation. Paper presented at the annual meeting of the National Communication Association, San Francisco, CA

Scollon, R. and S. W. Scollon. (2001). *Intercultural Communication, A Discourse Approach*. Blackwell Publishing, Second Edition.

Searle, J. R. (1969). *Speech Acts*. Cambridge University Press, London-New York.

Shaikh, S., T. Strzalkowski, S. Taylor and N. Webb. (2010). 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. 2010.

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

Stolcke, A., K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer. (2000). Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics* 26(3), 339–373.

Strzalkowski, T., Broadwell, G. A., Stromer-Galley, J., Shaikh, S., Taylor, S., & Webb, N. (2010) Modeling socio-cultural phenomena in discourse. In *Proceedings of The 23rd International Conference on Computational Linguistics. Beijing, China.*

Taylor S., Ting Liu, Samira Shaikh, Tomek Strzalkowski, Jenny Stromer-Galley, Aaron Broadwell, Umit Boz, Xiaoai Ren, Jingsi Wu and Feifei Zhang. (2012). Chinese and American Leadership Characteristics - Discovery and Comparison in Multi-party On-line Dialogues. In *Proceedings IEEE International Conference on Semantic Computing, Special Session, in Palermo, Italy.*

Twitchell, D. P., J. F. Nunamaker Jr., and J. K. Burgoon. (2004). Using Speech Act Profiling for Deception Detection. *Intelligence and Security Informatics*, LNCS, Vol. 3073

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. (2003). Feature-Rich

Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL 2003*, pp. 252-259.

Webb, N. and M. Ferguson. (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. 2010.