# Modeling Influence in Online Multi-Party Discourse

Samira Shaikh[1], Tomek Strzalkowski[1,2], Jenny Stromer-Galley[1], George Aaron Broadwell[1], Sarah Taylor[3], Ting Liu[1], Veena Ravishankar[1], Xiaoai Ren[1] and Umit Boz[1]
[1]State University of New York – University at Albany, NY 12222 USA
[2]Polish Academy of Sciences
[3]Sarah Taylor Consulting, LLC
samirashaikh@gmail.com tomek@albany.edu

***Abstract*— In this article, we present our 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 Influence, which is the focus of this paper. We show robust performance results by comparing our computational results to participants' own perceptions and rankings of influence. We use weights learnt from correlations with known influence rankings to compute and score sociolinguistic behaviors and show performance significantly above baseline for two data sets and two different languages.**

***Keywords: computational socio-linguistics, online dialogues, social phenomena, linguistic behavior, influence, multi-disciplinary artificial intelligence, social computing***

## I. Introduction And Related Work

Our objective is to model high-level sociolinguistic phenomena such as Influence in discourse. This research project aims to develop a computational approach that uses linguistic features of conversational text to detect and model social behaviors of conversation participants. Given a representative dialogue of multi-party conversation, our prototype system automatically classifies the participants by the degree to which they engage in sociolinguistic behaviors, which include Topic Control and Disagreement. These behaviors are mid-level phenomena that are deployed by discourse participants in order to assert higher-level social roles or behaviors such as Influence. Our approach to this problem combines robust computational linguistics methods and established empirical social science techniques.

We define influence in terms of the participants' ability to be a "thought leader" in the group. An influential person is one 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.

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 [2], [21] provides a generalized framework of multiple levels of discourse analysis; work on dialogue analysis [3], [6], [24] focuses on information content and structure of dialogues. However, the effects of speech acts on social behaviors and roles of conversation participants have not been systematically studied.

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. Theories of communication have noted the function of language to exert force on others. Austin's [2] speech act theory advances an informative view of language, noting that language acts on or has a force on others in communication, while Searle's theory articulates categories of speech with distinct forces on interlocutors, creating a power differential between them.

The use of language by participants as a feature to determine interpersonal relations has been studied by Bracewell et al. [4] 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. [12] 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 data may be rich in sociolinguistic phenomena. Strzalkowski et al. [25] and Broadwell et al. [5] have demonstrated using a two-tier approach which operates over the entire discourse that certain mid-level social behaviors and subsequently complex roles like leadership and group cohesion may be accurately inferred by computational modeling of language features. We have adopted their two-tier approach with an enhancement of adding the evidence learnt from correlations of indices and measures to compute weights through which sociolinguistic behaviors may be combined appropriately to infer social phenomena as Influence.

In this paper, we describe our approach to model Influence in online multi-party task-oriented chat dialogues. Our approach works on a variety of data genres including formal meetings and face-to-face discussion transcripts as well as asynchronous online discussion threads. We show how our models were developed on evidence from online English chat dialogues and then adjusted to work on asynchronous threaded discussions. We also conducted these analyses on Mandarin Chinese online chat dialogues as well as Mandarin asynchronous threaded discussions. Performance on both types of data genre and languages is very encouraging, as we shall discuss in the Evaluation section.

## II. Sociolinguistic Behaviors to Model Influence

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 like Influence. Our mid-level behaviors are Topic Control, Disagreement, 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 III. 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 the behaviors, the component indices and the corresponding measures in this section.

### A. Topic Control Measure (TCM)

Our hypothesis is that a participant with a high degree of Topic Control has a high degree of Influence in the group, where Topic Control is defined as attempts of participants to impose a topic of conversation. 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 [13], may be equated with any substantive noun phrases introduced into discourse that are subsequently mentioned again via repetition, synonym, or pronoun. Some of these local topics may be talked about in only a couple of turns, while others may persist for much longer; some of them will be relevant to the overall discussion, while others may appear like digressions. Who introduces local topics into conversation and who continues to talk about them, and for how long are some of the indicators of topic control in dialogue.

These indicators are the basis of the following four indices:

*1) Local Topic Introductions (LTI)*

Participants who introduce more local topics exert more topic control in dialogue. This index 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.

*2) Subsequent Mentions of Local Topics (SMT)*

Participants, who introduce local topics that are subsequently widely discussed, exert a high degree of topic control in discourse. This 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

*3) Cite Score (CS)*

Participants, who introduce topics that are subsequently more frequently discussed by others, rather than merely by themselves, exert more topic control in discourse. This index calculates the percentage of subsequent mentions of local topics first introduced by each participant, but excluding the self-mentions by this participant.

*4) Turn Length (TL)*

The final measure of topic control is the average turn length per participant: participants who have, on average, longer dialogue turns exert more topic control in discourse. This index calculates the average utterance length (in words and other symbols) for each participant, relative to other participants.

Figure 1 shows a fragment of chat 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. 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 1. Fragment of chat with selected topics highlighted.

We have developed multiple indices to compute a sociolinguistic behavior in order to find the indices with the best predictive power. This predictive power can be measured by how well the indices to compute a single measure correlate with each other. Another argument for having multiple indices is to test the accuracy and the utility of automatically computing them from conversational text.

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 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 initial system prototype, the TCM score is computed by taking the mean of the component index scores.

Thus, multiple indices are used to indicate a single behavior. We take the same approach on our second tier, where we have developed multiple measures to indicate Influence. This multiplicity is intentional; in order to determine different aspects of Influence we use different behaviors. The behaviors, implemented in our system as measures, are validated by social science literature as well as their correlation with each other. We describe three other measures of Influence next.

### B. *Cumulative Disagreement Measure (CDM)*

Disagreement has a role to play with regard to influence in that it is possible that an influential 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 [10], [20]. 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 index 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. According to this index, called the **Disagree-Reject Index (DRI)**, a participant who makes more utterances of disagreement, disapproval, or rejection is considered to produce a higher degree of disagreement in discourse. There are additional indices that we have developed for this measure, which are described in a separate larger publication.

### C. *Network Centrality Measure (NCM)*

Another measure of Influence is the degree to which a participant is a "center hub" of the communication within the group. In other words, the influencer is someone whom most others direct their comments to as well as whose topics are most widely cited by others. Three of the indices used to compute this measure are described:

*1) Communication Links Measure (CLM)*

This 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.

*2) Citation Rate Index (CRI)*

This index 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, 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. Participants with higher CRI scores have a higher degree of Network Centrality.

*3) Meso-topic Introduction (MTI)*

This index is a variant of LTI index of Topic Control but is applied to meso-topics rather than all local topics. Meso-topics are the most persistent local topics, in other words they are widely cited in long stretches of discourse. Participants who introduce more meso-topics have a higher degree of Network Centrality because these topics are widely cited by others. In our current prototype, meso-topics are those local topics that have more than 10 subsequent mentions.

### D. *Measure of Argument Diversity (MAD)*

Social science research [14] indicates that influential participants use a broader range of arguments in conversation. This may be signaled by a wider vocabulary, including citations of authoritative sources as well as use of specialized terminology. 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.

Thus, the four measures we described in this section, which have been developed to compute and measure the corresponding sociolinguistic behaviors – Topic Control, Cumulative Disagreement, Network Centrality and Argument Diversity; are now combined to compute a measure of Influence. This is done by taking a linear combination of participant scores on the measures and calculating a single score for Influence. The induced ranking of participants is their degree of Influence. The participant with the highest score has the highest degree of Influence in

the conversation. All the above measures and indices have been validated by running correlations as well as by comparing against the rankings produced by discourse participants on a survey questionnaire designed specifically for this type of research. First, we elaborate on the data corpus and annotation process used to build our computational models.

## III. Data, 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 [22], [18]. 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 [23] and the AMI meeting corpus [7], however these are spoken language resources rather than online chat. Where corpora of online chat do exist, like the NPS Internet chat corpus [11] and StrikeCom corpus [27], they do not contain any information about the participants themselves or their reactions to the discussion. In Tomlinson et al. [26], a corpus has been created from Arabic asynchronous threaded discussions where they have attempted to perform annotations the social act level. 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 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. One such question pertaining to Influence is shown in Figure 2. There are similar questions regarding other behaviors we are interested in modeling and we refer the reader to the cited paper [22] 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: Very Influential --- Not Influential.*

Figure 2. Post session survey question related to Influence in the MPC chat corpus

Participants rated each other as well as themselves on this question on a scale of 1-10. Using this rating, we can compute a ranking of Influence by taking the mean of scores obtained by a participant for this question. This ranking serves as the ground against which we can compare our system performance.

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 [16]). We briefly explain three of the categories that annotation was performed on:

*1) Communication Links:*

It is important and very challenging to determine 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 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 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).

*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 [1] and SWBD-DAMSL [15], 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. Annotated data from this process is used to train a cue-phrase based dialogue act classifier adapted from Webb and Ferguson's [28] 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.

*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 [17] to automatically detect nouns from text. Princeton's Wordnet [19] 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.

For a complete discussion on how annotation was performed and how our computational modules are trained to recognize linguistic features of text, we refer the readers to a more detailed publication [5].

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 behavior 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 consistently for each participant. 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 behavior for each participant.

Having multiple indices for each behavior helps us account for the error introduced from our 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.

## IV. Correlations

### A. Correlations between proposed indices and behaviors

We compute the scores for each participant of a dialogue for each proposed index. For example, in Table I, 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 [8], [9] for the scores shown in Table I is 0.96, which is extremely high. 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.

TABLE I. Correlation among selected Topic Control indices for a sample online chat dialogue

| | 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 | | | |

### B. 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 Influence. Figure 3 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. On the x-axis are the participant names DE, NT and so on and the y-axis is the score on each of the four measures, normalized as percentages.

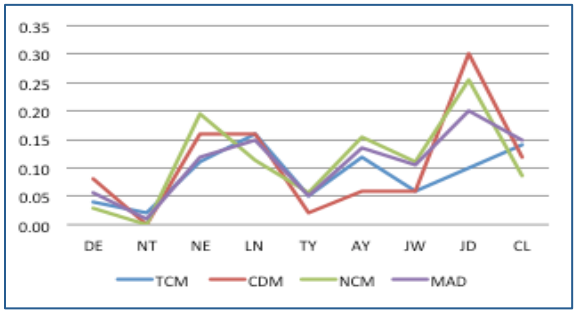


Figure 3. Correlation between selected Influencer measures for a typical chat dialogue

Table II shows the correlations between all proposed Influencer measures. 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. 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.

TABLE II. CORRELATION AMONG MEASURES FOR A SAMPLE CHAT DIALOGUE

| | 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 |

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 Influence. We demonstrate our evaluation and results in the next section.

## V. EVALUATION AND RESULTS

We compute the scores for each participant for the four proposed measures. Shown in Table III, are the scores for five participants on four measures – TCM, CDM, NCM and MAD normalized as percentages. The 3rd participant scores the highest among others on Influence in the participant ratings, score of 6.84. 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 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 Influencer score for all participants by taking the mean of our four measures and deriving an Influencer score. According to system score, the 3rd and 4th participant in Table III score high, the 4th participant scoring slightly higher (0.243 and 0.244). The system does not predict the Influencer correctly in this instance. Using this simple weighting scheme of taking an average across all measures of Influence, our accuracy is ~71% 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 others with survey ratings. Consequently, we devised a weighting scheme that reflects the evidence found from our analysis of correlations against survey ratings.

In this weighting scheme for English chat dialogues:

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

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

Using this weighting scheme, we compute the Influencer scores again, as shown in Table IV, and can correctly predict the top Influential person. The accuracy of our predictions on our test data set improves to 78.50% for English chat dialogues 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 Influencer score. In essence, the higher the correlation, the greater the weight given to the measure.

Next, we wish to apply the models learnt from analysis of online chat discourse to a different data genre – asynchronous threaded discussions. For this experiment, we downloaded 30 Wikipedia discussions talks, centering on a variety of topics, with a sizable number of turns (posts) and a sufficient number of participants. While selecting the data, we were careful in defining and choosing only those threads where the characteristics of discourse were satisfactory, such as the ratio of turns to participants should be higher than 3. Since we could not obtain participants own ratings of sociolinguistic phenomena on these threaded discussions, we asked trained assessors to study the discussions and mark participants they observed to be highly influential. Note that these assessments are not analogous to annotations; we did not ask assessors to indicate every instance of sociolinguistic behavior in the discourse, rather we asked them to study the discussion as a whole and rate the participants on their degree of influence. It could occur that no individual participant would be prominently influential; in which case, there would be no prominent influencer in the discourse. Our prototype system should be able to detect such instance and return appropriate results. We have created this corpus of assessed Wikipedia threaded discussions in English and Chinese; the discussions are chosen where multiple assessors agree with regard to the Influential participant. Inter-assessor agreement for this task is above 80%.

In conducting these analyses, we applied the same procedures as with online chat dialogues of learning the weights of social-linguistic behaviors based on correlations with the assessors ratings of influence. Table V shows the weights learnt for the various types of data and languages we applied our models to. English and Chinese chats refer to the English and Mandarin chat dialogues from the MPC corpus, English and Chinese wiki refer to the asynchronous discussion threads we acquired from English and Chinese Wikipedia. As expected, different correlations hold across languages – hence possibly cultures, since the participants are native speakers – and across data genres.

TABLE III. INFLUENCER SCORE COMPUTED BY TAKING THE MEAN ACROSS PROPOSED MEASURES ON A SAMPLE ENGLISH CHAT DIALOGUE

| Survey Score | TCM | CDM | NCM | MAD | Influencer |
|---|---|---|---|---|---|
| 5.5 | 0.11 | 0.15 | 0.06 | 0.17 | 0.12 |
| 5 | 0.13 | 0.15 | 0.19 | 0.11 | 0.15 |
| 6.84 | 0.26 | 0.25 | 0.33 | 0.13 | 0.243 |
| 4.67 | 0.23 | 0.25 | 0.21 | 0.28 | 0.244 |
| 5 | 0.15 | 0.1 | 0.16 | 0.14 | 0.14 |

TABLE IV. INFLUENCER SCORE COMPUTED USING LINEAR COMBINATION WITH WEIGHTS ON A SAMPLE ENGLISH CHAT DIALOGUE

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

TABLE V. WEIGHTING SCHEMES FOR COMBINING SOCIOLINGUISTIC BEHAVIORS LEARNT FROM CORRELATION ANALYSES

| Weights | TCM | CDM | NCM | MAD |
|---|---|---|---|---|
| English Chat | 0.75 | 0.15 | 0.5 | 0.4 |
| English Wiki | 0.3 | 0.7 | 0.2 | 0.45 |
| Chinese Chat | 0.75 | 0.1 | 0.34 | 0.75 |
| Chinese Wiki | 1.0 | 1.0 | 1.0 | 1.0 |

We can see from the weighting schemes that different behaviors account for Influence in different cultures and types of data. The Cumulative Disagreement measure has a higher correlation with Influence in asynchronous threaded discussions as compared to online chat dialogues. Also, we found that there were no significant differences in correlations for Chinese wiki threaded discussions. Thus, the weighting scheme for Chinese wiki discussions is the same as taking an un-weighted mean.

In Table VI, we show performance of detecting the top influential participant across languages and data types 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, both for online chat and Wikipedia discussion threads, 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 Influencer. However, we saw the similar performance for such baselines as the random one.

Table VI shows that by applying a social science-informed approach aided by computational linguistics, we can achieve significant performance gain over baseline. On average, we are able to accurately predict the top influential participants 66% of the time across two different languages and types of discourse. With the addition of weighting schemes learnt from correlation of our proposed measures with human assessments of influence, we are further able to improve our performance. Our system can also account for and predict cases where no prominent Influencer is found. Accuracy scores seen in this table are computed against human ratings (in the case of chat dialogues) and assessments (in the case of wiki discussions). As seen in Table VI, accuracy is 90% for Chinese chat dialogues, while for English chat dialogues and wiki discussions accuracy is 78.5% and 85% respectively. Further detailed investigation is needed for Chinese threaded discussions, to understand where the models may be missing evidence of Influence and which appropriate weighting scheme may be applied.

TABLE VI. PERFORMANCE OF SYSTEM AGAINST RANDOM BASELINE, WITH AND WITHOUT WEIGHTING SCHEME

| Performance | Baseline | Without Weights | With Weights |
|---|---|---|---|
| English Chat | 17.85% | 71.40% | 78.50% |
| English Wiki | 10% | 75% | 85% |
| Chinese Chat | 12.5% | 69% | 90% |
| Chinese Wiki | 5% | 50% | 50% |
| Average | 11.33% | 66.35% | 75.8% |

## VI. CONCLUSION

We have a 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 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, our behaviors seem to be missing evidence to predict adequately Influential behavior in Chinese wiki dialogues. We can remedy this by modeling additional sociolinguistic behaviors, in addition to TCM, CDM, NCM and MAD, which would then only be applied to that specific data corpus. Such sociolinguistic analysis is impractical in a straightforward 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 applied directly on linguistic features may be brittle and not easily transferable to different genres. Some measures turn out to be more predictive in a given data genre, and when applied appropriately, perform well at

predicting Influence as rated and understood by human assessors. We note that there may be some variance as to how humans percieve the concept of 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, when trying to model additional higher-level sociolinguistic phenomena beyond Influence.

## ACKNOWLEDGMENT

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

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

[2] Austin, J. L. *How to do Things with Words*. 1962. Clarendon Press, Oxford.

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

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

[5] George Broadwell, Jennifer Stromer-Galley, Tomek Strzalkowski, Samira Shaikh, Sarah Taylor, Umit Boz, Alana Elia, Laura Jiao, Ting Liu and Nick Webb. "Modeling Socio-Cultural Phenomena in Discourse". 2012. *Journal of Natural Language Engineering, Cambridge Press.*

[6] Carberry, S. and L. Lambert. "A Process Model for Recognizing Communicative Acts and Modeling Negotiation Dialogue". 1999. *Computational Linguistics*, 25(1), pp. 1-53.

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

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

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

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

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

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

[13] Givon, T. Topic continuity in discourse: A quantitative cross-language study. 1983. Amsterdam: John Benjamins.

[14] Huffaker, D. "Dimensions of leadership and social influence in online communities". 2010. *Human Communication Research, 36*, 596-617.

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

[16] Krippendorff, Klaus. 'Content analysis: an introduction to its methodology. "2004. SAGE.

[17] Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network". In *Proceedings of HLT-NAACL 2003*, pp. 252-259. 2003.

[18] Liu, T., Samira Shaikh, Strzalkowski, T., Broadwell, A., Stromer-Galley, J., Taylor, S., Boz, Umit., Ren, Xiaoai. and Jingsi Wu. "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). 2012.

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

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

[21] Searle, J. R. *Speech acts: an essay in the philosophy of language*. 1969. Cambridge: The University Printing House.

[22] Shaikh, S., T. Strzalkowski, S. Taylor and N. Webb. "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.

[23] Shriberg, E., R. Dhillon, S. Bhagat, J. Ang and H. Carvey. "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

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

[25] Strzalkowski, T., Broadwell, G. A., Stromer-Galley, J., Shaikh, S., Taylor, S., & Webb, N. "Modeling socio-cultural phenomena in discourse". The 23rd International Conference on Computational Linguistics. Beijing, China. 2010.

[26] Marc Tomlinson, David B. Bracewell, Mary Draper, Zewar Almissour, Ying Shi, and Jeremy Bensley. "Pursing power in arabic on-line discussion forums." In Proceedings of the Eighth Conference on International Language Resources and Evaluation. 2012.

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

Webb, N. and M. Ferguson. "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.