VERBAL BEHAVIOR OF THE MORE AND THE LESS INFLUENTIAL MEETING PARTICIPANT

Rutger Rienks
University of Twente
Enschede, The Netherlands
rienks@ewi.utwente.nl

Anton Nijholt
University of Twente
Enschede, The Netherlands
anijholt@ewi.utwente.nl

Dirk Heylen
University of Twente
Enschede, The Netherlands
heylen@ewi.utwente.nl

ABSTRACT

We test the strength of the relationship between the way that people behave in a discussion and their level of influence on the basis of some empirical grounds. We use the data sources that were collected from the AMI corpus for the experiments in the areas of argumentation, dialogue-act and influence research. Statistical dependencies and (cor)relations between the tags are mined for possible relationships.

1. INTRODUCTION

Amongst the many definitions that exist for argumentation Van Eemeren et al. [7] define argumentation as a social, intellectual, verbal activity that serves to justify or to refute an opinion, consisting of a constellation of statements and that is directed towards obtaining the *approbation* of an audience. The interesting word here is approbation which indicates that the goal of argumentation is to get something approved, regardless of its truth value. Through approbation, a change of attitude, or belief towards a certain issue, is to be established. So, independent of the fact whether the message one is trying to convey is true by itself, one could still bring arguments trying to *persuade* the other. Persuasion in turn, entails some level of influence, as it guides or forces people towards the adoption of an idea, a claim, an attitude, or an action. By using valid arguments, one may change the attitude of the other. A change in attitude can, vice versa, be regarded as a sign of influence. It is therefore not unlikely to expect a relationship between the phenomena of argumentation and influence.

The aim of this paper is to test the strength of the relationship between the way that people behave in a discussion and their perceived level of influence on the basis of some empirical grounds. Using the data sources that were collected from the AMI corpus for the experiments in the areas of argumentation [9], dialogue-act [8], and influence research [6], statistical dependencies and (cor)relations between the tags are mined for possible relationships. Here we present the results. The remainder of this paper is organized as follows: First the data that is used is presented before rule induction is applied to look for regular patterns in the collection of the combined data sets. Next it is examined in greater detail where, at the meeting level, differences exist amongst both the tag distributions of dialogue acts and argumentation for the various influence types. The experiments are concluded by examining if the (significant) differences that were found could be exploited for classification purposes by cross-fertilizing features. Given the results of these experiments the final section tries to provide a profile of 'Highly' influential participants in relation to 'Lowly' influential participants.

ICMI 07 Workshop on Tagging, Mining and Retrieval of Human-Related Activity Information, November 15, 2007, Nagoya Japan.

2. THE DATA

One of the major deliverables of the AMI project is the development of a meeting corpus [3] that aims to benefit a range of research communities, including those working on speech, language, gesture, information retrieval, and object tracking, as well as organizational psychologists and sociologists interested in how groups of individuals work together as a team. Progress in the interaction research requires a large data set that allows for empirical observations and on which the foreseen technologies can be developed.

These requirements resulted in a scenario for design meetings in which four persons, as a project team, in a sequence of four meetings had to develop a design for a remote control [5]. Capturing meetings 'in the wild' would have resulted in a too wide variety of meetings. The scenario was used to achieve controlled yet natural interaction between the meeting participants, rather than using predefined scripts that told participants explicitly what to do and how to behave. In the scenario, four participants play the roles of employees of an electronics company that has to develop a new type of television remote-control in order to create an attractive, user-friendly remote-control that could beat the unattractive and old-fashioned ones currently on the market. The participants were told that they were joining a design team whose task, over a day of individual work and group meetings, is to develop a prototype. The participants were assigned four distinct roles: Project Manager (PM), Marketing Expert (ME), User Interface Designer (UI) and Industrial Designer (ID).

Over one hundred hours of meetings were recorded that followed the same scenario. This hub corpus comprised a series of 30 completed scenarios, 120 different meetings in total. Participants were neither professionally trained for design work nor experienced in their role. This was done to

assure the same starting point whilst aiming for replicable behavior for all those playing the same role [3]. English is used in the meetings. Several participants were non native speakers. Given the size of the data set it has been further assumed that other individual differences were levelled out.

Typical sensors that were used for capturing the data were cameras (recording global and close-up views), lapel microphones, microphone arrays, a whiteboard and smart pens. The corpus provides meta-information such as the seating arrangement, and the (powerpoint) presentations that were used have been collected. The recorded data, including layers of annotation such as manually created transcripts, dialogue acts and summaries are all publicly available[1]. Figure 1 shows a global camera view of a meeting.

Figure 1: An overview image from one of the AMI meetings recorded at IDIAP

In the following paragraphs we provide some further information on the subsets of the AMI corpus and the annotation schemes that will be used in the present study.

2.1 The Argumentation Annotations

The Twente Argument Schema (TAS) is an annotation schema designed to create argument diagrams from meeting transcripts. It identifies the argumentative functions of the different contributions made by debating participants and labels the relations that exist between these contributions. Following most of the existing diagramming techniques, application of the method results in a structure with labelled nodes and edges. The nodes of the tree contain complete speaker turns or parts of it, whereas the edges represent the type of relation between the nodes. An overview of the complete annotated set is shown in Table 1. For more information on the specific meaning of the labels consider [9].

2.2 The Dialogue-act Annotations

Most of the meeting transcripts comprising over 100,000 utterances, have been annotated for dialogue acts[2] The AMI dialogue act scheme consists of 15 dialogue acts. For an overview of this data consider Table 2.

Node labels	Amount
Statement (STA)	4077
Weak statement (WST)	194
Open issue (OIS)	232
A/B issue (AIS)	69
Yes/No issue (YIS)	443
Other (OTH)	1905
Total	6920
Relation labels	Amount
Positive	2319
Negative	471
Uncertain	259
Request	223
Specialization	131
Elaboration	689
Option	601
Option exclusion	14
Subject-to	190
Total	4897

Table 1: Distribution of TAS labels.

2.3 The Influence Annotations

In 40 meetings, the participants were asked to rank all participants of their meeting, including themselves, from most to least influential by assigning them unique nominal values ranging from one (most influential) to four (least influential). Participants were not allowed to rank people equivalently. The collected permutations of the numbers one, two, three and four, were quantized into three classes as described in [6]. The resulting data set had a total of 160 labels (40 meetings times four participants) resulting in 34 observations for 'Low', 91 for 'Normal', and 35 for 'High'.

2.4 The Dataset Used

As not all of the annotation levels are available for all meetings, our investigation can use only a subset. The combined influence - dialogue act annotations were available for 30 AMI meetings and the combined influence - argumentation information was available for 29 discussions distributed over 18 meetings. All in all 865 of the total of 6920 (12.5%) TAS unit labels were covered with influence information. 263 of these TAS units were produced by a 'Highly' influential participant, 474 by a 'Normally' influential participant and 155 by a 'Low' influential participant. The distribution of the unit labels comprised 4 A/B Issues (an issue were the solution is restricted to a fixed number of choices), 24 Open Issues, 51 Yes/No Issues, 464 Statements, 20 Weak Statements and 302 Others. To increase the number of samples per category, the argument labels were grouped into three main categories (Issues, Statements and Other). All in all it resulted in the data set that is shown in Table 3.

A first exploration reveals that the distribution of argument labels as a function of the influence values does not turn out to be significant ($\chi^2(4, N = 864) = 4.73, P < 0.31$), nor do ANOVA tests on the individual labels show any sig-

[1]http://corpus.amiproject.org

[2]The procedure is described in the AMI Guidelines for Dialogue Act and Addressee Annotation V1.0, Oct 13, 2005.

Label	Amount	**Label**	Amount
Fragment	14348	Assessment	19020
Backchannel	11251	Comment about understanding	1931
Stall	6933	Elicit assesment	1942
Inform	28891	Elicit comment about understanding	169
Elicit Inform	3703	be positive	1936
Suggest	8114	be negative	77
Offer	1288	Other	1993
Elicit Offer or Suggestion	602		
Total	102198		

Table 2: Distribution of Dialogue acts in the AMI corpus.

	low	Normal	High	Total
Issues	12	40	27	79
Statements	78	254	152	484
Other	65	153	84	302
Total	155	447	263	865

Table 3: Distribution of label combinations for combined argumentation (merged) and influence data.

nificant results. As a consequence, one might conclude that both phenomena seem to be independent.

Not taken aback by this somewhat discouraging result we took some closer looks to investigate for possible other interdependencies. The results are presented in the following sections.

3. RULE INDUCTION

In this section an unsupervised mining method known as association rule mining is used to explore the data for patterns. Association rule mining finds associations and/or correlation relationships among large sets of data items. These resulting association rules bring to light feature value conditions that co-occur in any given data set. Association rules contain a precondition (antecedent) and a conclusion (consequent). The precondition is a series of constraints that is laid over the features and the conclusion generally gives the label that applies to instances covered by the constraints. An association rule can typically be expressed by an 'If a Then b' clause, where the preconditions are specified in the a part and the conclusions in the b part [11].

Many different association rules can be derived from even a tiny data set, not all of them of interest. Only rules that apply to a reasonably large number of instances and have a reasonably high accuracy on the instances they apply to are worth considering. The rules that are found are therefore usually ranked according to their 'strength'.

The 'Tertius' algorithm [4] we used for rule mining provides two measures for the strength of the rule: the confirmation value[3] and the frequency of counter-instances (the number of counter-instances divided by the total number of data items). A rule is said to be better than another if it has a higher confirmation value.

[3]The confirmation value trades off the decrease in counter-instances from expected to observed and the ratio of expected but non observed counter-instances (see [4] for more detail).

For this experiment the influence class labels and the fractions of the various argumentation labels per meeting were used. To allow the data to be used for rule induction, the label fractions were quantized in three nominal categories 'High', 'Normal' and 'Low' using WEKA's simple binning algorithm [11]. This was done to get hold of the argumentation label distributions per influence category. The top three rules are shown per influence category in Table 4.

I	II	Antecedent	Consequent
0,164	0,448	Sta = 'normal'	LOW
0,155	0,405	Iss = 'low' and Sta = 'normal'	LOW
0,103	0,155	Sta = 'normal' and Oth = 'high'	LOW
0,145	0,000	Iss = 'low' and Sta = 'low'	NORMAL
0,112	0,043	Sta = 'low'	NORMAL
0,110	0,026	Sta = 'low' and Oth = 'high'	NORMAL
0,130	0,293	Oth = 'normal'	HIGH
0,101	0,216	Sta = 'normal' and Oth = 'normal'	HIGH
0,084	0,000	Iss = 'high' and Sta = 'normal'	HIGH

Table 4: Induced rules with the Tertius algorithm, where the consequent is an influence class. I= confirmation value of the rule, II= the observed frequency of counter-instances of the rule in the data set.

Table 4 shows that the fraction TAS unit label distribution sums to one for all the individual influence type categories. This means that if one particular TAS unit class has a relatively low fraction, another class automatically has a relatively high fraction. From Table 4, one can distill that it seems that a high 'Issue' frequency in combination with a low 'Other' frequency seems to be more representative for highly influential people. People of low influence, on the other hand, score high on the 'Other' units and low on the 'Issues'. As could be expected from the confirmation values, post-hoc statistical analysis revealed that these hypotheses do not prove to be statistically significant (cf. Section 4.1).

A second experiment was performed with a data set containing the influence values added to all TAS unit labels and its associated features (including the relation that attaches the node to the tree). All of the features were again binned into the three (high, normal and low) bins. The top three rules[4] that were induced from the data for each influence category are reported in Table 5.

[4]Note that confirmation rank is dependent on the number of features and that rankings of rules between tables can therefore not be compared.

I	II	Antecedent	Consequent
0,079	0,009	ORT = 'low' and LL = STA and LL2 = YIS	HIGH
0,079	0,009	LL = STA and LL2 = YIS and NS = 'high'	HIGH
0,076	0,003	QMT = 'high' and L = 'low' and node = STA	HIGH
0,094	0,007	L = 'low' and rel = OPT and DB = 'normal'	NORMAL
0,091	0,007	QMT = 'low' and rel = OPT and DB = 'normal'	NORMAL
0,091	0,007	rel = OPT and DB = 'normal' and node = STA	NORMAL
0,093	0,621	QMT = 'low' and ORT = 'low' and L = 'low'	LOW
0,088	0,593	QMT = 'low' and ORT = 'low' and LPS = 'low'	LOW
0,086	0,635	ORT = 'low' and L = 'low' and LPS = 'low'	LOW

Table 5: The top three induced rules with the Tertius algorithm where the consequent is an influence class. I= confirmation value of the rule, II= the observed frequency of counter-instances of the rule in the data set. Features: ORT = Or-token present, LL = value of the last label, LL2 is value of the label before LL, NS = new speaker, QMT = Question mark token present, DB = Branch Depth, L = Length of Current Segment, LPS = Length Previous Segment.

From the deduced rules shown in Table 5 one could conclude that relatively high influential people respond to people who provide responses to Yes/No issues. People with a relatively low influence level seem to use fewer question marks, use the word 'or' less frequently and provide relatively short responses. This seems to align with the finding reported above that influential participants seem to raise more 'issues' and generally provide less units that can be labelled as 'other'.

Besides rule induction, one can use other methods to look for correlations.

4. A CLOSER LOOK

This section reports on experiments that were conducted to find out other dependencies between the TAS scheme and the participants influence rankings than those that could emerge via rule induction.

4.1 TAS units and influence

We started by conducting three different kinds of experiments to see whether, and if so which, aspects in relation to the TAS unit labels could be (cor)related to the various influence levels. Examined for possible relationship with the influence rankings were: the total number of units, the average unit duration, and the unit type distributions.

4.1.1 Examining the number of TAS units

When considering the number of TAS units uttered per person per meeting, an average of 7.27 was found with a standard deviation of 3.56. No significant differences were found with respect to the number of turns for each type of influence level. When zooming in on the contribution of turns along the discussion (split up in five bins of equal time intervals) we obtain Figure 2.

Apart from the fact that no difference exists in the total number of TAS units uttered per influence level, no significant difference for the various influence levels when considering the number of TAS units uttered per bin were found. A significant positive correlation, however, was found between the fraction of turns and the progress of the discussion for all influence levels combined (Pearson's correlation coefficient r=0.22, with a significant regression model $F(1) = 30.34, P < 0.001$) as well as for the separate influence levels (r between 0.24 and 0.19, $P < 0.01$ for 'Medium' and $P < 0.03$ for 'Low' and 'High').

This finding shows that towards the end of the discussion people tend to talk in shorter turns. A logical explanation for this might be that people reach agreement towards the end, and that contributions in terms of 'yeah' and 'sure' occur more frequently. Another, but perhaps less likely explanation could be that people start to run out of time and therefore try to limit the length of their contributions.

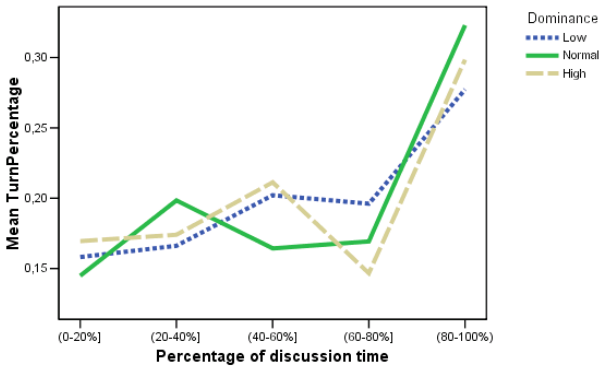

Figure 2: The fraction contributions divided over five time intervals per influence type.

4.1.2 Examining the duration of the TAS units

To examine in more detail the finding that turns towards the end of a discussion seem to be a little shorter, the average duration of the turns was computed for the same discussion intervals. The results are shown in Figure 3.

Statistical analysis on this data revealed a significant decrease in turn duration as the meeting progresses for all the influence levels individually (Pearson's correlation coefficient r between -0.18 and -0.11, $P < 0.01$ for 'High' and $P < 0.03$ for 'Normal' and 'Low') as well as for all levels combined (r=0.15 with a significant regression model F(1)=19.66, $P < 0.001$). This was expected, when looking at our earlier finding that the number of turns increases along with the meeting. One could, when considering Figure 3, get the idea that less influential people generally resort to shorter turns more quickly than more influential people.

This is interesting, because most decisions are taken at the final stage of a discussion. However, when considering the individual time intervals, one-way ANOVA with post-hoc Fisher-LSD testing showed no significant differences between the average duration of the turns for the various influence levels. This could be due to our relatively scarce data set.

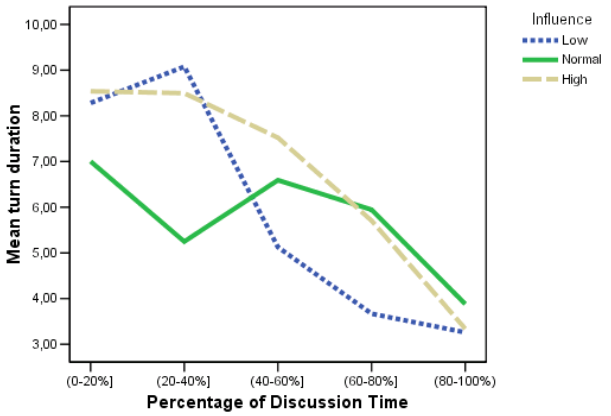

Figure 3: The average turn duration over a discussion per influence type.

4.2 Dialogue acts and influence

It was shown in the previous section that for the three grouped argument labels no significant differences existed in their distributions over the three influence categories. A possible explanation for this could be the relative scarcity of examples. This, combined with the fact that we grouped the unit labels, was the reason to conduct some extra experiments. It was decided to examine more closely whether and how, certain categories of dialogue-acts can be related to the various influence rankings over the course of a meeting. Dialogue act annotations were available for 30 of the 40 meetings with influence rankings.

As a first attempt the fractions for the occurrence of all dialogue-acts was computed for all participants. These results were subsequently merged for each of the influence levels. The resulting average fractions are shown in Figure 4.

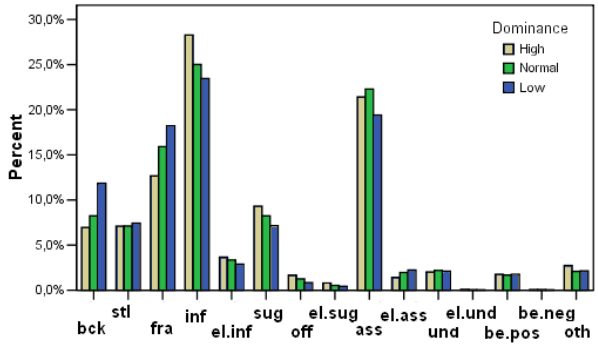

Figure 4: The fraction of dialogue acts per influence level.

Figure 4 seems to suggest that there are some interesting differences between the various dialogue act distributions. Statistical analysis by means of ANOVA showed that on the P<0.05 level significant differences exist for the labels 'fragment' ($F(2)= 7.87$, $P < 0.001$), 'back-channel' ($F(2)= 6.01$, $P < 0.003$), 'elicit-suggestion' ($F(2)= 3.94$, $P < 0.022$) and 'suggestion' ($F(2)= 3.19$, $P < 0.045$).

For all of these some intuitive explanations can be given. Starting with the 'fragment' label, it appears that people who are highly influential utter less fragments than people who have low influence. This finding is in line with the finding from [2] who stated that people who are interrupted more than others are likely to be of a lower social status, and hence likely to be less influential. For the 'Back-channel' label it appeared that people who are 'Low' on dominance back-channel more than people who are 'high' on dominance. One could say that those that back-channel signal to others that they follow, or that they express listeners' behavior [12]. By providing back-channels people signal that they understand the messages produced by others. One could therefore say that it can be related to their *participation* level and hence to a lesser degree of talkativeness. This aligns with [1] who observed that people who talk more than others are likely to be more dominant. Both of these dialogue-act labels are related to the meeting process. The remaining two labels 'suggest' and the 'elicit-suggest' show that both types of utterances are uttered relatively more by people who are 'High' on dominance than by people who are 'Low' on dominance. Both the elicitation of suggestions, as well as making suggestions during a meeting, or a discussion, relate to the fact that people provide options, or ideas, that could be solutions to the problems, or issues at hand. This finding, hence seems to provide evidence for the hypothesis that dominance and argumentation are related.

The data was again transformed into a feature set for training some classifiers. For this experiment a data set was used containing 120 samples, out of which 25 were labelled 'High', 69 were labelled 'Normal' and 26 were labelled 'Low'. The results are shown in Table 6.

FeatureSet	J48	SVM	NB
All Dialogue-acts	56.66	58.33	45
Fragment and Suggest*	55.83	57.5	53.3

Table 6: Results on automatic influence level classification using the fraction of dialogue act labels as features. $*$ = best subset.

Given the majority class baseline of 57.5% it appears that, although some of the feature values differ significantly, the features themselves are unable to outperform the baseline. Also after applying a post-hoc feature analysis this turned out to be impossible[5].

In Section 4.1 TAS units were related to level of influence. Next we turn to the relation between TAS labels on the relations and influence.

4.3 TAS Relations and Influence

[5]Note that the optimal feature set contains the 'fragment' and 'suggest' labels which, given the significance levels and their complementarity in distinctiveness (see Figure 4), is a logical choice.

This section reports on attempts to relate the various relations that exist between nodes in the argument diagrams to the levels of influence. Similar to the previous sections, for each participant, for each meeting, the percentage of relation labels was sampled. The combined data resulted in a data-set of 59 participants, participating in 15 meetings (not in all meetings were discussions, nor did all participants participate in all discussions). 13 of the participants were labelled as 'High', 33 of them were labelled as 'Normal' and 13 of them were labelled as 'Low'.

An overview of the 95% confidence interval of the mean percentage of the six most frequently occurring relation labels is shown in Figure 5.

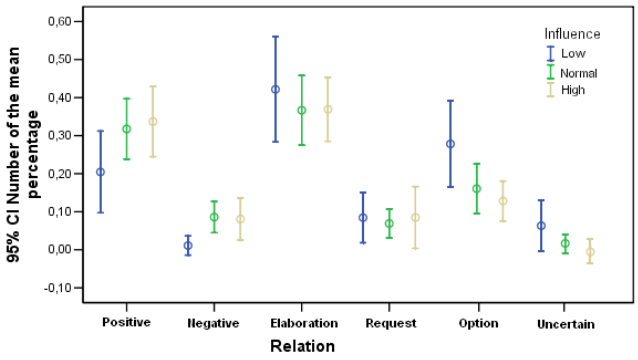

Figure 5: The mean number of relation occurrences per influence level.

ANOVA testing showed a significant dependency between the 'uncertain' relation category and the influence levels ($F(2)=3.52$, $p<0.037$). It appears that the lower the participant's influence, the more uncertain, or unclear, his or her contributions to the discussion are. Spearman's correlation coefficient ρ however did not prove significant. Here the relatively low number of samples is bothering us once more. For all the other relations we therefore cannot draw any hard conclusions.

When considering Figure 5 one could, however, construct the hypothesis that evaluative contributions, in terms of 'positive' and 'negative', seem to occur more frequently for higher influential participants. So if you give your opinion on things you might become more influential. But again, this is just a tendency that can be observed from the figure and this is not based on significant evidence.

Another interesting observation that can be made is that it seems that people of low influence seem to provide more 'options' to the discussions. These options suggest possible answers to issues that were raised. This is quite remarkable, because on the one hand this is perfectly in line with [10] who stated that the more dominant participants ask the questions. But on the other hand, it seems to contradict the statistically significant finding from the previous section saying that suggestions are mostly put forward by influential participants. However, one should note here that the dialogue acts that were considered go beyond the discussion boundaries and that the suggestion label, as is formulated in the annotation manual is applied in relation to "when the speaker expresses an intention relating to the *actions* of another individual, the group as a whole, or the group in a wider environment". This shows that the suggestion label covers more than suggestions for solutions to particular issues (options) and also that a bigger 'force' lies behind it in a sense that it really steers towards an action, rather than just raising an idea.

5. CROSS-FERTILIZING FEATURES

A typical question we want to answer from a machine learning point of view when considering Figure 5, deals with the extent to which the different distributions of certain (class) labels are useful for the classification process. Even more, since the previous section also showed that indeed some regularities seem to exist between the level of influence of a participant and the way that argumentation enfolds in a discussion. This section therefore aims to investigate the usefulness of (the features of) one of the phenomena of influence and argumentation as predictor, or feature, for the other phenomenon in a machine learning context.

One should note that both phenomena are higher level phenomena and that, in a real applications, it is not a clever choice to predict higher level concepts with other higher level concepts. This is true for at least two reasons. In the first place, the recognition process of the phenomenon that serves as a feature has to be recognized itself. This in turn requires more and other directly observable features. In the second place is it quite unlikely that the recognized higher level concepts are free of errors. The aim of this section is therefore just to investigate the extent to which one phenomenon theoretically could improve the recognition of the other and to explore the consequences of interrelation for the classification performance.

Theoretically, one could expect that whenever a certain feature f aids a classification task C, a classifier would be better able to distinguish the class labels and hence the recognition rate would increase. If this feature, however, represents a class label that itself can also be recognized by a different feature set $\{f_1, ..., f_n\}$ one could choose to replace f with the set of features that was devised to recognize f itself. This is interesting because one could expect that the recognition performance of C will be influenced by the fact that the function from $\{f_1, ..., f_n\}$ to f is not error-free, and hence it could be that the performance of C will be higher when using manually assigned values for f, rather than a whole set of automatically obtained features that are only to a certain extent able to represent f. However, as the feature set contains more than one feature, it could also perfectly well be that a certain feature of f (e.g. f_3) is more beneficial to C than f itself. For this experiment we confined ourselves to the manually assigned class labels that were elicited from the meeting participants and the manual annotations.

5.1 Predicting influence with argumentation

The first experiment tries to predict the influence level (dependent variable) making use of just the argumentation label distributions (independent variables).

As influence was measured on a meeting level, the feature vectors that contained the argumentation labels were also created on a meeting level by taking the label fraction distributions for the individual participants as feature values to predict the influence label of the associated participant.

This resulted in 59 samples[6] with a baseline of 55.93%. Machine learning algorithms were trained and evaluated using 10-fold cross validation. The results are shown in Table 7.

FeatureSet	J48	SVM	NB
STA-WST-OTH-OIS-AIS-YIS (unbal)	55.93	55.93	54.24
STA-WST-OTH-OIS-AIS-YIS (bal)	25.64	25.64	25.64
STA-OTH-ISS (unbalanced)	54.23	55.93	52.54

Table 7: Results on automatic influence level classification using the fraction of argument labels as features.

From Table 7 it appears that on the balanced corpus none of the tested classifiers outperforms the baseline. Not with the class labels added as feature, nor with the features that predict the class label, nor after merging the different issues and the different statements.

To explore this finding, a multiple linear regression model was instantiated from the data. Not surprisingly it appeared that none of the coefficients proved significant, nor for the individual labels, nor after merging the statements and the issues (the stronger the correlation coefficients, the more discriminating the feature).

5.2 Predicting argumentation with influence

For the second experiment the influence labels were used to see whether they could aid the prediction of TAS labels (both units and relations). So in this case the class labels were the TAS labels and the influence value of the speaker was added as a feature. The results are shown in Table 8.

Class	Feature set	J48	SVM	NB
Nodes	DOM	53.64	53.64	53.64
	QMT-ORT-L-LL-NS	73.53	68.21	64.05
	QMT-ORT-L-LL-NS-DOM	71.91	68.32	64.05
Relations	DOM	34.48	34.48	34.48
	TT	39.24	39.24	39.62
	TT-DOM	38.86	39.43	38.29
	TT-WT	44.95	39.80	44.00
	TT-WT-DOM	43.62	42.67	44.57

Table 8: Results on automatic TAS unit labelling with and without the dominance (DOM) feature. Features: TT = Target Type, WT = # Words in Target

The results indicate that the dominance feature does not seem to be of any use to the classifier. For the nodes of the TAS schema, the dominance feature itself does not score above the baseline of 55.95% (most frequent class is statement (484) amongst a total of 865 labels.). When adding the dominance feature to a set of more useful features, the performance does not increase either. For the relations of the TAS schema the baseline is set by the elaboration relation (181 occurrences amongst a total of 525 relations) to 34.4%. Again here the dominance feature does not prove useful, neither in combination with a set of other features that have proved useful in [9].

[6]13 were labelled as 'High', 33 as 'Normal', and 13 as 'Low'.

6. CONCLUSIONS

Given the results from the statistical investigations, the results on the classification performance and the rules that were induced, one could try to construct a tentative profile of how influential participants, as experienced by actual meeting participants, distinguish themselves from less influential participants. When considering the previous sections, one could say that:

- Influential participants seem to raise more issues.
- Influential participants leave the provision of options, or possible solutions, to others.
- Influential participants seem to provide more evaluative information with respect to the contributions of others.
- Influential participants seem to respond to statements from others that follow after Yes/No Issues.
- Influential participants significantly elicit and provide more suggestions for action over the course of a meeting.
- Influential participants significantly provide less backchannels over the course of a meeting.
- Influential participants seem to provide less 'other' TAS units.
- Influential participants provide fewer unfinished utterances, or speech fragments over the course of a meeting.
- Influential participants seem to resort later in a discussion to shorter turns.

So it seems that if a participant raises issues, elicits solutions, evaluates these solutions and then steers towards a choice amongst the possible solutions, one indeed gets an intuitive sense of a person who is highly influential, and who controls the course of discussion. On the other hand, if someone provides options, back-channels a lot to others, resorts to shorter contributions in the decision phase of the discussion indeed, then an intuitive profile of a less influential participant appears.

Exploitation of these profiles and the interrelation between both phenomena, however, do not prove to be sufficiently distinctive, in such a way that cross-fertilization of (features of) phenomena can yield machine learning algorithms to significantly improve their recognition performance. This result underlines that features have to correlate more than slightly with the phenomena of interest and also that 'just adding' features to the data set does not automatically improve the performance, in a sense that complementarity also plays a part.

7. ACKNOWLEDGEMENTS

This work is supported by the European IST Programme Project FP6-033812. This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein.

8. REFERENCES

[1] R. Bales. *Interaction Process Analysis.* Addison-Wesley, 1950.

[2] R. Bales, F. Strodbeck, T. Mills, and M. Roseborough. Channels of communication in small groups. *American Sociological Review*, 16:461–468, 1951.

[3] J. Carletta, S. Ashby, S. Bourban, and et al. The AMI meetings corpus. In *Proceedings of the Measuring Behavior 2005 symposium on "Annotating and measuring Meeting Behavior"*, 2005.

[4] P. Flach and N. Lachiche. Confirmation-guided discovery of first-order rules with tertius. *Machine Learning*, 1(42):61–95, 2002.

[5] W. Post, A. Cremers, and O. Henkemans. A research environment for meeting behavior. In *Proceedings of the third workshop on Social Intelligence Design (SID)*, 2004.

[6] R. Rienks, D. Zhang, D. Gatica-Perez, and W. Post. Detection and application of influence rankings in small group meetings. In *Proceedings of the Eighth International Conference on Multimodal Interfaces (ICMI06)*, 2006.

[7] F. van Eemeren, R. Grootendorst, and T. Kruiger. *Handbook of Argumentation Theory.* Foris publications, 1987.

[8] A. Verbree, R. Rienks, and D. Heylen. Dialogue-act tagging using smart feature selection: results on multiple corpora. In *The first International IEEE Workshop on Spoken Language Technology (SLT)*, Palm Beach, Aruba, Dec. 2006.

[9] A. Verbree, R. Rienks, and D. Heylen. First steps towards the automatic construction of argument-diagrams from real discussions. In *Proceedings of the in 1st International Conference on Computational Models of Argument*, 2006.

[10] J. Wang. Questions and the exercise of power. *Discourse and Society*, 17(4):529–548, 2006.

[11] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools with Java implementations.* Morgan Kaufmann, 2000.

[12] V. Yngve. On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, pages 567–577, 1970.