Dominance Detection in Meetings Using Easily Obtainable Features

Rutger Rienks and Dirk Heylen

Human Media Interaction (HMI),
University of Twente, Enschede, The Netherlands
{rienks, heylen}@ewi.utwente.nl
http://hmi.ewi.utwente.nl/

Abstract. We show that, using a Support Vector Machine classifier, it is possible to determine with a 75% success rate who dominated a particular meeting on the basis of a few basic features. We discuss the corpus we have used, the way we had people judge dominance and the features that were used.

1 Introduction

In many cases it is beneficial for the effectiveness of a meeting if people assume a cooperative stance. Grice [1975] formulated four maxims that hold for cooperative conversations. The maxims of quantity, quality, relevance and manner state that one should say nothing more or less than is required, speak the truth or say only things for which one has enough evidence, only say things that are relevant for the discussion at hand and formulate the contribution such that it can be easily heart and understood by the interlocutors. These maxims are all formulated from the perspective of producing utterances in a conversation. One could define similar maxims for cooperative behavior, more generally. One can also think of several tasks of chairpersons in meetings as being guided by such maxims. The chair should facilitate the participants to have their say, to cut off people who make their contribution too long or to intervene when contributions are not relevant to the discussion at hand. Discussions should be properly organized to have arguments develop, so that all positions are put to the fore, and all relevant pros and cons are raised. People that are too dominant in meetings may violate one or more of the cooperative maxims and thereby frustrate the process of collective decision making for which many meetings are intended. The chair of the meeting should avoid this from happening or intervene when it does.

Nowadays, in order to maximize the efficiency, meetings can be assisted with a variety of tools and supporting technologies [Rienks et al., 2005]. These tools can be passive objects such as microphones facilitating better understanding or semi-intelligent software systems that automatically adjust the lighting conditions. In the near future, meetings will be assisted with various similar sorts of active, and perhaps even autonomous, software agents that can make sense of what is happening in the meeting and make certain interventions [Ellis and Barthelmess,

2003]. An example of such meeting assisting agents could be an agent that signals possible violations of cooperative maxims in the decision making process to the chairperson. One of the major issues to be addressed in this case is how the agent can detect that there is such a disturbance. In the research described in the following sections we looked at a way to automatically detect the relative level of dominance of meeting participants on the basis of a set of simple features. We start with introducing the concept of dominance (Section 2). To establish a corpus, we asked several people to rank a collection of meetings. We investigated whether the rankings by different people were similar (Section 3). Next, we describe the features we used for our classifier (Section 4), how we obtained the feature values from our corpus (Section 5) and what the performance of our classifier was when using the best features (Section 6).

The only work that we are aware of which is in some sense comparable is described in Zhang et al. [2005] who created a two-level influence model. A dynamic Bayesian network (DBN) was proposed to learn the influence of each participant in meetings using both acoustic and language features. As 'ground truth' input for their model, a set of thirty meetings of about five minutes each was used together with the averaged results of three annotators.

2   Dominance

According to Hoffmann [1979], there are three types of behavioral roles that can be identified in groups or teams. These roles can be classified as task-oriented, relation-oriented and self-oriented. Each group member has the potential of performing all of these roles over time. *Initiators, Coordinators* and *Information Givers* are task-oriented roles that facilitate and coordinate the decision making tasks. The Relations-Oriented role of members deals with team-centered tasks, sentiments and viewpoints. Typical examples are : *Harmonizers, Gatekeepers* and *Followers.* The Self-Oriented role of members focusses on the members' individual needs, possibly at expense of the team or group. Examples here are *Blockers, Recognition Seekers* and *Dominators.* The Dominator is a group member trying to assert authority by manipulating the group or certain individuals in the group. Dominators may use flattery or proclaim their superior status to gain attention and interrupt contributions of others. According to Hellriegel et al. [1995], a group dominated by individuals who are performing self-oriented sub-roles is likely to be ineffective.

In psychology, dominance refers to a social control aspect of interaction. It involves the ability to influence others. One can refer to it as a personality characteristic - the predisposition to attempt to influence others - or one can use the term to describe relationships within a group. Dominance is a hypothetical construct that is not directly observable. However, there appear to be certain behavioral features displayed by people that behave dominantly that make it possible for observers of these behaviors to agree on judgments of dominance. In Ellyson and Dovidio [1985] the nonverbal behaviors that are typically associated

with dominance and power are investigated. In several of the papers in that volume, human perceptions of dominance are discussed as well. Behaviors such as proxemic relations, facial expressions and gaze were investigated.

In 'A System for Multiple Level Observation of Groups' (SYMLOG), [Bales and Cohen, 1979], Bales distinguishes three structural dimensions in group interactions: status, attraction and goal orientation. Goal orientation refers to the question whether people are involved with the task or rather with socio-emotional behaviours. This dimension was already present in Bales' earlier work on Interaction Process Analysis [Bales, 1951]. The attraction dimension refers to friendly versus unfriendly behaviours. The status dimension has to do with dominant versus submissive behaviours. Bales developed a checklist that observers can use to structure their observations of groups. He has also developed a number of self-report scales that group members can use to rate themselves (and other group members) on these three dimensions. SYMLOG presents a questionnaire containing 26 questions from which 18 relate to the concept of dominance. The factors involved in these questions provide a way to explicate the concept. An overview of these factors in their most general form are shown in Table 1.

Table 1. Aspects of dominance according to SYMLOG

Positive contributions	Negative contributions
active, dominant, talks a lot	passive, introverted, said little
extraverted, outgoing, positive	gentle, willing to accept responsibility
purposeful, democratic task-leader	obedient, worked submissively
assertive, business-like, manager	self-punishing, worked too hard
authority, controlling, critical	depressed, sad, resentful, rejecting
domineering, tough-minded, powerful	alienated, quit, withdrawn
provocative, egocentric, showed-off	afraid to try, doubts own ability
joked around, expressive, dramatic	quietly happy just to be in group
entertaining, sociable, smiled, warm	looked up to others, appreciative

When we look at these factors we see that most of them are very hard to operationalize. For example to automatically determine when someone is 'purposeful' or 'alienated' is quite complex and highly dependent on human interpretative skills. For the automatic classification task, we need easy to extract and automatically detectable features that can be quantified and transformed as a series of digits into our system.

To train a classifier that can determine who is the person that dominated a meeting, we need a corpus of meeting recordings with the relevant features that the classifier is using either extracted or annotated and also we need to know how the participants of the various meetings scored on the dimension of dominance. We will provide more details on the corpus and the features used by the classifier in Section 4. Now, we will first describe how we established the dominance ranking for the meetings we used.

3   Dominance Judgements

For our study, We used a corpus of eight four-person meetings[1]. The meetings varied in length between 5 and 35 minutes. We collected 95 minutes in total. We used different kinds of meetings, including group discussions where topics had to be debated, discussions about the design of a remote control, book club meetings and PhD. evaluation sessions.

We asked ten people to rank the participants of the meetings with respect to their perceived dominance. Each person ranked four, i.e. half of, the meetings. We thus had a total of five rankings for every meeting. We simply told people to rate the four people involved in the meeting on a dominance scale. We did not tell the judges anything more about what we meant by that term. The results are shown in Table 2. The first cell shows that in the first meeting (M1), judge A1 thought that the most dominant person was the one corresponding to the fourth position in this list, second was the first person in this list, third the second person in the list and least dominant was the third person in the list: 2,3,4,1. If one looks at the judgements by the other judges for this meeting (A2 to A5), by comparing the different columns for this first row, one can see that A3's judgments are identical to A1's. All but A4 agree that the fourth person on the list was most dominant. All but A5 agree that the third person was least dominant. All but A2 agree that the first person was the second dominant person. This seems to suggest that on the whole judgements were largely consistent across judges at first sight.

Table 2. Rating of meeting participants for all the annotators per meeting

	A1	A2	A3	A4	A5	'Average'	'Variance'
M1	2,3,4,1	3,2,4,1	2,3,4,1	2,1,4,3	2,4,3,1	2,3,4,1	8
M2	2,3,4,1	2,3,4,1	2,3,4,1	2,3,1,4	3,2,4,1	2,3,4,1	8
M3	2,1,3,4	3,1,2,4	2,1,4,3	3,1,2,4	1,2,3,4	2,1,3,4	8
M4	2,4,3,1	2,4,3,1	1,4,2,3	2,3,4,1	1,4,3,2	1,4,3,1	4
	A6	A7	A8	A9	A10	'Average'	'Variance'
M5	4,3,2,1	4,3,1,2	3,4,1,2	4,3,1,2	3,4,1,2	4,3,1,2	6
M6	1,3,2,4	1,4,3,2	3,1,4,2	3,1,4,2	1,3,4,2	1,3,4,2	12
M7	1,4,3,2	2,4,3,1	3,2,1,4	2,4,1,3	1,4,3,2	1,4,2,3	14
M8	1,2,4,3	1,4,2,3	2,1,3,4	2,1,3,4	1,2,4,3	1,2,3,4	12

To establish the degree of agreement, we compared the variance of the judgements with the variance of random rankings. If the variance of the annotators is smaller than the variance of the random rankings, we have a strong indication that people agree on how to create a dominance ranking.

---

[1] The first three meetings are meetings from for the AMI project (cf. http://www.amiproject.org), M1 and M2 are the AMI pilot meetings AMI-Pilot-2 and AMI-Pilot-4, M3 is a meeting from the AMI spokes corpus (AMI-FOB 6). The last five are meetings recorded for the M4 project (cf. http://www.m4project.org: M4TRN1, M4TRN2, M4TRN6, M4TRN7 and M4TRN12).

If we add up the dominance scores for each person in the meeting, this results for the first meeting in scores 11, 13, 19 and 7, with results in an overall ranking of 2, 3, 4, 1. We call this the 'average' ranking. In case of similar scores, we scored them an equal rank by giving them both the highest value. The next one highest in the ranking was ranked with a gap of two. Example: if the sum of the total scores ended up 8, 10, 12, 10 the resulting ranking became 1,2,4,2. For each of the judges we compare how they differ for each person from this average.

As a measure for the variance we calculated the sum of all the (absolute) differences of each of the annotators judgments (A^i) with their corresponding average. The difference with the average was calculated as the sum of the pairwise absolute differences for all the annotators values of the meeting participants A_p with their corresponding average value $Average_p$. See Table 2 for the results.

$$\text{`}Variance\text{'} = \sum_{i=1}^{5} \sum_{p=1}^{4} |A_p^i - Average_p|$$

In this case A1 and A3 judgments are identical to the average. A2 made different judgments for the first person (scoring him as 3 instead of 2) and the second person (scoring him as 2 instead of 3). So this results in a variance of 2 adding up the variance 4 and 2 of judges A4 and A5 respectively this ends up in an overall variance of 8 for judgements on the first meeting.

When comparing the variance of the judges with the variance resulting from randomly generated rankings, the distribution of the variance of the annotators ($\mu = 9$, $\sigma = 3.38$, $n = 8$) lies far more left of the distribution coming from randomly generated rankings. ($\mu = 17.8$, $\sigma = 3.49$, $n = 1.0 * 10^6$). The two distributions appeared to be statistically significant ($p < 0.001$) according to the 2-sided Kolmogorov Smirnov test. It thus appears that judges agree very well on dominance rankings. We may have to be conservative to generalize this though as we have only a small (n=8) amount of real samples.

These results support our initial thoughts, where we expected humans to agree (to a reasonable extent) on the ranking of meeting participants according to their conveyed dominance level.

4   Features Used by the Classifier

Dominance can be regarded as a higher level concept that might be deduced automatically from a subset of lower level observations [Reidsma et al., 2004], similar to the assignment of the value for dominance by humans on the basis of the perception and interpretation of certain behaviours.

For our classifier we considered some easily obtainable common sense features that possibly could tell us something about the dominance of a person in relation to other persons in meetings. We deliberately did not use semantically oriented features. For each person in the meeting we calculated scores for the following features.

- The person's influence diffusion (IDM)
- The speaking time in seconds (STS)

- The number of turns in a meeting (NOT)
- The number of times addressed (NTA)
- The number of successful interruptions (NSI)
- The number of times the person grabbed the floor (NOF)
- The number of questions asked (NQA)
- The number of times interrupted (NTI)
- The ratio of NSI/NTI (TIR)
- Normalised IDM by the amount of words spoken. (NIDF)
- The number of words spoken in the whole meeting (NOW)
- The number of times privately addressed (NPA)

The *Influence diffusion model* [Ohsawa et al., 2002] generates a ranking of the participants by counting the number of terms, reused by the next speaker from the current speaker. The person who's terms are re-used the most is called the most influential.

Most of the features appear as simple metrics with variations that measure the amount to which someone is involved in the conversation and how others allow him/her to be involved. These are all measures that are easy to calculate given a corpus with appropriate transcriptions and annotations provided. Metrics used in the literature, as in SYMLOG, depend on the interpretation of an observer.

We defined a *successful interruption* in line with Leffler et al. [1982]. We counted as a successful interruption any occurrence where a speaker A starts talking while another speaker B is talking and speaker B finishes his turn before speaker A. So we did not make a distinction between overlap and interruption. A *floorgrab* was defined each time a participant started speaking after a silence larger than 1.5 seconds.

After the judges that rated our corpus had finished their ratings, we asked them to write down a list of at least five aspects which they thought they had based their rankings on. The following features were mentioned.

> Dominant is the person: who speaks for the longest time, who speaks the most, who is addressed the most, who interrupts the others the most, who grabs the floor the most, who asks the most questions, who speaks the loudest, whose posture is dominant, who has the biggest impact on the discussion, who appears to be most certain of himself, who shows charisma, who seems most confident.

From the features identified by the annotators we can see that e.g. *charisma* and *confidence* are again typical examples of features that are very hard to measure and to operationalize. Most of this is again due to the fact that a proper scale does not exist, and as a result the valuation becomes too subjective and values from one annotator might not correlate with the values from another annotator. Several of the other features are similar to the ones we are exploring for their predictive power in our classifier.

5   Acquiring and Preprocessing the Data

For each of the eight meetings ranked by our annotators, we calculated the values for the measures identified in the previous section. This was done on the basis of simple calculations on manual annotations and on the results of some scripts processing the transcriptions[2]. With respect to addressee annotation 25% of the data was not annotated due to the cost involved[3].

In order to make the values for the same feature comparable, we first made the feature values relative with respect to the meeting length. This was done in two steps. First the fraction, or share, of a feature value was calculated given all the values for that feature in a meeting.

$$\text{The share of a feature value } (F'_{Pn}) = \frac{F_{Pn}}{\sum F_{P1}..F_{P4}}$$

Then, according to the value of the fraction, the results were binned in three different bins. As we are dealing with four person meetings the average value after step 1 is 0.25 (=25% share). The features were grouped using the labels 'High' ($F'_{Pn} > 35\%$), 'Normal' ($15\% < F'_{Pn} < 35\%$), and 'Low' ($F'_{Pn} < 15\%$).

As a consequence, apart from the fact that features were now comparable between meetings, the feature values that originally had 'approximately' the same value now also ended up in the same bin. This seemed intuitively the right thing to do. Table 3 shows the value of the NOW feature ('The number of words used' per participant per meeting) before and after applying the process. If we look at the number of words used for person 2 (P2) and person 4 (P4) we see that they both end up labelled as 'High'. Although they did not speak the same amount of words, they both used more than 90000 words, which is a lot in comparison with P1 (38914) and P3 (26310), both ending up classified as 'Low'.

Now, as the feature values were made comparable, we were almost ready to train our model. The only step left was to define the class labels determining the dominance level. For this we decided to use the same technique as for the features, labelling them also as 'High', 'Normal' and 'Low'. We calculated the shares of each of the participants by dividing the sum of the valuations of all judges for this participant by the total amount of points the judges could spend $(5 * (1 + 2 + 3 + 4) = 50)$.

The results were then again binned using the same borders of 15 and 35 percent. When a share was smaller than 15% the dominance level was labelled as 'High'; if the share lay between 15% and 35% the dominance level was labelled 'Normal' and when it was higher than 35 % the label 'Low' was used. This way, also the persons who received more or less similar scores ended up in the same bin.

---

[2] All transcriptions used were created using the official AMI and M4 transcription guidelines of those meetings [Moore et al., 2005, Edwards, 2001].
[3] Addressee information takes over 15 times real time to annotate [Jovanovic et al., 2005].

Table 3. The feature 'Number of Words' before and after preprocessing for person 1,2,3 and 4 respectively for each meeting

	NOW before preprocessing				NOW after preprocessing			
	P1	P2	P3	P4	P1	P2	P3	P4
M1	38914	93716	26310	98612	low	high	low	high
M2	33458	11602	14556	37986	high	low	low	high
M3	3496	7202	8732	2774	low	high	high	low
M4	2240	1956	4286	7642	low	low	normal	high
M5	4470	1126	9148	1974	normal	low	high	low
M6	2046	17476	1828	4058	low	high	low	high
M7	4296	6812	8258	1318	normal	high	high	low
M8	1586	13750	1786	1540	low	high	low	low

This resulted in a data-set of 32 samples with twelve samples receiving the class label 'High', ten 'Normal' and ten 'Low'. We define our baseline performance as the share of the most frequent class label ('High') having a share of 37.5% of all labels.

6   Detecting Dominance

We wanted to predict the dominance level of the meeting participants with the least possible features, in accordance with Occam's razor [Blumer et al., 1987], trying to explain as much as possible with as little as possible. The fewer features we required, the easier it would be to eventually provide all information to the system. This way we reduced the risk of over fitting our model to the data as well. To decrease the amount of features we evaluated the features on their discriminative force using WEKA's Support Vector Machine (SVM) attribute evaluator.

The top five of most discriminative features appeared to be (in order of importance) NOF, NOT, NSI, NOW, and NOQ. We obtained the best performance by training a SVM using the two most discriminative features: NOF and NOT. Ten-fold cross validation resulted in a performance of 75%, much higher than our 37.5% baseline. This means, that given the number of times the meeting participants grab the floor after a silence together with the number of turns a participant has, our classifier is in 75 % of the cases able to correctly classify the behavior of the participants as being 'Low', 'Normal' or 'High'on dominance. The confusion matrix is shown in Table 4.

From the confusion matrix it can be seen that our classifier performs better on the classes 'Low' and 'High' than on the class 'Normal'. This seems in line with our intuition that people showing more extreme behavior are easier to classify. To test the significance it appeared that the 90% confidence interval for our classifier lies between a performance of 62% and 88%, having a lower bound much higher than the 37.5% baseline. This confidence interval is important due

Table 4. Confusion matrix using the features NOF and NOT. The rows are showing the actual labels and the columns the labels resulting from the classifier.

|	Low	Normal	High
Low	8	1	1
Normal	2	7	1
High	0	3	9

to the relatively small number of data samples. The fact that we would over fit our classifier when using all the features appeared when we trained on all the features. Ten fold cross validation resulted in that case in a performance of 50%.

It is interesting to see that the number of successful interruptions as a feature on its own results in a performance of 59% which, although not significant, implies a correlation with the concept of dominance. This is in line with the claim of West and Zimmerman [1983] calling interruption 'a device for exercising power and control in conversation'. Tannen [1993] on the other hand claims that interruptions not necessarily need to be a display of dominance as people can interrupt each other to show enthusiastic listenership and participation as well.

7   Applying the Model

Aware of the fact that our sample size is relatively small and that not all meetings follow the same format, we do think that our results suggest that it is possible to have a system analyzing the level of dominance of the meeting participants. If we look at the features used by our model, and the fact that their values should be just as informative during the meeting as after the meeting, we expect these systems not to function just after the meeting, but just as well in real time.

We crafted a very simplistic model based on the top three features: NOF, NOT and NSI. The model grants one point for each turn a participant takes in a meeting and if the turn is acquired, either after a silence greater than 1.5 seconds, or by an interruption, another extra point is given. This model enabled
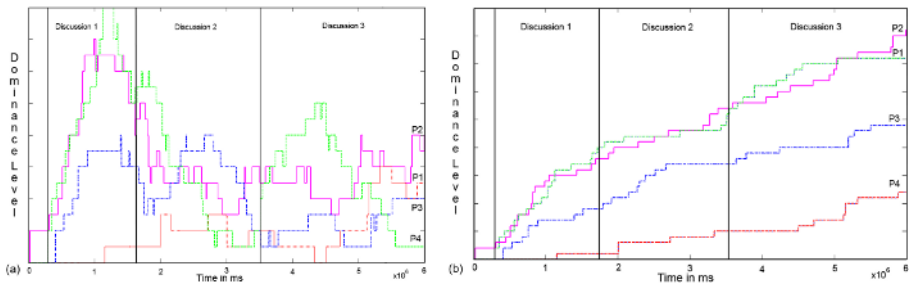
Fig. 1. Graphical outputs of the simplistic dominance model applied to M3 for a 100 seconds window (a) and for a window spanning the whole meeting (b)

us to produce figures similar to Figure 1, where we counted the points for all the participants of the AMI-FOB6 meeting in a time window of 100 seconds for each participant (a) and cumulatively counted the points for each participant over a whole meeting (b). It should be noted that the resulting heights of the participants levels correspond to the averaged annotator value of meeting M3 in Table 2.

8   Conclusions and Future Work

We have shown that in the future systems might be extended with modules able to determine the relative dominance level of individual meeting participants. We were able to reach an accuracy of 75% using just two easily obtainable features. The concept *dominance* appears in the meeting domain to be mainly reflected by the number of floorgrabs and the number of turns someone takes. As all the features are made relative to the total value of all participants, one is able to apply the model both during as well as after the meeting.

Possible directions for opportunities to improve our model could be to extend the feature set with more semantically oriented features, such as 'Who is using the strongest language?', or 'Who gets most suggestions accepted?'. Although these features seem very intuitive and might increase the performance, one does have to realize that being able to measure these, costly and complex inferencing systems have to be developed.

Another possible thing to look at is to use more samples, this will be more expensive on one side, but also decreases the confidence interval and thus further increase the reliability of the performance on the other side.

Typical applications of systems that track the dominance levels of participants are other systems using the dominance information in order to inform the meeting participants or a meeting chairman about this. With this information a chairman could alter his style of leadership in order to increase the meeting productivity. Combined with other information, recommender systems could be created that directly suggest how to change the leadership style. The next thing one could think of is a virtual chairman as mentioned in Rienks et al. [2005] which is able to lead a meeting all by itself, giving turns, keeping track of a time-line and most important: keeping the meeting as effective and efficient as possible.

Acknowledgements

References

R.F. Bales. *Interaction Process Analysis.* Addison-Wesley, 1951.

R.F. Bales and S.P. Cohen. *SYMLOG: A System for the Multiple Level Observation of Groups.* The Free Press, 1979.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Marmuthh. Occam's razor. In *Information Processing Letters*, pages 377–380. 24 edition, 1987.

J. A. Edwards. *Transcription in Discourse*, chapter in Handbook of Discourse, pages 321–348. Mass: Blackwell Publishers, 2001.

C.(S.) Ellis and P. Barthelmess. The Neem dream. In *Proceedings of the 2003 conference on Diversity in computing*, pages 23–29. ACM Press, 2003. ISBN 1-58113-790-7.

Ellyson and Dovidio. *Power, Dominance, an Nonverbal Behavior.* Springer Verlag, 1985.

H.P. Grice. *Logic and conversation*, chapter Syntax and Semantics: Speech Acts, pages 41–58. Academic Press, 1975.

D. Hellriegel, J.W. Slocum Jr., and R.W. Woodman. *Organizational Behavior, seventh edition.* West publising company, 1995.

L.R. Hoffmann. Applying experimental research on group problem solving to organizations. *Journal of applied behavioral science*, 15:375–391, 1979.

N. Jovanovic, R. Op den Akker, and A. Nijholt. A corpus for studying addressing behavior in multi-party dialogues. In *Proc. of The sixth SigDial conference on Discourse and Dialogue*, 2005.

A. Leffler, D.L. Gillespie, and C. Conaty. The effects of status differentiation on nonverbal behaviour. *Social Psychology Quarterly*, 45(3):151–161., 1982.

J. Moore, M. Kronenthal, and S. Ashby. Guidelines for AMI speech transcriptions. Technical report, IDIAP, Univ. of Edinburgh, February 2005.

Y. Ohsawa, N. Matsumura, and M. Ishizuka. Influence diffusion model in text-based communication. In *Proc. of The eleventh world wide web conference*, 2002. ISBN 1-880672-20-0.

D. Reidsma, R. Rienks, and N. Jovanovic. Meeting modelling in the context of multimodal research. In *Proc. of the Workshop on Machine Learning and Multimodal Interaction*, 2004.

R. Rienks, A. Nijholt, and D. Reidsma. *Meetings and Meeting support in ambient intelligence*, chapter in Ambient Intelligence, Wireless Networking, Ubiquitous Computing. Artech House, Norwood, MA, USA, 2005. In Press.

D. Tannen. *Interpreting Interruption in Conversation*, chapter in Gender and Discourse, pages 53–83. Oxford University Press, 1993.

C. West and D.H. Zimmerman. *Small Insults: A study of interruptions in cross-sex conversations between unaquainted persons*, chapter in Language, Gender and Society, pages 103–117. Newbury House, 1983.

D. Zhang, D. Gatica-Perez, S. Bengio, and D. Roy. Learning influence among interacting markov chains. *Advances in Neural Information Processing Systems (NIPS)*, 18, 2005.