Computing Backchannel Distributions in Multi-Party Conversations

Dirk Heylen
Human Media Interaction
University of Twente
heylen@cs.utwente.nl

Rieks op den Akker
Human Media Interaction
University of Twente
infrieks@cs.utwente.nl

Abstract

In multi-party conversations it may not always be obvious who is talking to whom. Backchannels may provide a partial answer to this question, possibly in combination with some other events, such as gaze behaviors of the interlocutors. We look at some patterns in multi-party interaction relating features of backchannel behaviours to aspects of the partipation framework.

1 Introduction

In this paper we present a summary of our investigations into the distribution of back-channels and some other forms of feedback and assesments in argumentative multi-party discourse. We are interested in such expressions for several reasons. First, the sheer utterance of a backchannel indicates the presence of an auditor that indicates "I am here, I am attending". The fact that it is being uttered by an auditor indicates intrinsically that the auditor *felt addressed in some way or another* by the speaker. For the analysis of multi-party conversations, it is important to establish *who is talking to whom* and backchannels, at least seem to give away the *whom* part. Second, the exact form, the kind of vocalisation, the intonation and the context may further invest the utterance with additional meanings, expressing various attitudes towards what has been said: skepticism, surprise, liking, agreement, and so on. So, when we look at back-channels in the context of multi-party dialogues they may tell us something about the participation framework on the one hand (who was talking to whom) and about the way utterances are being assessed by their audience.

The qualifier "in some way or another" with respect to feeling or being addressed is particularly important in the context of multi-party dialogues (i.e. dialogues with more than two persons present). Typically, an utterance by a speaker instantiates the performance of a speech act with a particular illocutionary and perlocutionary force. The speech act involves a request for uptake. However, as has been pointed out several times (Goffman (Goffman, 1981), Levinson (Levinson, 1988), Clark and Carlson (Clark and Carlson, 1992), Schegloff (Schegloff, 1988)), participants in a multi-party conversation can have a different role or status and they can be addressed in different ways.

In this paper we report on some of our investigations into the distribution of backchannels in multiparty interactions (for instance in relation to other phenomena such as gaze) and how this information can help us to uncover certain features of floor and stance taking automatically.

We will first describe the corpus and the annotations. Next we look at the annotations of utterances consisting of starting with "yeah" and try to see whether we can classify these utterances as continuers, i.e. neutral with respect to stance taking (Schegloff, 1981), or as assessments.

2 Corpus

The argumentative discourses that we are studying are part of the meeting corpus collected during the AMI project (McCowan et al., 2005). From a computational, technological perspective, the aims

Proceedings of the Workshop on Embodied Language Processing, pages 17–24,
Prague, Czech Republic, June 28, 2007.

of this research is directed at developing automatic procedures that can help to provide answers to any query users may have about what goes on in the meetings. The AMI corpus consists of meetings in which a group of four people discuss the design of a new remote control. T

The kinds of queries that we would like our procedures to be able to answer are related to these moves: what suggestions have been made; what were the arguments given and how much animosity was there related to the decision. In the AMI corpus, the meeting recordings have been annotated on many levels, allowing the use of machine learning techniques to develop appropriate algorithms for answering such questions. We focus on the dialogue act annotation scheme. This contains three types of information. Information on the speech act, the relation between speech acts and information on addressing.

The dialogue act classes that are distinguished in our dialogue act annotation schema fall into the following classes:

- Classes for things that are not really dialogue acts at all, but are present to account for something in the transcription that doesn't really convey a speaker intention. This includes backchannels, stalls and fragments

- Classes for acts that are about information exchange: inform and elicit inform.

- Classes for acts about some action that an individual or group might take: suggest, offer, elicit suggest or offer.

- Classes for acts that are about commenting on the previous discussion: assess, comment about understanding, elicit assessment, elicit comment about understanding

- Classes for acts whose primary purpose is to smooth the social functioning of the group: be-positive, be-negative.

- A "bucket" type, OTHER, for acts that do convey a speaker intention, but where the intention doesn't fit any of the other classes.

For our studies into feedback in the AMI corpus, the dialogue acts labelled as backchannesl are clearly important. They were defined in the annotation manual as follows.

In backchannels, someone who has just been listening to a speaker says something in the background, without really stopping that speaker. [...] Some typical backchannels are "uhhuh", "mm-hmm", "yeah", "yep", "ok", "ah", "huh", "hmm", "mm" and, for the Scottish speakers in the data recorded in Edinburgh, "aye". Backchannels can also repeat or paraphrase part or all of what the main speaker has just said.

The labels *assess* and *comment-about-understanding* are closely related. They were defined as follows.

An ASSESS is any comment that expresses an evaluation, however tentative or incomplete, of something that the group is discussing. [...] There are many different kinds of assessment; they include, among other things, accepting an offer, expressing agreement/disagreement or any opinion about some information that's been given, expressing uncertainty as to whether a suggestion is a good idea or not, evaluating actions by members of the group, such as drawings. [...] An ASSESS can be very short, like "yeah" and "ok". It is important not to confuse this type of act with the class BACKCHANNEL, where the speaker is merely expressing, in the background, that they are following the conversation.

C-A-U is for the very specific case of commenting on a previous dialogue act where the speaker indicates something about whether they heard or understood what a previous speaker has said, without doing anything more substantive. In a C-A-U, the speaker can indicate either that they did understand (or simply hear) what a previous speaker said, or that they didn't.

The Backchannel class largely conforms to Yngve's notion of backchannel and is used for the functions of contact (Yngve, 1970). Assess is used for the attitudinal reactions, where the speaker expresses his stance towards what is said, either acceptance or rejection. Comments about understanding are used for explicit signals of understanding or non-understanding.

In addition to dialogue acts also relation between dialogue acts are annotated. Relations are annotated between two dialogue acts (a later source act and an earlier target act) or between a dialogue act (the source of the relation) and some other action, in which case the target is not specified. Relations are a more general concept than adjacency pairs, like question-answer. Relation have one of four types: positive, negative, partial and uncertain, indicating that the source expresses a positive, negative, partially positive or uncertain stance of the speaker towards the contents of the target of the related pair. For example: a "yes"-answer to a question is an inform act that is the source of a positive relation with the question act, which is the target of the relation. A dialogue act that assesses some action that is not a dialogue act, will be coded as the source of a relation that has no (dialogue act as) target.

A part of the scenario-based meetings (14 meetings) were annotated with addressee labels, i.e. annotators had to say who the speaker is talking to. The addressee tag is attached to the dialogue act. If a speaker changes his addressee (for instance, from group to a particular participant) during a turn the utterance should be split into two dialogue act segments, even if the type of dialogue act is the same for both segments.

3 Yeah

In this section we look at the distribution of *yeah* in the AMI corpus. "yeah" utterances make up a substantial part of the dialogue acts in the AMI meeting conversations (about 8%). If we try to tell group addressed dialogue acts from individually addressed acts then "yeah" is the best cue phrase for the class of single addressed dialogue acts; cf. (Stehouwer, 2006).

In order to get information about the stance that participants take with respect towards the issue discussed it is important to be able to tell utterances of "yeah" as a mere backchannel, or a stall, from yeah-utterances that express agreement with the opinion of the speaker. The latter will more often be classified as assessments. We first look at the way annotators used and confused the labels and then turn to see in what way we can predict the assignments to the class.

3.1 Annotations of yeah utterances

One important feature of the dialogue act annotation scheme is that the annotators had to decide what they consider to be the segments that constitute a dialogue act. Annotators differ in the way they segment the transcribed speech of a speaker. Where one annotator splits "Yeah. Maybe pear yeah or something like that." into two segments labeling "yeah." as a backchannel and the rest as a suggest, an other may not split it and consider the whole utterance as a suggest.

In comparing how different annotators labeled "yeah" occurrences, we compared the labels they assigned to the segment that starts with the occurrence of "yeah".

The confusion matrix for 2 annotators of 213 yeah-utterances, i.e. utterances that start with "yeah", is given below. It shows that backchannel (38%), assess (37%) and inform (11%) are the largest categories [1]. Each of the annotators has about 80 items in the backchannel class. In about 75% of the cases, annotators agree on the back-channel label. In either of the other cases a category deemed a backchannel is mostly categorized as assessment by the other and vice versa. For the assessments, annotators agree on about slightly more than half of the cases (43 out of 79 and 43 out of 76). The disagreements are, for both annotators split between the backchannels, for the larger part, the inform category, as second largest, and the **other** category.

The **other** category subsumes the following types of dialogue acts: summing up for both annotators: be-positive(9), suggest(8), elicit-assess(3), elicit-inform(2), comment-about-understanding(2). The dialogue act type of these **other** labeled utterances is mostly motivated by the utterances following "Yeah". Examples: "Yeah , it's a bit difficult" is labeled as Be-positive. "Yeah ? Was it a nice way to create your remote control ?" is labeled as an Elicit-Assessment .

Out of the 213 Yeah-utterances a number contains just "yeah" without a continuation. Below, the confusion matrix for the same two annotators, but now for only those cases that have text "yeah" only. In

[1] As the numbers for each of the classes by both annotators is about the same, we have permitted ourselves the license to this sloppy way of presenting the percentages.

yeah	0	1	2	3	4	SUM
0	59.0	2.0	17.0	0.0	2.0	80.0
1	0.0	9.0	4.0	2.0	2.0	17.0
2	21.0	3.0	43.0	7.0	5.0	79.0
3	2.0	0.0	7.0	13.0	4.0	26.0
4	1.0	0.0	5.0	0.0	5.0	11.0
SUM	83.0	14.0	76.0	22.0	18.0	213.0

Figure 1: Confusion matrix of two annotations of all Yeah utterances. labels: 0 = backchannel; 1 = fragment or stall; 2 = assess; 3 = inform; 4 = other. p0=0.61 (percentage agreement); kappa=0.44.

yeah-only	0	1	2	SUM
0	50.0	12.0	3.0	65.0
1	13.0	5.0	1.0	19.0
2	2.0	0.0	2.0	4.0
SUM	65.0	17.0	6.0	88.0

Figure 2: labels: 0 = bc 1 = assess 2 = other (subsuming: be-positive, fragment, comment-about-understanding). p0=0.65; kappa=0.14

the comparison only those segments were taken into account that both annotators marked as a segment i.e. a dialogue act realized by the word "Yeah" only.[2]

What do these patterns in the interpretation of "yeah" expressions tell us about its semantics? It appears that there is a significant collection of occurrences that annotators agree on as being backchannels. For the classes of assessments and other there also seem to be prototypical examples that are clear for both annotators. The confusions show that there is a class of expressions that are either interpreted as backchannel or assess and a class whose expressions are interpreted as either assessments or some other label. Annotators often disagree in segmentation. A segment of speech that only consist of the word "yeah" is considered to be either a backchannel or an assess, with very few exceptions. There is more confusion between annotators than agreement about the potential assess acts.

[2]The text segment covered by the dialogue act then contains "Yeah", "Yeah ?", "Yeah ," or "Yeah .".

3.2 Predicting the class of a yeah utterance

We derived a decision rule model for the assignment of a dialogue act label to yeah utterances, based on annotated meeting data. For our exploration we used decision tree classifiers as they have the advantage over other classifiers that the rules can be interpreted.

The data we used consisted of 1122 yeah utterances from 15 meetings. Because of the relative low inter-annotator agreement, we took meetings that were all annotated by one and the same annotator, because we expect that it will find better rules for classifying the utterances when the data is not too noisy.

There are 12786 dialogue act segments in the corpus. The number of segments that start with "yeah" is 1122, of which 861 are short utterances only containing the word "yeah". Of the 1122 yeahs 493 dialogue acts were annotated as related to a previous dialogue act. 319 out of the 861 short yeah utterances are related to a previous act.

The distribution of the 1122 yeah utterances over dialogue act classes is: assess (407), stall (224), backchannel (348), inform (95) and other (48 of which 25 comment-about-understanding). These are the class variables we used in the classification. The model consists of five features. We make use of the notion of *conversational state*, being an ensemble of the speech activities of all participants. Since we have four participants a state is a 4-tuple $< a, b, c, d >$ where a is the dialogue act performed by participant A, etc. A conversation is in a particular state as long as no participant stops or starts speaking. Thus, a state change occurs every time when some participants starts speaking or stops speaking, in the sense that the dialogue act that he performs has finished. The features that we use are:

- *lex* This feature has value 0 if the utterance consists of the word Yeah only. Otherwise 1.
- *continue* Has value 1 when the producer of the utterance also speaks in the next conversational state. Otherwise 0. This feature models incipient behavior of the backchanneler.
- *samespeaker* Has value 1 if the conversational state in which this utterance happens has the

Null	629.0
Assess	81.0
Inform	162.0
Elicit-Comment-Understanding	2.0
Elicit-Assessment	40.0
Elicit-Inform	73.0
Elicit-Offer-Or-Suggestion	2.0
Suggest	114.0
Comment-About-Understanding	13.0
Offer	5.0
Be-Positive	1.0

Figure 3: Distribution of the types of dialogue acts that yeah utterances are responses to.

same speaker, but different from the backchanneler, as the next state. Otherwise 0. This feature indicates that there is another speaker that continues speaking.

- *overlap* There is speaker overlap in the state where the utterance started.
- *source* This involves the relation labeling of the annotation scheme. *source* refers to the dialogue act type of the source of the relation of the dialogue act that is realized by the Yeah utterance. If the yeah dialogue act is not related to some other act the value of this feature is null. The possible values for this feature are: null, assess, inform, suggest, elicitation (which covers all elicitations), and other.

The distribution of source types of the 1122 yeah dialogue acts is shown in table 3.2. The table shows that 629 out of 1122 yeah utterances were not related to some other act.

We first show the decision tree computed by the J48-tree classifier as implemented in the weka-toolkit, if we do not use the source feature looks as follows. The tree shows that 392 utterances satisfy the properties: continued = 1 and short = 1. Of these 158 are misclassified as backchannel.

1. Continued ≤ 0
   (a) lex ≤ 0: bc(392.0/158.0)
   (b) lex > 0: as(56.0/24.0)
2. Continue $\rangle$ 0
   (a) samespkr ≤ 0
      i. overlap ≤ 0: st(105.0/27.0)
      ii. overlap > 0
         A. lex ≤ 0: st(76.0/30.0)
         B. lex > 0: bc(16.0/6.0)
   (b) samespkr > 0 : ass(477.0/233.0)

In this case the J48 decision tree classifier has an accuracy of 57%. If we decide that every yeah utterance is a Backchannel, the most frequent class in our data, we would have an accuracy of 31%. If we include the source feature, so we know the type of dialogue act that the yeah utterance is a response to, the accuracy of the J48 classifier raises at 80%. Figure 3.2 shows the decision tree for this classifier. The results were obtained using ten-fold cross-validation.

It is clear from these results that there is a strong relation between the source type of a Yeah dialogue act and the way this Yeah dialogue act should be classified: as a backchannel or as an assess. Note that since backchannels are never marked as target of a relation, **null** as source value is a good indicator for the Yeah act to be a backchannel or a stall.

We also tested the decision tree classifier on a test set that consists of 4453 dialogue acts of which 539 are yeah-utterances (219 marked as related to some source act). Of these 219 are short utterances consisting only of the word “Yeah” (139 marked as related). The utterances in this test set were annotated by other annotators than the annotator that annotated the training set. The J48 classifier had an accuracy on the test set of 64%. The classes which are confused most are those that are also confused most by the human annotators: backchannels and stall, and assess and inform. One cause of the performance drop is that in the test corpus the distribution of class labels differs substantially from that of the training set. In the test set yeah utterances were very rarely labelled as stall, whereas this was a frequent label (about 20%) in the training set. The distribution of yeah-utterance labels in the test set is: backchannels 241, stalls 4, assessments 186, inform 66 and other 42.

When we merged the train and test meetings and trained the J48 decision tree classifier, a 10 fold cross-validation test showed an accuracy of 75%. Classes that are confused most are again: backchannel and stall, and assessment and inform.

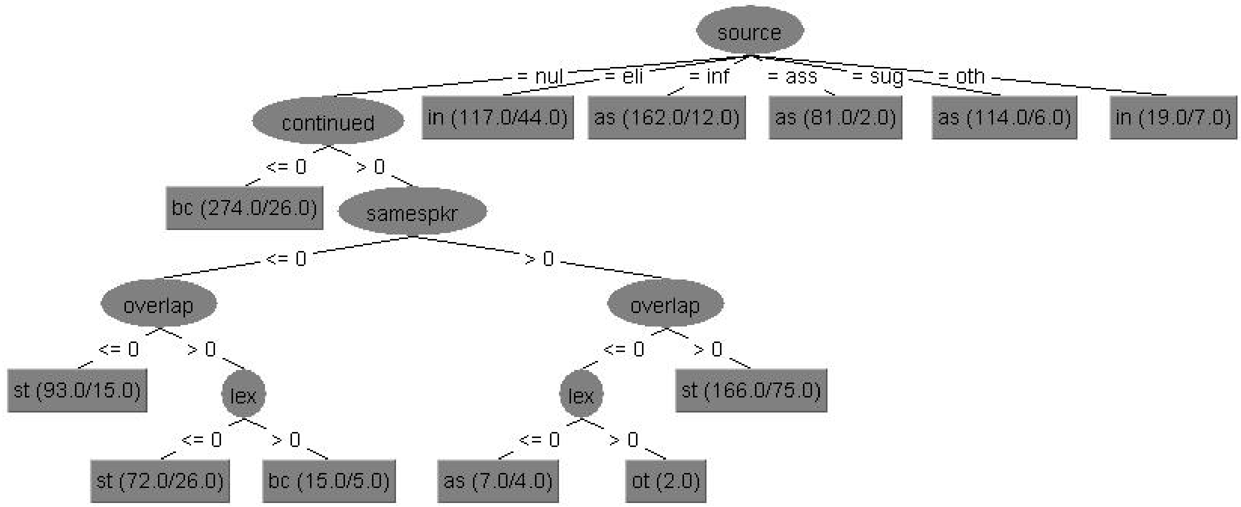

Figure 4: Decision tree for classification of yeah utterances when information about the source of the related dialogue act is used.

4 Measuring Speaker Gaze at Backchannelors

When thinking about the interaction between speaker and backchannelor, it seems obvious, as we said before, that the person backchanneling feels addressed by the speaker. We were wondering whether the backchannel was not prompted by an invitation of a speaker, for example, by gazing at the listener.

Gaze behavior of speaker and backchannelor is classified by means of the following gaze targets, a sequence of focus of attention labels that indicates where the actor is looking at during a period of time:

1. the gaze targets of the *speaker* in the period starting some short time ($DeltaTime$) before the start time of the backchannel act till the start of the backchannel act.
2. the gaze targets of the *backchannelor* in the period starting some short time ($DeltaTime$) before the start time of the backchannel act till the start of the backchannel act.
3. the gaze targets of the speaker during the backchannel act.
4. the gaze targets of the backchannelor during the backchannel act.

We set $DeltaTime$ at 1 sec, so we observed the gaze behavior of the speaker in the period from one second before the start of the backchannel act. Using these gaze target sequences, we classified the gaze behavior of the actor as follows:

0: the gaze before target sequence of the actor does not contain any person

1: the before gaze target sequence of the actor does contain a person but not the other actor involved: for the speaker this means that he did not look at backchannelor before the backchannel act started, for the backchannelor this means that he did not look at the speaker before the start of the backchannel.

2: the actor did look at the other person involved before that backchannel act.

Figure 4 show a table with counts of these classes of events. In the 13 meetings we counted 1085 backchannel events. There were 687 events with a single speaker of a real dialogue act. For this cases it is clear who the backchannelor was reacting on. This is the selected speaker. The table shows speaker data in rows and backchannel data in columns. The $MaxDownTime$ is $1sec$ and the $MinUpTime$ is 2 sec. The $DeltaTime$ for the gaze period is $1sec$. From the table we can infer that:

1. The selected speaker looks at the backchannelor in the period before the backchannelor act starts in 316 out of the 687 cases.

2. The backchannelor looks at the selected speaker in the period before the backchannelor act starts in 430 out of the 687 cases.

3. The selected speaker looks at someone else than the backchannelor in the period before the backchannelor act starts in 209 out of the 687 cases.

4. The backchannelor looks at someone else than the selected speaker in the period before the backchannelor act starts in 54 out of the 687 cases.

5. In 254 out of the 687 cases the speaker looked at the backchannelor and the backchannelor looked at the speaker.

We may conclude that the speakers look more at the backchannelor than at the other two persons together (316 against 209). The table also shows that backchannelors look far more at the selected speaker than at the two others (430 against 54 instances).

In order to compare gaze of speaker in backchannel events, we also computed for each of the 13 meetings for each pair of participants (X, Y): $dagaze(X, Y)$: how long X looks at Y in those time frames that X is performing a dialogue act.

$$dagaze(X, Y) = \frac{\sum OT(gaze(X, Y), da(X))}{\sum da(X)} \tag{1}$$

where summation is over all real dialogue acts performed by X,
$OT(gaze(X, Y), da(X))$ is the overlap time of the two events: $gaze(X, Y)$: the time that X gazes at Y, and $da(X)$ the time that the dialogue act performed by X lasts. The numbers are normalized over the total duration of the dialogue acts during which gaze behavior was measured.

Next we computed $bcgaze(X, Y)$: how long X looks at Y in those time frames that X performs a real dialogue act and the Y responds with a backchannel act.

$$bcgaze(X, Y) = \frac{\sum OT(gaze(X, Y), dabc(X, Y))}{\sum da(X, Y)} \tag{2}$$

where $dabc(X, Y)$ is the time that X performs the dialogue act that Y reacts on by a backchannel. Here normalization is with the sum of the lengths of all dialogue acts performed by X that elicited a backchannel act by Y.

Analysis of pairs of values $gaze(X, Y)$ and $bcgaze(X, Y)$ shows that in a situation where someone performs a backchannel the speaker looks significantly more at the backchannelor than the speaker looks at the same person in general when the speaker is performing a dialogue act ($t = 8.66$, $df = 101$, $p < 0.0001$). The mean values are 0.33

$sp\|bc$	0	1	2	T
0	103	4	55	162
1	46	42	121	209
2	54	8	254	316
T	203	54	430	687

Figure 5: Gaze table of speaker and backchannelor. $DeltaTime = 1sec$. Total number of backchannel events is 1085. In the table only those 687 backchannel events with a single speaker are considered (excluded are those instances where no speaker or more than one speaker was performing a real dialogue act in the period with a $MinUpTime$ of 2 sec and a $MaxDownTime$ of 1 sec.). Speaker data in rows; backchannelor data in columns. The table shows for example that in 121 cases the speaker looked at someone but not the backchannelor, in the period from 1 sec before the start of the backchannel act till the start of the backchannel act, while the backchannelor looked in that period at the speaker.

and 0.16.[3]

Perhaps we can use the information on gaze of the participants in the short period before the backchannel act as features for predicting who the backchannel actor is. For the 687 data points of backchannel events with a single speaker, we used gaze of participants, the speaker and the duration of the backchannel act as features. Using a decision tree classifier we obtained an accuracy of 51% in predicting who will perform a backchannel act (given that someone will do that). Note that there are three possible actors (the speaker is given). This score is 16% above the a priori likelihood of the most likely participant: A (36%).

Conclusion

In this paper, we have explored some questions about the possible use and function of backchannels in multiparty interactions. On the one hand backchannels can be informative about functions related to floor and participation: who is talking to whom. Obviously, a person producing a backchannel was responding to an utterance of speaker. For the semantic analysis of meeting data an important question is whether he was just using the backchannel as a continuer (a sign of attention) or as an assessment. We also checked our intuition that backchannels in the kinds of meetings that we are looking at might often be invited by speakers through gaze. Obviously, these investigations just scratch the service of how backchannels work in conversations and how we can use them to uncover information from recorded conversations.

References

H. H. Clark and T. B. Carlson. 1992. Hearers and speech acts. In Herbert H. Clark, editor, *Arenas of Language Use*, pages 205–247. University of Chicago Press and CSLI.

Erving Goffman. 1981. Footing. In Erving Goffman, editor, *Forms of Talk*, pages 124–159. University of Pennsylvania Press, Philadelphia, PA.

Stephen C. Levinson. 1988. Putting linguistics on a proper footing: explorations in goffman's concept of participation. In Paul Drew and Anthony Wootton, editors, *Erving Goffman. Exploring the Interaction Order*, pages 161–227. Polity Press, Cambridge.

I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M.Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner. 2005. The ami meeting corpus. In *Measuring Behaviour, Proceedings of 5th International Conference on Methods and Techniques in Behavioral Research*.

Emanuel A. Schegloff. 1981. Discourse as an interactional achievement: some uses of 'uh huh' and other things that come between sentences. In Deborah Tannen, editor, *Analyzing Discourse: Text and Talk*, pages 71–93. Georgetown University Press, Washington.

Emanuel A. Schegloff. 1988. Goffman and the analysis of conversation. In Paul Drew and Anthony Wootton, editors, *Erving Goffman. Exploring the Interaction Order*, pages 89–135. Polity Press, Cambridge.

J.H. Stehouwer. 2006. Cue-phrase selection methods for textual classification problems. Technical report, M.Sc. Thesis, Twente University, Human Media Interaction, Enschede, the Netherlands.

V.H. Yngve. 1970. On getting a word in edgewise. In *Papers from the sixth regional meeting of the Chicago Linguistic Society*, pages 567–77, Chicago: Chicago Linguistic Society.

[3]For 13 meeting and 4 participants we would have 156 pairs of values. We only used those 102 pairs of which both values are non-zero.