Are You Being Addressed? - real-time addressee detection to support remote participants in hybrid meetings

Harm op den Akker
Roessingh Research and Development
Enschede
the Netherlands
h.opdenakker@rrd.nl

Rieks op den Akker
Human Media Interaction Twente
Enschede
the Netherlands
infrieks@cs.utwente.nl

Abstract

In this paper, we describe the development of a meeting assistant agent that helps remote meeting participants by notifying them when they are being addressed. We present experiments that have been conducted to develop machine classifiers to decide whether "you are being addressed" where "you" refers to a fixed (remote) participant in a meeting. The experimental results back up the choices made regarding the selection of data, features, and classification methods. We discuss variations of the addressee classification problem that have been considered in the literature and how suitable they are for addressee detection in a system that plays a role in a live meeting.

1 Introduction

In order to understand what is going on in a meeting, it is important to know who is talking, what is being said, and who is being addressed (talked to). Here, we focus on the question of whom the speech is addressed to. We present results obtained in developing a classifier for real-time addressee prediction to be used in an assistant for a remote participant in a *hybrid* meeting, a meeting where a number of participants share a common meeting room and one or more others take part via teleconferencing software.

It is obvious that in order to effectively participate in a meeting, participants need to know who is being addressed at all times. For remote participants in hybrid meetings, understanding the course of the conversation can be difficult due to the fact that it is hard to figure out who is being addressed. But it is not only meeting participants who are interested in addressees. The question who is being addressed has long been of interest for science: group therapists (Bales, 1950), small group research, or *outside observers* who analyse recorded meetings.

How speakers address listeners, what kind of procedures speakers use to designate their audience and to make clear whom they address has been the focus of conversational analysis, sociolinguistics and ethnomethodology for quite some time. An analysis of addressee selection is presented in (Lerner, 1996). Addressing as a special type of multi-modal interactional referring expression generation behavior is considered in (op den Akker and Theune, 2008).

The problem of *automatic addressee detection* is one of the problems that come up when technology makes the move from *two-party* man-machine natural dialogue systems to systems for *multi-party* conversations. In this context the addressing problem was raised by Traum (2004).

Since Jovanović (2004), presented her research on addressee prediction in meetings at SigDial, quite a few publications on the topic appeared. Jovanović used a number of multi-modal meeting corpora developed in the European projects M4 and AMI. In (Jovanović et al., 2006b) the first multi-modal multi-party corpus containing hand labeled addressee annotations was presented. The public release of the multi-modal AMI meeting corpus (Carletta, 2007; McCowan et al., 2005), a 100 hour annotated corpus of small group meetings has already shown to be an important achievement for research; not only for conversational speech recognition and tracking of visual elements

but also for automatic multi-modal conversational scene analysis. The M4 and AMI corpora are the only multi-modal meeting corpora (partly) annotated with addressee labels. Addressee detection in robot-human interaction is studied in (Katzenmaier et al., 2004) and in multi-party dialogue systems in (Knott and Vlugter, 2008; van Turnhout et al., 2005; Bakx et al., 2003; Rickel et al., 2002). Addressing in face-to-face conversations is achieved by multi-modal behavior and addressee detection is thus a multi-modal recognition task. This task requires not only speech recognition but also gaze and gesture recognition, the recognition of deictic references, and, ideally, the understanding of the "what's going on" in the meeting. It requires the detection of who is involved in current (parallel) activities. Speakers show explicit addressing behavior when they are not confident that the participants they want to address are paying attention to their words. Analysis of the remote meetings recorded in the EC project AMIDA reinforces our experiences that this happens more in remote meetings than in small group face-to-face meetings.

In AMIDA, the European follow-up project of AMI, the two new research goals are: (1) real-time processing (real-time speech recognition (Hain et al., 2008), focus of attention recognition (Ba and Odobez, 2009), real-time dialogue act labeling (Germesin et al., 2008) and addressee detection); and (2) technology for (remote) meeting support. Technology based on the analysis of how people behave and converse in meetings is now going to re-shape the meetings, and hopefully make them more effective and more engaging. Social interaction graphs that show who is talking to whom and how frequently in a meeting may help the group by mirroring its interpersonal relations, dominance, and group dynamics, and understand social mechanisms as possible causes of ineffectiveness. Although, feedback about the social interactions may also be useful *during* meetings, it doesn't require the prediction of the speaker's addressees in real-time. A participant in a meeting, however, needs to know who is being addressed by the speaker *at "the time of speaking"*. This holds for humans as well as for an artificial partner, a robot or a virtual Embodied Conversational Agent in a multi-party conversation.

The problem of addressee prediction comes in different flavors, depending on the relations that the subject who is in need of an answer, has with the event itself. *Time* is one of the aspects that play a role here: whether the subject needs to know the addressee of an utterance in real-time or offline. But it is not only time that plays a role. The addressing problem is an *interactional problem*, meaning that it is determined by the role that the subject has in the interaction itself; if and how the speaker and others communicate with each other and with the subject. Is he himself a possible addressee of the speaker or is he an outside observer? What type of communication channels are available to the subject and which channels of communication are available to the conversational partners in the meeting? It is often harder to follow a face-to-face discussion on the radio than to follow a radio broadcasted multi-party discussion that was held via a point-to-point telephone connection.

What speakers do to make clear whom they are addressing depends on the status and capacities of the communication lines with their interlocutors. Discussion leaders in TV shows are aware of their TV audience. Every now and then, they explicitly address their *virtual* audience at home. They also design their questions so as to make clear to the TV viewer whom their questions are addressed to. Outside observers in the form of a video camera will, however, not affect the way speakers make clear whom they address as long as the camera is not considered as a participant interested in the speaker's intention. Because remote participants are often out of sight, speakers in the meeting room do not take them into account when they converse to others in the meeting room. Remote participants become a kind of outside observers and share the same problems that annotators have when they watch video recordings of meetings to see what is happening in the meeting and who is being addressed by the speaker.

In section 2 we will specify the particular type of addressing problem that we are trying to tackle here. We make clear how our problem and approach differ from those of other researchers and what this means for the applicability of previous results and available data. In section 3 we present the data we used for testing and training. We set a baseline for the performance of our classifiers as

well as a hypothesized maximum value, or ceiling, based on the complexity of the task at hand. In section 4 we discuss the experiments, for selecting the optimal features, classifiers, and parameters. In section 5 we present the experimental results. In section 6 we discuss how the currently implemented addressing module works in the meeting assistant and what is required to use all the features of the addressee predictor in a hybrid meeting.

2 The Addressing Problem Considered Here

Jovanović et al. (2004) and Jovanović et al. (2006a) describe the classifiers that have been trained and tested on the M4 and AMI corpora. The classification problem is to assign an addressee label to a dialogue act, a hand-labeled and hand-segmented sequence of words, which is obtained by manual transcription of a speaker's utterance. The output of the classifier is one of a set of possible addressee labels: Group, or P0,P1,P2,P3, which are the four fixed positions around the table of the four participants in the meeting. Since the AMI data contains several meetings of different groups of four people, the class value cannot be the name of a participant, as that is not an invariant of the meeting setting. Positions at the rectangular table are invariant. This implies that the classifiers can only be used for meetings with this setting and four participants. A comparison of the statistical classifier of Jovanović with a rule-based method using the same part of the AMI corpus is presented in (op den Akker and Traum, 2009). The same data is also used by Gupta et al. (2007) in their study of a related problem: finding the person the speaker refers to when he uses a second person pronoun (e.g. 'you' or 'your') as a deictic referring expression. Their class values are not positions at the table but "virtual positions" in the speaking order (e.g. next speaker, previous speaker), a solution that generalises to a broader class of conversations than four participants in a face-to-face meeting. In a more recent study, Frampton et al. (2009) use positions at the table relative to the position of the speaker as class values: L1, L2, L3. The reason for this is to alleviate the problem of class imbalance in the corpus.

We will also use the AMI corpus but we will look at a different variant of the addressing problem. This is motivated by our application: to support a remote participant in a hybrid meeting. The question that we will try to answer is "are you being addressed?", where "you" refers to an individual participant in a conversation. The possible answers we consider are "yes" or "no"[1]. The addressing classifier that solves this problem is thus dedicated to a personal buddy. Note that this makes the method useable for any type of conversational setting. Note also that the addressing prediction problem "are you being addressed?" for a meeting assistant who is not himself participating in the meeting is different from the problem "am I being addressed?" that a participant himself may have to solve. The meeting assistant does not have direct "internal" knowledge about the processes or attentiveness of his buddy participant; he has to rely on outside observations. Our view on the problem implies that we have to take another look at the AMI data and that we will analyse and use it in a different way for training, testing and performance measuring. It also implies that we cannot rely for our binary classification problem on the results of Jovanović (2007) with (dynamic) Bayesian networks.

3 The Data and How Complex Our Task Is

We use a subset of the AMI corpus, containing those fourteen meetings that have not only been annotated with dialogue acts, but where dialogue acts are also attributed an addressee label, telling if the speaker addresses the Group, or the person sitting at position P0,P1,P2 or P3[2]. They have also been annotated with visual focus of attention: at any time it is known for each partner where he is looking and during what time frame. Annotated gaze targets are persons in the meeting, whiteboard, laptop, table or some other object.

Another level of annotations that we use concerns the topic being discussed during a topic segment of the meeting. Participants in the AMI corpus play a role following a scenario, the group has to design a remote TV control and team members each have one of four roles in the design project: PM - project manager; UI - user interface designer; ID - industrial designer; or ME - marketing

[1] A 'yes' means that the dialogue act is addressed to 'you' only. Group-addressed dialogue acts are considered to be 'no' (not addressed to you only).

[2] Annotators could also use label *Unknown* in case they could not decide the addressee of the speaker, this is treated as Group-addressed or 'no'.

expert. In training and testing the classifiers we alternately take up the position in the meeting of one of the participants, who is treated as the target for addressee prediction.

3.1 Base-line and Ceiling-value

Because most of the dialogue acts are not specifically addressed to one and the same meeting participant, the baseline for the binary classification task is already quite high: 89.20%, being the percentage of all dialogue acts annotated with addressing information "*not* addressed to You", which is 5962 out of a total of 6648 dialogue acts.

The performance of a supervised machine learning method depends on (1) the selection of features (2) the type of classifier including the settings of the hyper-parameters of the classifiers (Daelemans et al., 2003), and (3) the quality and the amount of training data (Reidsma, 2008; Reidsma and Carletta, 2008). Since we measure the classifier's performance with a part of the annotated data it is interesting to see how human annotators (or, 'human classifiers') perform on this task.

One of the AMI meetings[3] has been annotated with addressing information by four different annotators. We will use this to measure how ambiguous the task of addressee labeling is. Table 1 shows the confusion matrix for two annotators: *s95* and *vka*. This shows the (dis-)agreements for labelling the 412 dialogue acts as addressed to A, B, C, D or to the Group. [4] However, because we use our data differently, we will look at the confusion matrices in a different way. We split it up into 4 matrices, each from the view of one of the four meeting participants. Table 2 is an example of this, taking the view of participant A (i.e. for the binary decision task "is **Participant A** being addressed?", and having annotator s95 as gold standard.

Table 2 shows that when taking annotator *s95* as gold standard, and considering annotator *vka* as the classifier, he achieves an accuracy of 92.23 (380 out of 412 instances classified correctly).

[3] IS1003d

[4] Note that the annotators first independently segmented the speaker's turns into dialogue act segments; then labeled them with a dialogue act type label and then labeled the dialogue acts with an addressee label. The 412 dialogues acts are those segments that both annotators identified as a dialogue act segment.

	A	B	C	D	Group	Total
A	**29**				10	39
B		**14**			8	22
C			**32**		7	39
D	1		1	**49**	18	69
Group	21	10	19	22	**171**	243
Total	51	24	52	71	214	412

Table 1: Confusion matrix for one pair of annotators ($\kappa = 0.55$).

	A	$\neg A$	Total
A	**29**	10	39
$\neg A$	22	**351**	373
Total	51	361	412

Table 2: Confusion matrix for one pair of annotators, considering addressed to A or not (derived from the matrix in Table 1).

We can argue that we can use these human annotators/classifiers scores as a measure of "maximum performance", because it indicates a level of task ambiguity. Classifiers can achieve higher scores, because they can learn through noise in the data. Thus, the inter-annotator confusion value is not an absolute limit of actual performance, but cases in which the classifier is "right" and the test-set "wrong" would not be reflected in the results. Since the inter-annotator confusion does also say something about the inherent task ambiguity, it can be used as a measure to compare a classifier score with. Table 3 contains the overall scores (taken over all 4 individual participants) for the 6 annotator pairs. The average values for Recall, Precision, F-Measure and Accuracy in Table 3 are considered as *ceiling* values for the performance measures for this binary classification task[5]. The Hypothesized Maximum Score (HMS) is the average accuracy value: 92.47.

Pair	Rec	Prec	F	Acc
s-v	73.37	62.63	67.58	92.78
m-s	59.75	70.59	64.72	91.87
m-v	69.92	74.78	72.27	93.11
m-d	37.77	81.61	51.64	91.79
v-d	42.04	80.49	55.23	92.22
s-d	43.68	77.55	55.88	93.02
Average:	**54.42**	**74.61**	**61.22**	**92.47**

Table 3: Recall, Precision, F-measure and Accuracy values for the 6 pairs of annotators.

[5] Inter-changing the roles of the two annotators, i.e. consider vka as "gold standard" in Table 2, means inter-changing the Recall and Precision values. The F-value remains the same, though.

The baseline (89.20 for all dialogue acts annotated with addressing) and the HMS (92.47) accuracy values will be used for comparison with the performance of our classifiers.

4 The Methods and Their Features

In the experiments, four different classifiers were created:

1. Lexical and Context Classifier

2. Visual Focus of Attention Classifier

3. Combined Classifier

4. Topic and Role Extended Classifier

For each of these classifiers a large number of experiments were performed with a varying number of 15 to 30 different machine learning methods -using Weka (Witten and Frank, 1999)- to select optimal feature sets. In this section we summarize the most important findings. For a more detailed analysis refer to (op den Akker, 2009). Because of the large number of features and classifiers used, the various classifier hyper parameters have largely been kept to their default values. Where it was deemed critical (Neural Network training epochs and number of trees in RandomForest classifier) these parameters were varied afterwards to make sure that the performance did not deviate too much from using the default values. It didn't.

4.1 Lexical and Context Classifier

The lexical and context based classifier uses features that can be derived from words and dialogue acts only. A total of 14 features were defined, 7 of which say something about the dialogue act (type, number of words, contains 1st person singular personal pronoun, and so on) and 7 of which say something about the context of the dialogue act (how often was I addressed in the previous 6 dialogue acts, how often did I speak in the previous 5 dialogue acts, and so on). Of these 14 features, the optimal feature subset was selected by trying out all the subsets. This was repeated using 15 different classifiers from the WEKA toolkit. The best result was achieved with a subset of 10 features, by the MultiLayerPerceptron classifier. In this way an accuracy of 90.93 was reached. Given the baseline of the used train and test set of 89.20 and the HMS of 92.47, this can be seen as 53% of what 'can' be achieved.

4.2 Visual Focus of Attention Classifier

The VFOA classifier uses features derived from a meeting participant's visual focus of attention. A total of 8 features were defined, such as: the total time that the speaker looks at me, the total time everyone is looking at me, and so on. The optimal time interval in which to measure who is looking at you was extensively researched by trying out different intervals around the start of a dialogue act, and training and testing a classifier on the feature. These optimal interval values differ for every feature, but is usually somewhere between a few seconds before the start of the dialogue act, to 1 second into the dialogue act. The difference in performance for using the optimal interval compared to using the start- and end times of the dialogue act is sometimes as much as 0.93 accuracy (which is a lot given a base score of 89.20 and HMS of 92.47). This shows, that when looking at VFOA information, one should take into account the participant's gaze before the dialogue act, instead of looking at the utterance duration as in (Jovanović, 2007; Frampton et al., 2009)[6]. The representation of feature values was also varied by either normalizing to the duration of the window or using the raw values. Again the optimal feature subset was calculated using brute-force. Because of the reduced time complexity for 2^8 possible feature subsets, 30 different classifiers from the WEKA toolkit were trained and tested. One of the best results was achieved with a feature set of 4 features again with the MultiLayerPerceptron: 90.80 accuracy. The train and test sets used for this classifier are slightly smaller than those used for the LexCont classifier because not all dialogue acts are annotated with VFOA. The base score for the data here is 89.24, and given the HMS of 92.47, this result can be seen as 48% of what can be achieved.

4.3 Combined Classifier

The third classifier is a combination of the first two. We tried three different methods of combining the results of the LexCont and VFOA classifiers. First we tried to train a classifier using all the features (14 lexical, 8 vfoa) which exploded the feature subset search space to over 4 million possibilities. A second approach was to combine the output of the LexCont and VFOA classifiers using a simple rule-based approach. The OR-rule

[6]Note that a dialogue act segment can be preceded by an other utterance unit of the same speaker.

(if either of the two classifiers thinks the DA is addressed to you, the outcome is 'yes') performed the best (91.19% accuracy). But the best results were achieved by training a rule based (Ridor) classifier on the output of the first two. For these experiments the test-set of the previous two classifiers was split again into a new train (3080 instances) and test set (1540 instances). The features are the outputs of the VFOA and LexCont classifiers (both class and class-probabilities). For this task, 35 classifiers have been trained with the best results coming from the Ridor classifier: 92.53 accuracy. The results of all the different techniques for combining the classifiers can be seen in Table 4. The baseline score for this smaller test set is 89.87, so given the HMS of 92.47, this result can be seen as 102% of what can be achieved. Note that this is not 'impossible', because the Hypothesized Maximum Score is merely an indication of how humans perform on the task, not an absolute ceiling.

4.4 Topic and Role Extended Classifier

As a final attempt to improve the results we used topic and role information as features to our combined classifier. In the AMI corpus, every meeting participant has a certain role (project manager, interface designer, etc...) and the meetings were segmented into broad topic (opening, discussion, industrial designer presentation). Now the idea is that participants with certain roles are more likely to be addressed during certain topics. As an illustration of how much these a-priori chances of being addressed can change, take the example of an industrial designer during an 'industrial designer presentation'. The a-priori probability of you being addressed as industrial designer in the entire corpus is 13%. This probability, given also the fact that the current topic is 'industrial designer presentation' becomes 46%. This is a huge difference, and this information can be exploited. For all combinations of topic and role, the a-priori probability of you being addressed as having that role and during that topic, have been calculated. These values have been added as features to the features used in the Combined Classifier, and the experiments have been repeated. This time, the best performing classifier is Logistic Model Trees with an accuracy of 92.99%. Given the baseline of 89.87 and HMS of 92.47, this can be seen as 120% of what 'can' be achieved, which is better by a fairly large margin than the results of the inter-annotator agreement values.

5 Summary of Results

Table 4 summarizes the results for the various classifiers. The LexCont and VFOA classifiers individually achieve only about 50% of what can be achieved, but if combined in a clever way, their performance seems to reach the limit of what is possible based on the comparison with inter-annotator agreement. The fact that the topic-role extended classifier achieves so much more than 100% can be ascribed to the fact that it is cheating. It uses pre-calculated a-priori chances of 'you' being addressed given the circumstances. This knowledge could be calculated by the machine learner by feeding it the topic and role features, and letting it learn these a-priori probabilities for itself. But the classifier that uses these types of features can not easily be deployed in any different setting, where participants have different roles and where different topics are being discussed.

Method	Acc	Rec	Prec	F	PoM
HMS	92.47	54.42	74.61	61.22	-
LexCont	90.93	33.10	66.02	44.09	53
VFoA	90.80	27.77	67.65	39.38	48
CombinedFeat	91.56	36.62	70.82	48.28	72
ClassOfResults	43.68	77.55	55.88	93.02	102
LogComb(AND)	90.24	9.86	94.23	17.85	31
LogComb(OR)	91.19	47.08	61.90	53.48	60
TopicRoleExt	92.99	41.03	80.00	54.24	120

Table 4: Performance values of the Methods discussed in this paper: Accuracy, Recall, Precision, F-measure and Percentage of Hypothesized Maximum Score (PoM).

6 How Does The Assistant Work?

At the time of writing, the assistant that has been implemented is based on the simple visual focus of attention classifier. The focus of attention is inferred from the head pose and head movements of a participant in the meeting room who is being observed by a close-up camera. The real-time focus of attention module sends the coordinates of the head pose to a central database 15 times per second (Ba and Odobez, 2009). The coordinates are translated into targets: objects and persons in the meeting room. For the addressing module most important are the persons and in particular the screen in the meeting room where the remote

participant is visible. The addressing module is notified of updates of who is speaking and decides whether the remote participant is being looked at by the speaker.

If the remote participant (RP) is not attentive (which can be detected automatically based on his recent activity) he is called when he is addressed or when the real-time keyword spotter has detected a word or phrase that occurs on the list of topics of interest to the RP. For a detailed description of the remote meeting assistant demonstrator developed in the AMIDA project refer to (op den Akker et al., 2009).

The meeting assistant allows the RP to distribute his attention over various tasks. The system can give a transcript of the fragment of the meeting that is of interest to the RP, so he can catch up with the meeting if he was not following. The simple focus of attention based addressing module works fine. The question is now if an addressing module that uses the output of the real-time dialogue act recognizer, which in turn uses the output of the real-time speech recognizer will outperform the visual focus of attention based addressee detector. Experiments make us rather pessimistic about this: the performance drop of state of the art real-time dialogue segmentation and labeling technology based on real-time ASR output is too large in comparison with those based on hand-annotated transcripts (Jovanović, 2007). For real-time automatic addressee detection more superficial features need to be used, such as: speech/non-speech, who is speaking, some prosodic information and visual focus of attention, by means of head orientation.

The most explicit way of addressing is by using a vocative, the proper name of the addressed person. In small group face-to-face meetings, where people constantly pay attention and keep track of others' attentiveness to what is being said and done, this method of addressing hardly ever occurs. In remote meetings where it is often not clear to the speaker if others are paying attention, people call other's names when they are addressing them. Other properties of the participant relevant for addressee detection include his role and his topics of interest. These can either be obtained directly from the participant when he subscribes for the meeting, or they can be recognized during an introduction round that most business meetings start with. For automatic topic detection further analysis of the meeting will be needed (see (Purver et al., 2007)). Probability tables for the conditional probabilities of the chance that someone with a given role is being addressed when the talk is about a given topic, can be obtained from previous data, and could be updated on the fly during the meeting. Only when that has been achieved will it be possible for our extended topic/role addressee classifier to be fully exploited by a live meeting assistant.

Acknowledgements

The research of the first author was performed when he was a Master's student at the Human Media Interaction group of the University of Twente. This work is supported by the European IST Programme Project FP6-0033812 (AMIDA). We are gratefull to the reviewers of SigDial 2009 for their encouraging comments, and to Lynn Packwood for correcting our English.

References

Sileye Ba and Jean-Marc Odobez. 2009. Recognizing human visual focus of attention from head pose in meetings. In *IEEE Transaction on Systems, Man, and Cybernetics, Part B (Trans. SMC-B)*, volume 39, pages 16–33.

I. Bakx, K. van Turnhout, and J. Terken. 2003. Facial orientation during multi-party interaction with information kiosks. In *Proceedings of 9th IFIP TC13 International Conference on Human-Computer Interaction (INTERACT)*, Zurich, Switzerland.

Robert Freed Bales. 1950. *Interaction Process Analysis; A Method for the Study of Small Groups*. Addison Wesley, Reading, Mass.

Jean C. Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation*, 41(2):181–190, May.

Walter Daelemans, Véronique Hoste, Fien De Meulder, and Bart Naudts. 2003. Combined optimization of feature selection and algorithm parameter interaction in machine learning of language. In *Proceedings of the 14th European Conference on Machine Learning (ECML-2003)*, Lecture Notes in Computer Science 2837, pages 84–95, Cavtat-Dubrovnik, Croatia. Springer-Verlag.

Matthew Frampton, Raquel Fernndez, Patrick Ehlen, Mario Christoudias, Trevor Darrell, and Stanley Peters. 2009. Who is you? combining linguistic and gaze features to resolve second-person references in

dialogue. In *Proceedings of the 12th Conference of the EACL*.

Sebastian Germesin, Tilman Becker, and Peter Poller. 2008. Determining latency for on-line dialog act classification. In *Poster Session for the 5th International Workshop on Machine Learning for Multimodal Interaction*, volume 5237.

Surabhi Gupta, John Niekrasz, Matthew Purver, and Daniel Jurafsky. 2007. Resolving "you" in multiparty dialog. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, Belgium, September.

Thomas Hain, Asmaa El Hannani, Stuart N. Wrigley, and Vincent Wan. 2008. Automatic speech recognition for scientific purposes - webasr. In *Proceedings of the international conference on spoken language processing (Interspeech 2008)*.

Natasa Jovanović and Rieks op den Akker. 2004. Towards automatic addressee identification in multiparty dialogues. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 89–92, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Natasa Jovanović, Rieks op den Akker, and Anton Nijholt. 2006a. Addressee identification in face-to-face meetings. In *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Trento, Italy.

Natasa Jovanović, Rieks op den Akker, and Anton Nijholt. 2006b. A corpus for studying addressing behaviour in multi-party dialogues. *Language Resources and Evaluation Journal*, 40(1):5–23.

Natasa Jovanović. 2007. *To whom it may concern: adressee identification in face-to-face meetings*. Ph.D. thesis, University of Twente.

M. Katzenmaier, R. Stiefelhagen, and T. Schultz. 2004. Identifying the addressee in human-human-robot interactions based on head pose and speech. In *Proceedings of International Conference on Multimodal Interfaces (ICMI)*, pages 144–151, State College, PA.

A. Knott and P. Vlugter. 2008. Multi-agent human-machine dialogue: issues in dialogue management and referring expression semantics. *Artificial Intelligence*, 172:69–102.

Gene H. Lerner. 1996. On the place of linguistic resources in the organization of talk-in interaction: "Second person" reference in multi-party conversation. *Pragmatics*, 6(3):281–294.

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 *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*.

Rieks op den Akker and Mariet Theune. 2008. How do I address you? - modelling addressing behavior based on an analysis of a multi-modal corpus of conversational discourse. In *Proceedings of the AISB 2008 Symposium on Multimodal Output Generation (MOG 2008), Aberdeen, UK*, pages 10–17.

Rieks op den Akker and David Traum. 2009. A comparison of addressee detection methods for multiparty conversations. In *Proceedings of DiaHolmia, 13th Workshop on the Semantics and Pragmatics of Dialogue*.

Rieks op den Akker, Dennis Hofs, Hendri Hondorp, Harm op den Akker, Job Zwiers, and Anton Nijholt. 2009. Engagement and floor control in hybrid meetings. In *Proceedings COST Action Prague 2008 (to appear)*, LNCS. Springer Verlag.

Harm op den Akker. 2009. On addressee detection for remote hybrid meeting settings. Master's thesis, University of Twente.

Matthew Purver, John Dowding, John Niekrasz, Patrick Ehlen, Sharareh Noorbaloochi, and Stanley Peters. 2007. Detecting and summarizing action items in multi-party dialogue. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, Belgium, September.

Dennis Reidsma and Jean C. Carletta. 2008. Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326, September.

Dennis Reidsma. 2008. *Annotations and Subjective Machines*. Ph.D. thesis, University of Twente.

J. Rickel, S. Marsella, J. Gratch, R. Hill, D. Traum, and W. Swartout. 2002. Towards a new generation of virtual humans for interactive experiences. *Intelligent Systems*, 17:32–36.

David Traum. 2004. Issues in multiparty dialogues. In *Advances in Agent Communication*, pages 201–211.

K. van Turnhout, J. Terken, I. Bakx, and B. Eggen. 2005. Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features. In *Proceedings of 7th International Conference on Multimodal Interfaces (ICMI'05)*, Trento, Italy.

Ian H. Witten and Eibe Frank. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1st edition, October.