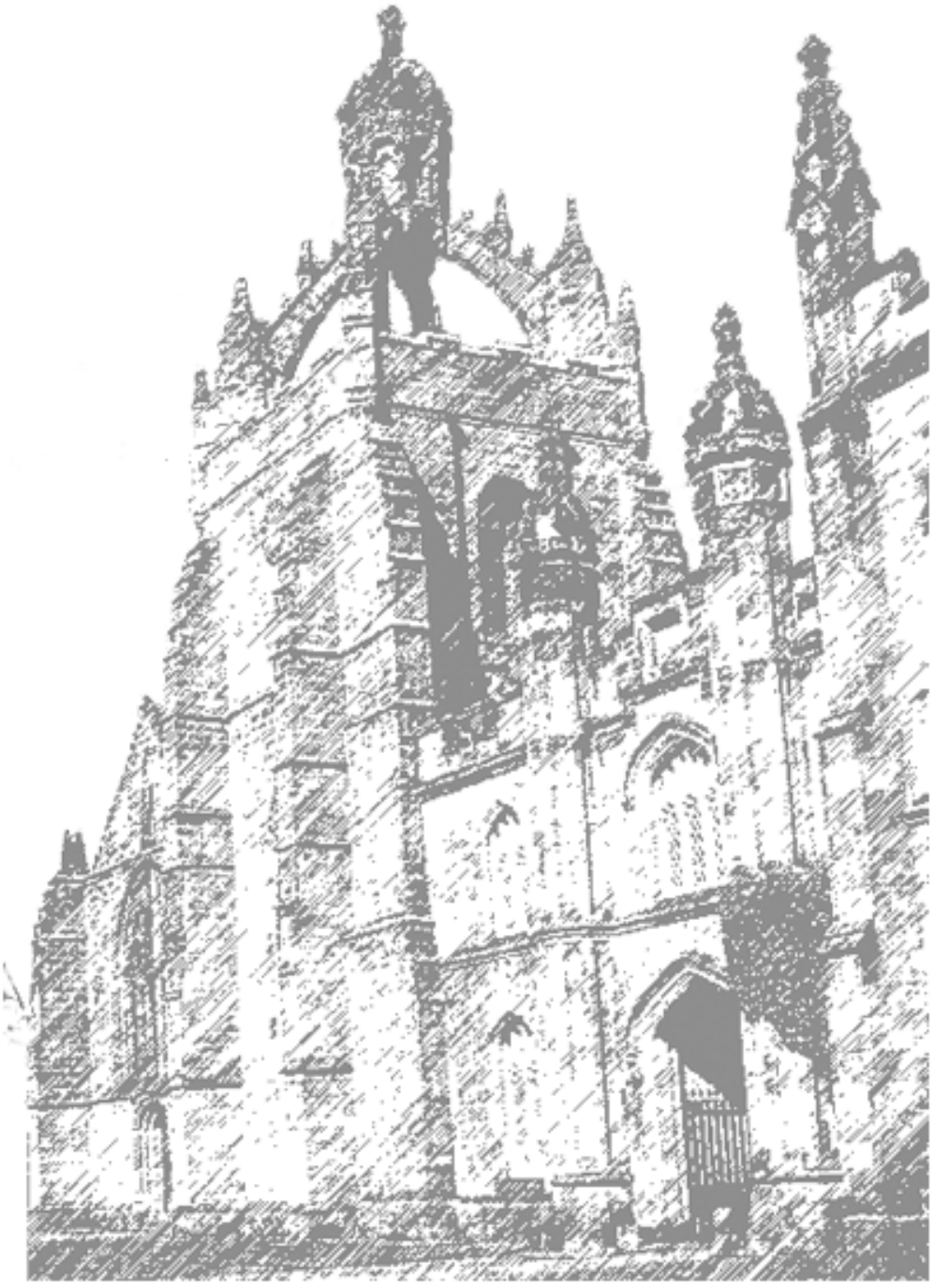
MOG 2008
Multimodal Output Generation

AISB 2008 Proceedings Volume 10

AISB '08



GLoRiClass

AISB ABERDEEN - 2008

UNIVERSITY OF ABERDEEN

AISB 2008 Convention
Communication, Interaction and Social Intelligence

1st-4th April 2008

University of Aberdeen

Volume 10:

Proceedings of the
AISB 2008 Symposium on Multimodal Output Generation (MOG 2008)

Published by
The Society for the Study of Artificial Intelligence and Simulation of Behaviour

http://www.aisb.org.uk/convention/aisb08/

ISBN 1 902956 69 9

How Do I Address You? Modelling addressing behaviour based on an analysis of multi-modal corpora of conversational discourse

Rieks op den Akker and **Mariët Theune** [1]

Abstract. Addressing is a special kind of *referring* and thus principles of multi-modal referring expression generation will also be basic for generation of address terms and addressing gestures for conversational agents. Addressing is a *special* kind of referring because of the different (second person instead of object) role that the referent has in the interaction. Based on an analysis of addressing behaviour in multi-party face-to-face conversations (meetings, TV discussions as well as theater plays), we present outlines of a model for generating multi-modal verbal and non-verbal addressing behaviour for agents in multi-party interactions.

1 INTRODUCTION

According to Gomperz [12], Hymes coined the notion of *communicative competence* to suggest that linguists, concerned as they are with communication in human groups, need to go beyond mere description of language usage patterns [15]. Gomperz then defines *communicative competence* as "the knowledge of linguistic and related communicative conventions that speakers must have to initiate and sustain conversational involvement" [12], p.326. In this study we consider the communicative competence of *addressing*, how partners in various conversational settings address each other. Addressing behaviour, i.e. behaviour that speakers show in order to make clear to others present who they address their talk to, is strongly related to, but to be distinguished from behaviour that speakers show to make clear who may or will continue talking; turn-taking is yet an other aspect of interactive behaviour. Speakers have the obligation to make clear to their listeners who they address their talk, their question for example, to and who they expect an answer from. Addressing is a special kind of *referring* and thus principles of multi-modal referring expression generation will also be basic for generation of address terms and addressing gestures for conversational agents. Addressing is a *special* kind of referring because of the different (second person instead of object) role that the referent has in the interaction.

Our analysis of addressing behaviour is partly based on the AMI meeting corpus, which provides audio and video recordings as well as handmade speech transcripts of four participants face-to-face meeting conversations [24]. But we also look at other conversational settings, such as TV discussions and theater plays. We will sketch outlines of a module for generating multi-modal addressing behaviour and addressing expressions in multi-party conversational settings.

[1] University of Twente, the Netherlands, email: infrieks@ewi.utwente.nl, m.theune@ewi.utwente.nl

When in a particular situation a speaker says *"Enriquez, how do you spell your name?"*, then the one who is requested to spell his name is the person *addressed by* the speaker when he performs the request. With the use of the pronouns 'you' and 'your', and the address term 'Enriquez', the speaker refers to the one he is talking to. In face to face conversations the speaker's selection of addressee is also embodied in speaker's gaze: he looks at the person he addresses. There are a number of different *categories* here:

1. the person who is expected to answer the question or to take up the request.
2. the referent of 'you',
3. the person gazed at by the speaker, and
4. the one the talk is directed to.

In a 'normal situation' they all co-refer to the same person in that situation. The distinction between 1. and 4. corresponds to Levinson's distinction between "the one the message is intended for" versus "the one the message is directed to" [23] (for a discussion of Levinson's classification of recipient roles we refer to Jovanovic' thesis [18]). In cases of *indirect* addressing in the sense of Clark and Carlson [6], these two are not the same person.

A conversational situation has a participation frame, a division of parties present in the following categories or modes of participation: speaker(s), addressee(s), co-participants and overhearers. They all stand in a different relation towards the speaker. When we talk we deal with the current participation frame – i.e. the current assignment of parties to these roles or categories – and we also have the opportunity to redesign the frame. *Recipient design* refers to "a multitude of respects in which the talk by a party in a conversation is constructed or designed in ways which display an orientation and sensitivity to the particular other(s) who are the co-participants" [26]. *Audience design* in the sense of [6] and [11] covers overhearers as well as co-participants.

In a quest for conversational rules for addressing, in this paper we discuss the main literature on addressing and present both qualitative and quantitative analyses of addressing in some multimodal corpora of multi-party, face-to-face conversations. Finally, we present a first sketch of a model for the generation of multimodal addressing behaviour.

2 ADDRESSING BEHAVIOUR

Linguists and conversational analysts describe *addressees* as those listeners who are expected (by speakers) to take up the proposed joint project [7]. The turn-taking theory of Sacks et al. has a rule saying

that speakers may select the next speaker by inviting them; if not, others do self-selection as next speaker [26]. Goffman's definition of addressee as "the one the speaker selects as the one he expects a response from, more than from other listeners" [11] refers to this *next speaker selecting notion of addressing*. From the point of view of addressee identification, the counterpart of generating addressing behaviour, one of the *observable* things indicative for who is being addressed is *who is talking next*. We may expect that if the speaker addresses a speech act to a *single* addressee that this addressed person will speak next. This motivates the use of the category of *next speaker* as feature (indicator) in systems for automatic addressee identification [19]. We come back to addressing and initiative when we discuss multi-party conversations in sections 3.2 and 5.

Lerner [22] distinguishes explicit methods of addressing, which are speakers' gaze and naming (the use of vocatives, address terms), from "tacit forms of addressing that call on the innumerable context-specific particulars of circumstance, content, and composition to select a next speaker" ([22], p.177). Lerner examines the context-sensitivity of addressing practices employed by a current speaker to make evident the selection of a next speaker. His discussions are restricted to those turns-at-talk that implement *sequence-initiating actions*, the first parts of adjacency pairs.

Addressing by *gaze* works only if the addressee notices the speaker's gaze and picks up the signal as a sign of addressing; moreover both have to believe that they share this common belief. Mutual gaze between speaker and addressee is basic for grounding in face-to-face conversations. Only mutual gaze between A and B is the most reliable way to establish the belief of A a) that A sees B, b) that A sees that B sees that A sees B, and c) that both share this belief. Accompanied with other messages sent by A (an utterance of a question for example, or a gesture) this may lead B to believe that A's gazing at her means that B is being addressed by A. By looking at B, A checks whether B is ready to receive his message. Others also have to understand that they are *not* selected as next speaker. Thus, "gaze is an explicit form of addressing, but its success is contingent on the separate gazing practices of co-participants" ([22], p.180).

According to Lerner there is one form of address that always has the property of indicating addressing, but that does not itself uniquely specify *who* is being addressed: the *recipient reference term* 'you':

> *The use of 'you' as a form of person reference separates the action of addressing a recipient from the designation of just who is being addressed. In interactional terms, then, 'you' might be termed a recipient indicator, but not a recipient designator. As such, it might be thought of as an incomplete form of address.* [22], p.182.

The speaker will try to complete the addressing act by gazing at the selected recipient, a completion that needs the joint gazing of the intended recipient, and of others present as well, so that they know they are not selected. Thus, for addressing to be complete it requires the joint actions of all participants. This is illustrated by the following fragment from the AMI meeting corpus [24].[2] In the first utterance by speaker P3 *'you"* is not supported by disambiguating gaze; both conversation participants P2 and P0 are gazed at by P3. P2 feels addressed and responds, but P0 also. P2's response overlaps with the elicitation and he is interrupted by P0. It is as if P2 then recognizes that *not he* but P0 was selected as next speaker. P2 and P0's *"Uh"* may also indicate the confusion in the situation.

P3>P0: What do you think, is it fancy?
P2>P3: Uh, it's really
P0>P3: Uh, I think that fancy, we can say it is fancy.

[22] discusses an example of use of referring 'you' directed to a specific individual in a multi-party conversation, where the addressing does not need the support of speaker's gaze at the intended addressee. In a situation where four people are having dinner together, and everybody knows who has prepared the dinner, and the speaker assumes that everybody knows that, the speaker asks: *"Did you cook this all the way through?"* ([22], p.192). Here, the content and context are sufficient to determine the identity of the addressee without the need for explicit addressing behaviour.

The most explicit form of addressing is by use of an address term (which may or may not take the form of a name). This is either used in pre-position, in post-position, or in mid-position, as illustrated by the following examples.

So, *mister money*, what's your opinion according to this remote control?
What do you think, *Ed*?
They wake up fast, *Jessie*, if they have to.

In almost all usages of address terms in talk in face-to-face conversations their function is not purely to call the addressee's attention. If it is, the term is used in pre-position, more often than elsewhere, but most often it seems to be used to put more stress on the addressing, maybe to signal the addressing to co-participants, or to express some affective or social relation with the recipient.

3 ADDRESSING AS A FORM OF MULTIMODAL REFERENCE

3.1 'Changing ideas about reference'

In their paper 'Changing ideas about reference' Clark and Bangerter list some common assumptions concerning reference made by David Olson[3] and his contemporaries ([8], p.26):

1. Referring is an *autonomous act*. It consists of planning and producing a referring expression, which speakers do on their own.
2. Referring is a *one-step process*. It consists of the planning and uttering of a referring expression, and nothing more.
3. Referring is *addressee-blind*. It depends on the context – the set of alternatives in the situation – but doesn't otherwise depend on beliefs of the addressees.
4. Referring is *ahistorical*. It doesn't take account of past relations between speakers and their addressees.
5. The referent belongs to a *specifiable set of alternatives*.

They then proceed to show that each of these assumptions is wrong, arguing that referring (in conversation) is a *cooperative* and *interactive* process involving among other things the establishment of common ground between dialogue participants and the formation of *conceptual pacts* [4], object descriptions which participants jointly decide to use throughout the conversation.

Remarkably, the assumptions listed by Clark and Bangerter still apply to most current approaches to the generation of referring expressions (GRE), which are for the most part direct descendants of Dale and Reiter's classic Incremental Algorithm [9] (see [28] for an

[2] In this and following examples, A > B indicates that A addresses B.

[3] D.R. Olson, Language and thought: Aspects of a cognitive theory of semantics. Psychological Review 77(4):257-73.

overview). These approaches focus on the generation of a definite description of a target object in a visual scene, by selecting properties that apply only to the target referent and not to any of its alternatives (the 'distractors').

So far, GRE has been mainly investigated in the context of *text* generation, where the user is a "distant reader" ([8], p.36) and not a conversation partner. As a consequence, most approaches to GRE do not take conversational factors into account. A few recent exceptions are [17] and [14], who applied GRE in a dialogue setting and took the notion of *conceptual pact* [4] into account by reusing features from earlier references to the same object. Other moves towards a more situated approach to GRE are those by [20] and [28], who developed algorithms for the generation of multimodal referring expressions that combine a verbal description with a pointing gesture, taking the relative locations of speaker, target object and distractors into account to achieve an optimal combination of verbal and nonverbal modalities.

In spite of this recent trend towards more dialogue-oriented, situated approaches to GRE, overall still not much attention is paid to social and conversational aspects of reference. However, for the generation of multimodal *addressing* references, such aspects cannot be ignored, as shown in the following section.

3.2 Addressing as an interactive process

Addressing is a special kind of *multimodal reference*, and here we show that the five assumptions listed by Clark and Bangerter [8] hold even less for addressing than for reference in general. Our examples come from a Dutch TV programme ("B&W", 2002) in which a discussion leader (W) discusses with his guests whether *foie gras* (goose liver) should be banned from restaurant menus for reasons of animal cruelty. Let us go through the assumptions one by one and show that their opposite holds true.

1. *Addressing is **not** an autonomous act.* As we have seen in section 2, for addressing to be complete it requires the joint actions of all participants. The addressee has to pick up on the fact that she is being addressed, and give evidence of this by for example returning the speaker's gaze, nodding, back-channeling, or answering the question that was addressed to her. At the same time, the other participants in the conversation need to know they are not being addressed. They can give evidence of this by for example leaning back, away from the speaker. As Figure 1a shows, body posture and gestures can clearly show that speaker and addressee are 'tuned in' on each other, whereas the co-participants are literally keeping more distance.

2. *Addressing is **not** a one-step process*. It consists of at least two phases: the addressing act by the speaker and the acknowledgement by the addressee. However, when addressing is not immediately successful, additional phases may be involved. In the following example from our TV discussion, W initially relies only on gaze and the content of the dialogue act when addressing B, and then turns to a more explicit form of addressing after this initial attempt fails.

W>B: (gazing at B) Als u nou dat van de kaart haalt, dan is er toch nog voldoende lekkers te eten bij u
If you take this off the menu, won't there still be plenty of nice things to eat at your place
M>W: Nee maar het alternatief is er niet, dat is nou wat ik
No but there is no alternative, that is just what I
W>M: Zei ik tegen meneer B
I said to Mr. B

3. *Referring is **not** addressee-blind.* Addressing involves not only the person(s) being addressed but also those who are not being addressed, and the beliefs of all these participants have to be taken into account. For example, in Lerner's dinner scene discussed in section 2, the speaker asking *"Did you cook this all the way through?"* assumes that A was the cook, and he assumes that everybody knows that A was the cook. So he expects that he can safely address A using *"you"*, without supporting gaze. In other situations, where the participants' common ground is limited, more explicit references may be necessary. In our TV discussion, for example, each time when W addresses one of the discussion participants for the first time, he does this in a particular fashion. He leans strongly toward the addressee (see Figure 1b) and explicitly mentions the addressee's name, followed by a statement or question about his or her identity:

W >B: Mijnheer B, u bent van B Restaurant in O
Mr B, you are from B Restaurant in O
...
W >M: Mijnheer M, u importeert het?
Mr. M, you import it? (where it = foie gras)

W's nonverbal behaviour makes it very clear who is being addressed, also for the co-participants who might not know the addressee's name. The accompanying verbal reference has a double function: besides addressing, it also serves to inform the co-participants and overhearers (the TV audience) of the addressee's name and occupation. This information becomes part of the common ground and may be used for later reference. (The first example of this section showed an unsuccessful attempt by W to use the shared knowledge that B owns a restaurant when implicitly addressing B.)

4. *Addressing is **not** ahistorical.* As shown above, speakers make use of past information from the conversation when they are addressing. Also (as we will discuss in more detail in section 5) dialogue history plays an important role, in that the most likely addressee of the second half of an adjacency pair is the previous speaker.

Our TV discussion contains many sequences where two speakers hold the floor in the conversation, arguing with each other and taking turns in addressing each other. In these bilateral exchanges, the speakers tend to simply refer to their addressees as 'you', relying on nonverbal cues and their shared dialogue history to disambiguate this term. At some points, however, discussion leader W breaks in and assigns the turn to a non-salient participant, using the addressee's name (e.g. *"Is this how it should be done, Mr. C?"*), sometimes supported by a gesture. This is illustrated in Figure 1c, where W (on the left) explicitly assigns the next turn to M (in the middle) by pointing at him and addressing him by name. (At the same time M is pointing at D (on the right), trying to address him while D is talking to someone else.)

5. *The referent does not necessarily belong to a specifiable set of alternatives.* At first sight it seems obvious that all participants in a conversation are potential addressees. However, it is not always clear exactly who the participants are. For example, in our TV discussion the audience does not actively participate in the discussion, but they can be addressed nevertheless: at the start of the programme, discussion leader W. welcomes the viewers using an explicit address term *"Ladies and gentlemen*" and at the end he says goodbye to them. (Note that this illustrates that addressees are not necessarily the next speakers.) The borderline between addressees and non-addressees can be vague. As discussed above, speakers not only take the addressee but also the co-participants into account when designing their address terms. We also sometimes see cases of indirect addressing behaviour [6], as in the following example. Here, W interrupts an argument between animal activist D and culinary journalist S by asking D a question. D answers the question, thus 'technically' addressing


(a) Two participants in discussion

(b) Discussion leader W while addressing

(c) Addressing with pointing gestures

Figure 1. Screen-shots from the B&W discussion programme

W, but he keeps gazing at S while he continues accusing her:

D>S: Daar heeft u uw excuses voor gemaakt aan ons op een halve manier, eh
You apologized to us for that in a half-hearted way, eh
W>D: Maar wat heeft dat met die foie gras te maken
But what does that have to do with foie gras?
D>W: (gazing at S) Nou kijk, uhm, die mevrouw die die is niet een onafhankelijk journalist, zij is meer een promotor van dierenmishandeling als ik het zo hoor
Well, you see, uhm, that lady she she isn't an independent journalist, she's more of a proponent of animal cruelty from what I hear

In his second utterance, D refers to S in the third person, so she cannot be the 'official' addressee [22]. However, his fixed gaze on her shows that she is really the intended recipient of his message.

4 SOCIAL, AFFECTIVE AND COGNITIVE DIMENSIONS OF ADDRESSING

Addressing is an aspect of any form of communication and as such addressing has cognitive, social as well as affective implications and its realisation depends on the capacities of sender, receiver and the communication channels.

Allan Bell [2] points at the impact that the audience has on language style. Style is what an individual speaker does with a language in relation to other people. According to Bell speakers design their style primarily for and in response to their audience. Within a single language audience design does not only refer to variations in speech, it also involves choice of personal pronouns or address terms, but audience design also applies to shift of codes, dialects or languages, in bilingual situations.

4.1 Affective impact of using address terms

In Marsha Norman's play *'night, Mother* the mother has several address terms for her daughter Jessie, who in turn has only a couple of standard ones for her mother. According to Bernardy [3] Mama's use of 'Jessie' mirrors her changing mental state. When overwhelmed or dismayed and unable to articulate a more complete reply, Mama cries out the single word *"Jessie!"* at various points throughout the play.

Since *'night, Mother* has only two players, obviously the primary function of the usage of address terms by Jessie and the mother is not to select who is being addressed, as is suggested by the choice of the term *"address term"*. Most prevalent and decisive for the choice of the terms used is the *affective* value that the usage of the address terms carries and expresses. We are inclined to say that the usage of these terms of endearment in these types of affective conversations have nothing to do with addressing *in the sense of speaker's selecting the intended receiver of his speech act*.

The choice of address terms also plays an important role in expressing different forms of *politeness* [5]. Many languages make a distinction between non-honorific (T) and honorific (V) pronouns[4] where the latter are used to show respect if there is a social distance or difference in power between the speaker and the addressee. On the other hand, when addressing a non-familiar alter, T pronouns can be used to claim solidarity, similar to the use of in-group address terms such as *mate, buddy, pal, honey, dear* ([5], p.107). Such an informal way of addressing can be used as a strategy to save the addressee's positive face, i.e., his need of approval by others. In strategies aimed at saving negative face, a person's need for autonomy, we see the reverse: speakers tend to use honorific address terms in particular when carrying out face-threatening acts such as a requests. E.g., *Excuse me, sir, but would you mind...?* In contrast, the use of honorifics in a non-threatening utterance is much less natural: *Goodness, sir, that sunset is amazing* ([5], p.183).

Given the important affective and social value of address terms usage we expect that careful selection of address terms will improve the believability of synthetic conversational characters [29] and make them appear more socially intelligent [1].

4.2 Channels of communication

Addressing behaviour depends on the conversational setting a part of which is the available "hardware" for communication, the capacities of the channels, auditive, visual, and the capabilities of acting and perception. The mechanics of a conversation, and thus the mechanics of addressing behaviour, depends very much on the communication channels and information made available to the partners. Addressees, over-hearers, remote meetings participants, differ in the ways they participate and interact with other participants in a conversation. It is hard for outside observers to tell who is being addressed in a face to face conversation if they don't share with the other participants the communication channels or the knowledge about the conversation.

Gupta et al. [13] present experiments into the resolution of 'you' in multi-party face-to-face dialogue. They distinguish between *generic*

[4] From the French 'tu' (T) and 'vous' (V).

and *referential* uses of 'you'; and they try to classify the referential uses automatically by identifying the referred-to addressee(s): either one of the participants, or the group. Their results were obtained without the use of visual information; obviously, that makes a hard job because the people in conversation did not have to cope with partners that could not see them. As expected their results on face-to-face meetings were much worse than those they had obtained earlier on the Switchboard corpus, where the parties only had audio communication.

Announcing the name of the addressee helps blind people know that they are being spoken to. (See Irrizary's analysis of the Spanish play *En la ardiente oscuridad* by Buero Vallejo, which has only two sighted characters [16].) Similarly, in discussions for broadcast radio, the discussion leader who feels responsible for informing the audience about who is being addressed, will use more names as address terms than is strictly required for informing his intended addressee that she is addressed.

5 ADDRESSING IN SMALL GROUP MEETINGS

For this study we analysed a number of hand annotated conversations from the AMI meeting corpus [24]. In the scenario-based AMI meetings, design project groups of four players have the task to design a new remote TV control. Group members have roles: project manager (PM), industrial designer (ID), user interface design (UD), and marketing expert (ME). Every group has four meetings (20-30 min. each), dedicated to a subtask. In the first meeting, partners introduce themselves and they use the white board to draw their favorite animal, in the last meeting a clay prototype is presented and evaluated. Conversations in brainstorm sessions, presentations using slide show and laptop, discussions that should lead to a group decision about some detail of the design of the remote control, are embedded in non-conversational activities. Most of the time during meetings partners sit at a square table.

The meetings were recorded in a meeting room stuffed with audio and video recording devices, so that close facial views and overview video, as well as high quality audio is available. Speech was transcribed manually, and words were time aligned. The corpus has several layers of annotation and is easily extendible with new layers. The *dialogue act* (DA) layer segments speaker turns into dialogue act segments, on top of the word layer, and they are labeled with one of 15 dialogue act type labels, following an annotation procedure. A part of the corpus is also annotated with addressee information: DAs are either addressed to the group (*G-addressed*) or to an individual (*I-addressed*). Sub-group addressing hardly occurs and was not annotated. Another layer contains *focus of attention* information (derived from head, body and gaze observations), so that for each partner, at any time instant, it is known who she is looking at; table, white board, or some other participant. In our search for patterns of addressing behaviour, that could inform better models for addressing we noticed a number of interesting fragments. We focused on those dialogue acts in which the speaker takes initiative and tries to elicit some response from some partner that he addresses. It is to be expected that the addressed person will take the turn and respond to the elicit act. The exceptional cases we encountered point at interesting addressing phenomena in small group face-to-face conversations. Recall however that a speaker addressing an individual listener does not necessarily imply that she yields turn to the addressee. For example in the following fragment B makes a proposal to A and C in (0). A addresses her objection to B in (1,3).

(0) B>A,C: I think we should ...
(1) A>B: Okay, but as Carla just said,
(2) A>C: *correct me if I'm wrong*,
(3) A>B: that is too costly for us ...

The embedded invitation to correct her is, however, addressed to C. But A does not give away the floor, instead she continues her objection towards B's proposal. While uttering (2), A gazes at C shortly to notice her non-verbal response. Speakers also invite listeners to give feedback only by briefly gazing at the listeners, especially if they refer to them in the course of their arguing, as A does to C in (1) above. It becomes clear that addressing comes in various flavors, and that conversational acts like asking for feedback can be done non-verbally as well as verbally. The annotators of the meetings were asked to tell if the speaker addressed his dialogue act in particular to some individual in the sense that it was more for her than for others present. Answers to questions are in a sense always addressed to the one who asked the question, and we see that answerers indeed gaze at the previous speaker, but if the issue is a group concern the inform act is addressed to the group.

5.1 Addressing in initiating acts

Addressing in initiating acts is more explicit than addressing in responsive acts, in as far as the speaker who takes initiative in an exchange also has to make clear whom he selects as addressee(s). Dialogue act sequences (we forget about speaker overlap, and multiple floors for a while) have a structure that reflects the fact that partners are interacting, they temporarily participate in changing participation frames around a shared task: to resolve some issue introduced by one of them. Based on conversational analysis we may indeed expect that *the structure of the dialogue gives the most indicative cues to addressee: forward-looking dialogue acts are likely to influence the addressee to speak next, while backward-looking acts might address a recent speaker*. A classical way to model the interaction is in terms of *adjacency pairs*, and Galley et al. [10] used the dialogue structure present in these smallest units of interaction as indicative for addressees: the speaker of the *a-part* of the pair would likely be the addressee of the *b-part* of the pair, and the addressee of the *a-part* would likely be the speaker of the *b-part*. In the one dimensional DA schema that is used in the AMI meeting corpus there is no clear distinction between Backward Looking and Forward Looking DA classes. However, the *elicit types* are primarily FL types of DAs. Typical BL DA types are backchannels, comments about understanding and assessments.

The total number of DAs in the addressee annotated part of the corpus is 9987, of which 6590 are real DAs (i.e. excluding stalls, fragments, backchannels, which do not have an addressee label). Of these, 2743 are addressed to some individual (*I-addressed*); the others are addressed to the Group (*G-addressed*, 3104) or the addressee label is Unknown (which means that the annotator could not tell). In 1739 (63%) of the 2743 *I-addressed* dialogue acts, the addressed person is the next speaker. In our corpus of 652 elicit acts, 236 are *G-addressed*, and 387 are *I-addressed*. Elicits are more *I-addressed* than other DAs. *I-addressed* elicits contain more referring *"you"* than G-addressed elicit acts[5]. In 302 cases (78% the addressee is the next speaker. Thus, FL DAs that are *I-addressed* are more selective for next speaker than *I-addressed* DAs in general, as we expected. When we looked at the instances in the other 22% of the cases, in

[5] If we use "more than", we always mean "significantly more than"; in this case ($\chi^2(df = 1) = 30.66, p < 0.0005$).

which the next speaker did not coincide with the addressed person, we found some interesting addressing phenomena.

Some activities center around one specific actor; a presenter, or someone drawing on the white board, or someone holding the clay prototype that is being discussed. If someone says *"is it heavy?"* it is clear who is being addressed. Or, when a person is drawing his favorite animal on the white board and the speaker makes a guess *"a horse?"*, asking the artist to reveal his secret animal. Actors of activities that are in focus are more salient than others for addressing. Moreover, these actors can tacitly be addressed by others when they comment on, or ask about, the action they perform.

Sometimes, the speaker uses the wrong name or a wrong attribute for his addressee. In such cases an unaddressed listener might feel more entitled to answer the question than the addressee. The speaker uses the referent term 'you' and gazes at P2 to make clear whose identity he is after. But the real marketing guy is called by the attributive use of *"marketing guy"*.

P3>P2: You are *the marketing guy* ? Or
P0>P3: I'm marketing .

In the following fragment it is unclear who is being addressed: a non-addressed attendant tries to answer but is interrupted by the addressee. The speaker indeed gazes at P1 at the end of his question which could easily be taken as if he has selected P1 to speak next.

P2>P0: so how many units should we sell to have a
P1: Well . Uh
P0>Group: Well each unit is is sell uh twenty five Euros .

Another example of unclear addressing: a *you*-utterance without speaker gaze to select the designated addressee

P0>P3: D D Is is there anything you want to add ?
P2>Group: Is there any fruit that is spongy ?

We have seen a number of cases that make clear that proper addressing uses beliefs that speakers have about saliency of persons because of their role in the activity that the group is busy with. Successful addressing is constrained by the general conditions about sharing beliefs about who are salient and who are gazed at as a signal of addressing.

5.2 Reliability

Are the judgments about *"what happens in these conversations"*, about who is being addressed, purely a matter of personal taste? The dialogue annotations and addressee annotations of the part of the AMI corpus we used contains parts made by three different annotators, who followed a documented annotation procedure, using dedicated annotation tools that allows listening to audio, reading the hand made transcripts, as well as looking at video recordings, showing front views of the individual participants as well as an overview of the meeting room. One meeting was annotated by all three annotators. Table 1 shows for each pair of annotators involved Krippendorff alpha values for inter-annotator agreement [21]. For the group of annotators alpha is 0.35 for addressing. The statistics are based on comparing DA-labels of completely agreed DA-segments. Most confusions in the addressing labeling are between *I-addressed* and *G-addressed*, between I and U and between G and U; there is hardly any confusion between annotators about who is addressed when they agree that the DA is *I-addressed* (see also [19]). The table shows that annotators agree more on the addressing of elicit acts than on DAs in general. For the subset of elicit acts we see hardly any U labels used, and when annotators agree that an elicit is *I-addressed* (which happens in 50-80% of the agreed elicit acts), they agree on who is addressed, without exception. Annotators agree more on the addressee of a DA in situations where the speaker clearly gazes at the addressed person. We did not find any indication that annotators systematically confused speaker's gaze with addressing. Addressing is a complex phenomenon and we believe that the low agreement between addressee annotations is due to this complexity.

Table 1. Krippendorff alpha values (and numbers of agreed DA segments) for the three pairs of annotators; for addressing, addressing of elicit acts, dialogue acts (all 15 DA classes), and elicit vs non-elicit acts.

pair	adr	adr-eli	da	da-eli
a-b	0.50(412)	0.67(31)	0.62(756)	0.69
a-c	0.37(344)	0.58(32)	0.58(735)	0.64
b-c	0.33(430)	0.62(53)	0.55(795)	0.80

Focus of attention annotation was done with high agreement, so we may conclude that the annotated data allows a good starting point for research of multi-modal conversational behaviour involved in addressing of eliciting acts and the responsive behaviour that follows in multi-party face to face conversations in general.

6 SYNTHESIS: GENERATING ADDRESSING BEHAVIOUR

We now come to the synthetical part of our project. The four aspects to be considered in the generation of referring expressions (REs) are according to Ielka van der Sluis ([28], p.21-22):

1. costs of the cooperative effort of both speaker and listener
2. accessibility of the object in its context
3. salience of objects
4. responsibility of the speaker for the effectiveness of his choice for the way of referring

All these are relevant in generating multi-modal expressions used in addressing as well. In fact the problem of finding an RE in a conversational setting (see our analyses in section 3 based on the five points made by Clark and Bangerter) is a special case of the more general problem of (language use in) communication. In [25] Prashant Parikh developes a model of communication using frameworks of situation theory and game theory in which he gives an "approximate" set of necessary and sufficient conditions for communication. The general problem of communication is "to find the necessary and sufficient conditions (which involves finding the inferential mechanism) for A to communicate some proposition p to B" ([25], p.475). We consider the *addressing problem* (AP) as an instance of this general communication problem. In a given *conversational situation* the *addressing problem* (AP) *for agent A who wants to address agent B* is to find REs so that if he *uses* them in the given situation, the effect is that he addresses B, or, more formally:

AP is: to find REs $< \phi_i >$ such that if A uses $< \phi_i >$ it has the effect that (A and B share the belief that) A addresses B by means of $< \phi_i >$.

$ADR(A, \phi, B)$ (A addresses B by using RE ϕ) is established when (see the conditions for usage of referentials by Schegloff [27]):

$$Bel_A(Ref(A, \phi, B))$$

$$Bel_A(Bel_B(use_A(\phi) \Rightarrow Ref(A, \phi, B)))$$

$$Bel_A(SB_{<A,B>}(ADR(A, \phi, B)))$$

where $Bel_A(p)$ means: A believes that p; $Ref(A, \phi, B)$: A refers to B by ϕ; $SB_{<A,B>}(p)$: A and B have shared belief p; $use_A(\phi)$: the *act* of using RE ϕ by A.
Parikh did not take co-participants into account, but addressing only becomes a problem in case more than two partners are present. We have seen that the speaker also has to make clear to non-addressed participants who are the ones he has selected as addressee, so that they also know they are not addressed. Formally, if C is a non-addressed side-participant, the additional intention of A is that:

$$Bel_A(SB_{<A,C>}(ADR(A, \phi, B)))$$

If we restrict to single addressing, if C knows B is addressed by A then C knows that he is not addressed. As we have seen in our corpus analyses, the speaker will take this final goal into account when selecting his selection of REs to refer to his addressee(s). For example, a speaker may address by using the name of the addressee and simultaneously point at his addressee. The pointing gesture can be redundant for the addressee but not for side-participants that do not know the name of the addressed person. The intended effect is then that non-addressed partners know who the person is that is being addressed.

Corpus analysis shows that what a speaker has to do when he wants to address someone is to check if there is a communication line open. If not he has to call the intended addressee. In case the speaker does not know the name of the intended addressee, a standard method for GRE can be used to find a referring expression for addressing. For example as in the referential installment: "*The lady in the red shirt in the back*. Could you please close the door?" An RE isn't required when the conversational situation and contents of the dialogue act make it clear who is being addressed; as was the case in Lerner's 'cooking dinner' example.

The AP concerns the selection of the multi-modal means for referring (the REs) that the speaker will and can (given the communication channels available) use for this. The REs are either verbal, or non-verbal. The set of verbal REs contains special types: *proper names* and (politeness) forms of *"you"*. The latter we denote by you_T, you_V or simply as **you**. Proper names (PN) are special in that in any situation if an agent A has PN ϕ, when someone uses ϕ then A will be *called*. For addressing two special types of non-verbal REs are deictic pointing at (**pat**) and gaze at (**gat**). We don't go into the details of the (conditions of) usages of the various types of REs, but to recall that visual and auditive *distance* between speaker and the others is a factor that plays in selecting one of them (impact of perceptual capabilities of parties). We use $< \phi_i >$ for a set of different REs (for example a PN and **gat**), although the temporal order of use of these REs can have effects (at least side effects, i.e. not directly on addressing; recall the affective impact of using PNs in certain positions or in repetition).

A few remarks are in order. In discussions about AP, we should distinguish the *selection* of the best REs for agent A to address B in a given situation from the actual *usage* of the selected REs, and these form the beliefs that A has about the effects of using an RE. Using a name means making a sound or writing the name down in some situation. It is an action that is observable by others and thus others will make inferences about the simple fact that they observe an agent act.

If agent A is involved in an exchange with agent B then A need not use REs to address B if B may suppose that his speech act is addressed to B. For example, if A responds in a face-to-face conversation to a question asked by B, then the situation is such that B is the preferred addressee, or most salient. This motivates a special RE $<>$ (null). If an agent A chooses $<>$ to address then either *no one* is addressed in particular, *or the one who is most salient for* A to be addressed is addressed by A. Every agent has a set of beliefs about the order of saliency of other partners. The selection of REs is determined by this (shared belief of) salience order.

There will be many REs (many verbal REs in particular) that may solve AP for A. The choice is constrained by preferences related to *costs*, and *affective values* of using some RE ϕ:

- A prefers RE ϕ for addressing B above ϕ' when $cost(\phi)$ is less than $cost(\phi')$.
- A prefers RE ϕ for addressing B above ϕ' when ϕ is a better term for the affective relation of A and B than ϕ'.

We take affective values here broadly and also include social values such as politeness. Affective values constrain the selection of address terms, but affective address terms like *"friend"* can also have distinguishing value and help to rule out alternatives from being addressed.

Properties of the *cost* function are: $cost(< \phi_i >) = \sum cost(\phi_i)$, $cost(<>) = 0$. Intuitively, it holds that $cost$(**gat**)> 0, and for all REs ϕ: $cost$(gat)$< cost(\phi)$. (or, gazing at is cheap).

Pointing at the addressee is more precise than gazing at and thus supports the addressing act when the target is at a certain distance from the speaker and there are alternatives (i.e. non-addressed participants) in the neighborhood of the target. The cost of pointing at will be higher the larger the distance to the addressee. The cost of pointing will be larger the shorter the distance between the target and the alternatives. Let l be the line through speaker and addressee; let l' be the line through speaker and the closest non-addressed person. The cost of pointing at is higher the smaller the angle between l and l'. [6] Obviously, there is a trade-off between *saliency* of the intended addressee and the costs of the RE that should do the job: the higher the saliency, the lower the energy that needs to be put in the addressing act. We have seen that saliency of partner depends on the (beliefs the agent has about) the current participation frame; this is, regarding AP, the essential aspect of the current conversational situation.

The proposed approach to solving AP meets all the five points mentioned by [8]. There is no guarantee that the REs selected by the speaker will effectively address the intended recipient, because assumptions that the speaker makes, for example about the state of mind of the recipient, may not hold. Thus agents use repair strategies. Indeed, a solution to AP should not be seen as a one step process. Conversations are *joint projects* [7] of multiple agents, and a solution of AP will thus take the form of an *addressing strategy* that involves the joint work of multiple agents, be it that the speaker takes the initiative in this "project" and he will mostly choose an RE that he

[6] Note that in [28], p.88, the cost of pointing depends on the distance between speaker and target as well as on the size of the target. In the special case of addressing size is hardly variable, but the distance between target and non-targeted alternatives does matter.

believes to be a one step solution to the problem.[7]

The approach outlined here is also valid in situations where for example there is only audio communication. If the speaker believes that the listener can't see him he will conclude that gazing at and gestures are not effective for addressing and these acts will not be chosen. In that case the speaker will rely on vocal means or he believes that there is already a line of communication with the intended addressee so that an empty RE is sufficient for addressing his message.

7 CONCLUSION AND FURTHER WORK

Addressing has up to now not gained much attention in research devoted to the generation of multi-modal referring expressions in multi-agent systems. Addressing involves a special way of referring to partners in conversations, and the social and affective dimension is so prevalent that it strongly determines the choice of address terms and other referring expressions used in addressing.

We presented some outlines of a method for generating addressing behaviour, based on an analysis of a variety of natural multi-party conversations. We have made a first step towards extending existing methods for generating multi-modal referring expressions as in [28] so that they can be used for addressing in multi-party conversations. We formulated AP initially as a problem of a single agent, the agent in the speaker role. But we should take it as a *joint* problem, a problem that can only be solved by joint acts; by speakers as well as listeners. That an act is potentially a joint act is something however, that can only become clear after it has already been initiated, performed by an agent as an attempt, a proposal to interact, as well as taken up by the listeners, in particular the intended addressee(s). The elaboration of this is the next step in our joint project.

ACKNOWLEDGEMENTS

We would like to thank the referees for their comments which helped improve this paper. We are grateful to Dirk Heylen for providing the video and transcripts of the B&W TV discussion. "This work is supported by the European IST Programme Project FP6-0033812. This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein."

REFERENCES

[1] Elisabeth Andre, Matthias Rehm, Wolfgang Minker, and Dirk Buhler, 'Endowing spoken language dialogue systems with emotional intelligence', in *Affective Dialogue Systems*, LNCS 3068, pp. 178–187, (2004).

[2] Allan Bell, 'Language style as audience design', in *Sociolinguistics, a reader and course book*, eds., Nikolas Coupland and Adam Jaworski, Palgrave, (1997).

[3] Maria Lee Bernardy, *Beyond Intuition: Analyzing Marsha Norman's 'night, Mother; with Concordance Data and Empirical Methods.*, Ph.D. dissertation, Iowa State University, 1996.

[4] Susan E. Brennan and Herbert H. Clark, 'Conceptual pacts and lexical choice in conversation', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **22**, 1482–1493, (1996).

[5] Penelope Brown and Stephen C. Levinson, *Politeness - Some universals in language usage*, Cambridge University Press, 1987.

[6] H. H. Clark and T. B. Carlson, 'Hearers and speech acts', in *Arenas of Language Use*, ed., Herbert H. Clark, 205–247, Chicago: University of Chicago Press and CSLI, (1992).

[7] Cost as well as expected effectiveness of actions together determine the selection of referring acts needed for addressing.

[7] Herbert H. Clark, *Using language*, Cambridge: Cambridge University Press, 1996.

[8] Herbert H. Clark and Adrian Bangerter, 'Changing ideas about reference', in *Experimental Pragmatics*, eds., Ira A. Noveck and Dan Sperber, 25–49, Palgrave Macmillan, Basingstoke, (2004).

[9] Robert Dale and Ehud Reiter, 'Computational interpretation of the Gricean maxims in the generation of referring expressions', *Cognitive Science*, **19**(2), 233–263, (1995).

[10] Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg, 'Identifying agreement and disagreement in conversational speech: Use of Bayesian networks to model pragmatic dependencies.', in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, Barcelona, Spain, (2004).

[11] Erving Goffman, 'Footing', in *Forms of Talk*, 124–159, Philadelphia: University of Pennsylvania Press, (1981).

[12] John J. Gomperz, 'The linguistic bases of communicative competence', in *Analyzing Discourse: Text and Talk*, ed., Deborah Tannen, 323–334, Georgetown University Press, (1981).

[13] Surabhi Gupta, John Niekrasz, Matthew Purver, and Daniel Jurafsky, 'Resolving "you" in multi-party dialog', in *Proceedings of SigDial Workshop of European Chapter of the ACL, Prague*, pp. 1–4, (2007).

[14] Surabhi Gupta and Amanda Stent, 'Automatic evaluation of referring expression generation using corpora', in *Proceedings of the Workshop on Using Corpora for Natural Language Generation (UCNLG)*, Morristown, NJ, USA, (2005). Association for Computational Linguistics.

[15] Dell H. Hymes, *Foundations in Sociolinguistics: An Ethnographic Approach*, Philadelphia: University of Pennsylvania Press, 1974.

[16] Estelle Irizarry, 'Some approaches to computer analysis of dialogue in theater: Buero vallejo's en la ardiente oscuridad', *Journal Computers and the Humanities*, **25**(1), (February 1991).

[17] Pamela W. Jordan and Marilyn A. Walker, 'Learning content selection rules for generating object descriptions in dialogue', *Journal of Artificial Intelligence Research*, **24**, 157–194, (2005).

[18] N. Jovanovic, *To whom it may concern. Addressee identification in face-to-face meetings*, Ph.D. dissertation, University of Twente, Enschede, The Netherlands, March 2007.

[19] N. Jovanovic, R. op den Akker, and A. Nijholt, '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, (2006).

[20] Alfred Kranstedt, Andy Lücking, Thies Pfeiffer, Hannes Rieser, and Ipke Wachsmuth, 'Deictic object reference in task-oriented dialogue', in *Situated Communication*, eds., G. Rickheit and I. Wachsmuth, 155–207, Mouton de Gruyter, Berlin, (2006).

[21] Klaus Krippendorff, 'Reliability in content analysis: Some common misconceptions and recommendations', *Human Communication Research*, **30(3)**, 411–433, (2004).

[22] Gene H. Lerner, 'Selecting next speaker: The context-sensitive operation of a context-free organization', *Language in Society*, **32**, 177–201, (2003).

[23] S. C. Levinson, 'Putting linguistics on a proper footing: Explorations in goffman's participation framework', in *Erving Goffman: Exploring the interaction order*, eds., P. Drew and A. Wootton, 161–227, Oxford: Polity Press, (1987).

[24] 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, 'The ami meeting corpus', in *Measuring Behaviour, Proceedings of 5th International Conference on Methods and Techniques in Behavioral Research*, (2005).

[25] Prashant Parikh, 'Communication and strategic inference', *Linguistics and Philosophy*, **14**, 473–514, (1991).

[26] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson, 'A simplest systematics for the organization of turn-taking for conversation', *Language*, **50**, 696–735, (1974).

[27] Emanuel A. Schegloff, 'Some practices for referring to persons in talk-in-interaction: A partial sketch of a systematics', in *Studies in Anaphora*, ed., B. Fox, 437–485, Amsterdam: John Benjamins, (1996).

[28] Ielka F. van der Sluis, *Multimodal Reference, studies in automatic generation of multi-modal referring expressions*, Ph.D. dissertation, University of Tilburg, 2005.

[29] M. Walker, J. Cahn, and S. Whittaker, 'Linguistic style improvisation for lifelike computer characters', in *Entertainment and AI / A-Life, Papers from the 1996 AAAI Workshop.*, (1996).