Turn taking, Utterance Density, and Gaze Patterns as Cues to Conversational Activity

Kristiina Jokinen
Institute of Behavioural Sciences
Siltavuorenpenger 3
FIN-00014 University of Helsinki
kristiina.jokinen@helsinki.fi

ABSTRACT
Conversational activity can be measured in many ways, and often it is related to the participants' turn-taking and feedback activity: the more the participants speak, i.e. produce independent contributions, overlap with each other and provide feedback, the more active and livelier the conversation appears to be. In this paper we discuss activity measures that refer to the frequency of speech and eye-gazing. The goal is to evaluate if the fairly simple "density" measures can produce useful information concerning the participants' communicative activity, and be used to analyse their roles, strategies, and individual differences in conversations.

Categories and Subject Descriptors
H 5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems – *Evaluation/methodology*.

General Terms
Experimentation, Human Factors, Languages, Theory.

Keywords
Non-verbal information, eye-gaze, speech interaction, utterance density, conversational activity

1. INTRODUCTION
Previous research has shown that conversational activity is a complex, highly coordinated process which involves many non-verbal and paralinguistic cues, such as head and hand gesturing, gaze and mutual attention, body posture and spatial proximity, besides verbal utterances [1,2,3,4]. Often conversational activity can be associated with the interlocutors' turn-taking and feedback activity: the more the participants speak, i.e. produce own contributions, and the more they overlap with each other, i.e. produce contributions simultaneously or almost simultaneously, the more active and livelier the conversation appears to be. Laughing, eye-gaze, and body movements are also associated with behaviour which indicates the interlocutors' engagement in the conversation. Activity can be estimated e.g. by visualizing the interlocutors' body and hand movements and comparing their time alignment against each other as in [5]. In this study it was noticed that there are clear peaks in the speaker's body movement at the start of their speaking, and there appears less movement in the interlocutor while listening, i.e. the speakers also move more than their listeners.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ICMI-MMI 2011, November 14-18, 2011, Alicante, Spain.
Copyright 2010 ACM 1-58113-000-0/00/0010…$10.00.

In this paper we discuss activity measures that calculate density, or average frequency of certain conversational behaviors within a time unit. The goal of the study is to evaluate if the fairly simple "density" measures can produce useful information concerning the participants' conversational activity, and be used in the analysis of communication, e.g. in estimating the participants' engagement in the conversation. We assume that conversational activity is related to the interlocutors' intention to build shared knowledge in a given interaction space (cf. [6]), and that the frequency of certain behaviour patterns is an indication of their conversational activity within that space. This kind of constructive dialogue management [7] means that the interlocutors need to both observe the partner's signals concerning how the content of their message is taken up, and also be able to produce appropriate signals themselves, in regard to the partner's message. One of the main tasks in interaction management is thus the grounding of information: finding the intended referents for the partner's expression, and regulating the flow of information with respect to one's own goal. We thus define conversational activity with respect to the participants' constructive activity, i.e. how eagerly they are observed to provide feedback that indicates their awareness and attention to the issues discussed.

Considering the fact that eye-gaze shows the interlocutors' focus of attention, it can be hypothesized that eye-gaze is also an important signal in conversational activity measures: eye-gaze indicates if the speaker currently focuses their attention on the partner, and is willing to contribute to the construction of the shared knowledge. Eye-gaze is thus an important cue in the coordination of social interactions [1,2,3,8,9,10,11], and mutual gazing is effectively used to manage turn taking: if the speaker wants to yield the turn, she looks at the listeners and if one of them also happens to look at the speaker, turn taking can be agreed by mutual gaze. If the listener takes the turn, they usually break the mutual gaze and look away, but if they do not wish to take the turn, they gaze away before turn taking can happen.

It can be assumed that the more frequent mutual gazing patterns, the more active conversation. There are obviously differences in the individual strategies and the participants' conversational roles. We will set to study if gaze activity correlates with speech and turn-taking activity, and to draw support for the hypothesis that the two modalities support each other in signalling engagement in conversational activity, and can thus be used to estimate conversational activity in general. There has been much research concerning eye-gaze and conversation modelling. For instance, [11] describe a gaze model for virtual agents, while [12] and [13] use gaze behaviour to estimate the user's conversational engagement. [8] describe how visual attention is focussed on the partner's face in different social settings (face-to-face and two

different video conditions), and how only a minority of gestures draw eye-fixations, suggesting that social effects play a role in overt gaze-following. Also, gaze behaviour is culture specific: e.g. [14] took cultural differences into account in their studies on virtual agent communication, while [15] compared cultural differences in various feedback functions.

In multi-party dialogues, conversation takes place within a context which is not directly shared by all the interlocutors, and models for mutual knowledge and coordination of interaction thus become more complex. For instance, [9,10] showed that eye-gaze information significantly improves classification accuracy of turn-taking possibilities, compared with the use of speech only features or dialogue acts, but they also noticed that in multiparty conversations head turning is important in turn management, since head movement may function as a more visible signal of the speaker's focus of attention and willingness to take or yield the turn. As already pointed out by [2,16], in multiparty interactions one of the participants is usually the primary addressee of the speaker's message, whereas the others, or the secondary addressees, remain observers and may be unaware of a particular aspect of interaction which the others are aware and may also consider important. The notion of "interaction space" is crucial in multiparty conversations, and the different roles of the participants also affect their means of interaction coordination. Recent work on the interlocutors' behaviour uses a motion tracker [17], and it was noticed that the speaker's behaviour differs from that of the addressees, and that the primary and secondary addressees also differ in the frequency and type of their head and hand movements.

In what follows we study the reciprocal relationship among speech and gaze as signals of conversational activity. We first present the data and annotations used in the studies in Section 2, and introduce gaze and speech activity in the data in Section 3. We then define the "density" model in Section 4, together with its application to the data. The results extracted from the turn-taking data and the different types of information that the measurements seem to provide are discussed in Section 4. The relevance of the technique is also presented. We finally compare information coming from the three measures and their combination, and conclude with an outline future work in Section 5.

2. DATA

We used a portion of the three-person conversational eye-gaze data collected at Doshisha University [18]. The data includes 28 natural conversations among Japanese students, balanced with familiarity and gender, and also some English conversations. The corpus amounts almost 5 hours of data, and it contains 14 conversations with familiar participants and 14 with unfamiliar participants. In each conversation, three participants sit in a triangle formation as shown in Figure 1. One of the participants has his eye movements tracked using an eye-tracker, while the two other participants are videotaped with a digital camera. The eye-tracked person is referred to as ES, and the two other participants are referred to as the left-hand speaker (LS) and the right-hand speaker (RS), accordingly. LS and RS provide a reference point to what ES sees and where his gaze is focused on. ES is always a different person, and to avoid the participants' accommodation to each other, the others rotate so that no group has exactly the same participants. In our experiments we used six 10 minutes long conversations among familiar partners.

The eye-gaze is tracked by the NAC EMR-AT VOXER eye-tracker which can be seen on the table in Figure 1. We used a desktop version of the eye-tracker, but this did not seem to affect naturalness of the conversational interactions or the participants' activities. The sitting around a table is already a natural setup for small group discussions, and as the optics of the eye-tracker is rather robust, ES could move head rather freely, and still be accurately recorded. Some loss of data was caused by blinking, as well as when the participants laughed, since ES's eyes become small and the relevant eye-patterns could not be found.

Figure 1 Data collection setup. The eye-tracked person sits on the right in front of the eye-tracker.

Data are annotated using the Anvil software [19], according to the MUMIN annotation scheme [20]. The scheme has been applied in various languages, to annotate hand gestures, facial expressions, and body posture, and their relation to speech, and it focuses on the shape and the communicative function of the multimodal events, especially in relation to turn taking and feedback giving activity. We adapted the scheme for our specific goals, by adding a level for eye-gaze events. The attributes related to hand gestures and body postures are not used, for the obvious reason that the video focuses on the participants' head and upper body only. The annotation labels used in the experiments are given in Table 1.

Table 1 Annotation features.

Annotation features	*Feature values*
GazeObject (only for ES)	RS, LS, Other
GazeToInterlocutor	ES-Speaking,ES-NotSpeaking, PartnerSpeaking, PartnerNotSpeaking, Away
HeadMovement	Nod, Jerk, Backward, Forward, Tilt, TurntoPartner, TurnSide, Waggle, Other
HeadRepetition	Single, Repeated, None
Handedness	Both, Single
TrajectoryRightHand	Forward, Backward, Side, Up, Down, Complex, Other
TrajectoryLeftHand	Forward, Backward, Side, Up, Down, Complex, Other
HandRepetition	Single, Repeated, None
Turn	Give, Take, Hold, Noturn
Dialogue Act	Backchannel, Stall, Fragment, BePositive, BeNegative, Ask, Inform, Suggest-offer, Other

Gaze is coded with respect to what the person is looking at, in this case, if the speaker is looking at one of the partners or away from the partners. The label GazeObject is used for ES only, and it is based on the gaze path given by the eye-tracker. If the gaze shifts or the gaze path breaks for longer than 0.2 seconds, the two gaze elements are considered two separate gazing events. If the gaze path breaks but the break is shorter than 0.2 seconds, the two gaze elements are regarded as part of the same gaze event. The gaze

events of LS and RS are manually approximated from the video data. Although the gaze events are estimated in two different methods for ES and for LS/RS, this is not considered a problem for the current investigations since we do not combine the annotations but study ES and LS/RS separately. Although Cohen's Kappa coefficient may not be a good statistic for showing that the annotation is independent of the annotator's subjective views [21], we used it to calculate the intercoder agreement on different annotation events, and got the average kappa value of 0.46, which corresponds to a moderate agreement.

3. SPEECH, TURN-TAKING AND GAZE

We first look at the turn-taking and gaze activity with respect to their frequency among the interlocutors. Turn-changes occur either so that the partners wait for each other to finish their utterance or they overlap with each others' speech. Overlappings are considered examples of increased attention and cooperation: the next speaker anticipates the end of the current speaker's utterance and aligns behaviour with that of the partner. They make about 19% of all the turn takings in our corpus (not including backchannels), and in the figures below they are grouped together with the clean ones: in both cases the pattern seems to be the same so that ES gives turns rather equally to the two partners, but LS and RS mostly talk to ES. This can be seen in Figure 2 where the numbers show the absolute frequency of turn-changes.

Figure 2 Turn-changes by each speaker.

As for the gaze activity, Figure 3 shows the gaze distributions among the interlocutors. As with turn-taking, ES also seems to gaze both partners equally, whereas LS and RS both gaze at ES more than each other. However, this may be due to the setup which favoured frontal viewing of ES by LS and RS, but also to the general setting where ES has a special, important role of being the one who is eye-tracked. As already mentioned, the ES gaze events correspond to eye-tracker information while the LS and RS gaze events are manually annotated (the *GazeToInterlocutor* feature), and the speakers are thus reported separately.

For LS and RS we also distinguished the speaking and non-speaking conditions. Figure 4 shows gaze distribution with respect to the speaking time, and an interesting tendency can be seen: the interlocutors tend to avoid looking at the partner when they are speaking themselves, but when they are not speaking, they tend to look at the partner. This can be interpreted so that the interlocutors show their interest in the speaker by directing their attention and awareness to the speaker by gazing at them, whereas when they are speaking themselves, their attention is focused on their own planning and production of what they want to say. Compared with the results by [17], this seems to indicate that the gaze functions differently from the hand and body movements: while the latter support the speaker's own communication management, gazing is the main channel for receiving input from the partner, and thus mainly used while listening to the partner.

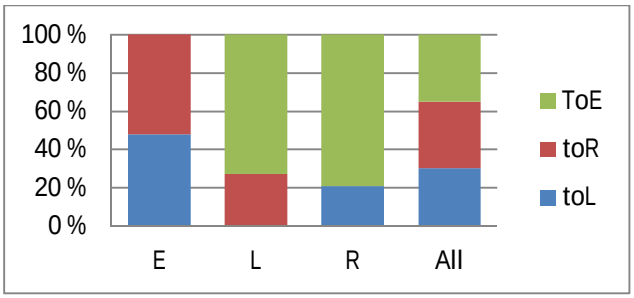

Figure 3 Gaze patterns among the participants.

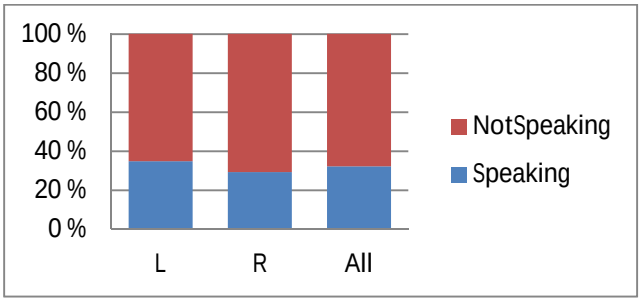

Figure 4 Gaze patterns when speaking.

Finally, by depicting individual turn-taking activities in the six conversations separately, we can compare and visualize their relative contributions to the conversations. Figure 5 shows how turn-taking activity differs in the six conversations. The dialogues are identified by the codes HYI, ISY etc., and we can notice big differences depending on whether turn-taking takes place between ES and LS or RS, or between LS and RS. For instance, in the fourth conversation OHM, most of the conversational activity involves ES and there is hardly any mutual turn-taking between LS and RS. However, in the second conversation ISY, about half of the conversational activity takes place between LS and RS. (In the figure, total numbers are between LS and RS, and between ES and LS or RS, thus the difference compared with Figure 2.)

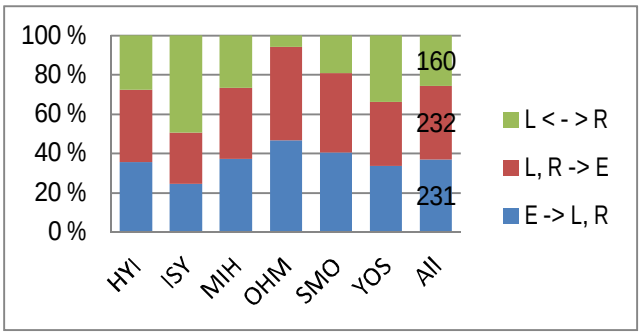

Figure 5 Turn taking activity in each conversation.

4. CONVERSATION DENSITY MODEL

In order to study conversational activity and timing details, [22] produced *utterance density* or the relative speech activity per unit of time. This is calculated by dividing each utterance duration by the sum of the previous and following pause durations.

We calculated utterance density for each of the six speakers and for each conversation in our data, and the results are shown in Figure 6. In the figure, the speakers are organized in groups of three according to their conversational triads. The numbers indicate the conversation (1-6), while ES, LS and RS show the

role of the speaker in the conversation, and the actual speaker is indicated by the letters H, Y, I, S, M, O at the end of the name.

We can see that there are no significant differences in the interlocutors' behaviour with respect to their conversational roles, i.e. all the six interlocutors seemed to behave in similar way when they conversed as ES, LS, or RS. However, there are differences between the conversations as a whole, which could be expected as they are composed of different participants which obviously have different impact on each other. Figure 6 shows the utterance density of each of the three speakers in the conversations marked by the numbers 1-6. The conversations 3 and 5 seem to have high utterance density while conversation 6 has very low density. If these measures are compared with perceptions of the interactions by an outside observer, we can confirm that the conversations 3 and 5 indeed correspond to the interlocutors' lively speech, whereas conversation 6 is less intensive in general.

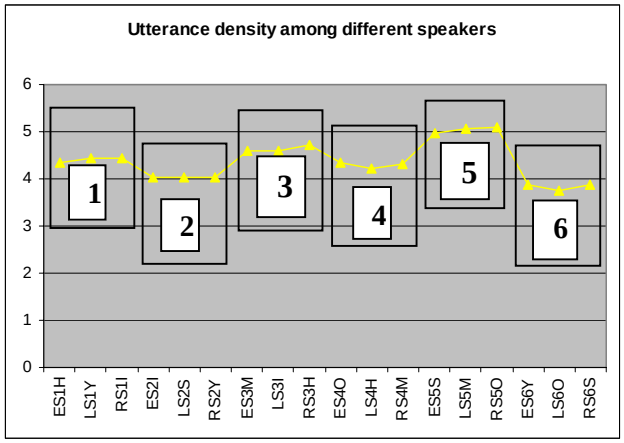

Figure 6 Utterance density rates. The numbers refer to the conversations, the letters H, Y, I, S, M, O to the speaker, and ES, LS and RS to the roles of the speakers.

If we compare utterance density with the turn-taking activity in Figure 5, we notice that the dialogues with high utterance density (1,3, and 5) have fairly balanced turn taking activity among the participants; especially there are turn-takings from ES to LS and RS, and among LS and RS. The low utterance density in the conversations (2,4, and 6) seem to correspond to dialogues where there is a dominant speaker: in 2, the interaction is between LS and RS, whereas in 4, turn-taking is mostly between ES and the other partners.

Turn-taking can indeed be different from utterance density: the speaker may speak a lot without much yielding the turn because the speaker is especially interested in topic, or the speaker's role in the activity (e.g. chair in the meeting, pupil at school) indicates the speakers' social responsibility to the interaction in general. We can conclude that the interlocutors' activity, if measured by the density of speaking time or by the frequency of turn transitions, does not necessarily appear to coincide. For instance, the relative speaking times as depicted with utterance density in general, can be low as in conversations 2 and 4 in Figure 6, but there can be a lot of turn-taking between some participants in the conversation (the corresponding conversations ISY and OHM in Figure 5). This indicates that the different aspects of conversational activity provide different types of information concerning the repertoire available for the interlocutors to express their feedback and construct the shared context: turn-taking activity is a local measure between two participants, while utterance density is global measure that covers average behaviour in a stretch of time. It also shows that in multiparty conversations conversational activity is a complex phenomenon, and differs from that in two-party dialogues. As discussed above, in multiparty conversations, some partners may choose to listen to the others or just say very little, which is reflected in their turn-taking activity but not necessarily in the overall speech activity of the whole dialogue.

Differences can also be seen if we visualize the conversational activity of the participants relative to each other within the conversation duration. In Figure 7, the percentage speech activity by the speakers in the lively conversation 5 (SMO) is plotted against the time axis, and we can see that in the particular stretch of time, all three participants seem to contribute to the dialogue almost equally (RS has a long speech stretch at the end of the clip and thus dominates the end). In Figure 8, however, the low density conversation 6 (YOS) has a different pattern: ES has very small contributions and the whole conversation is dominated by one person with the addressees providing only some feedback.

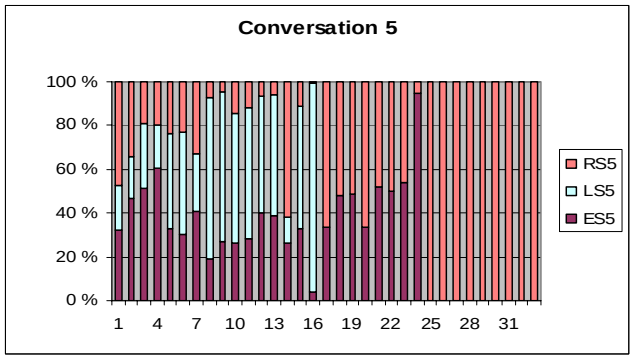

Figure 7 Percentage of speech activity in conversation 5.

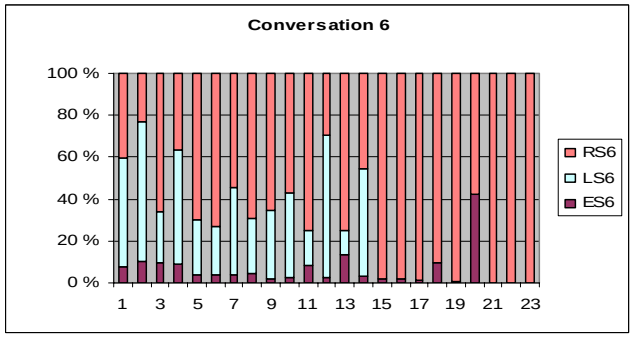

Figure 8 Percentage of speech activity in conversation 6

Yet, if we look at the similar speech activity pattern of conversation 1 (Figure 9), with fairly high utterance density and fairly balanced turn-taking activity, an opposite situation seems to prevail: the speaker ES dominates the conversation, and the others only provide short feedback. Figure 10 brings more complexity to the issue, by showing that Conversation 2 has rather balanced conversational activity among the interlocutors when looking at the speaking time, whereas utterance density is low and turn-taking activity is dominated by turn changes between LS and RS.

We confirm that utterance density does not directly correspond to turn-taking activity. For instance, the two opposite conversations in terms of utterance density, 3 and 6, seem to have rather similar profiles when it comes to overlapping turn-taking activity: they correspond to conversations MIH and YOS in Figure 5, respectively.

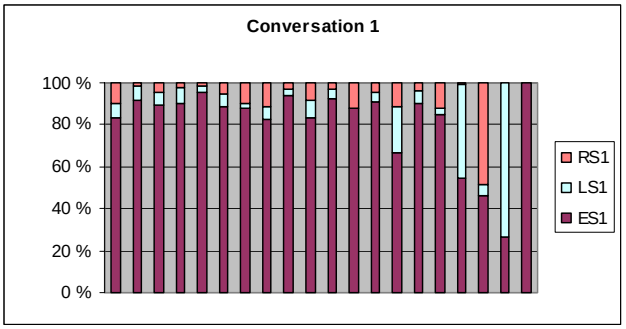

Figure 9 Percentage of speech activity in conversation 1.

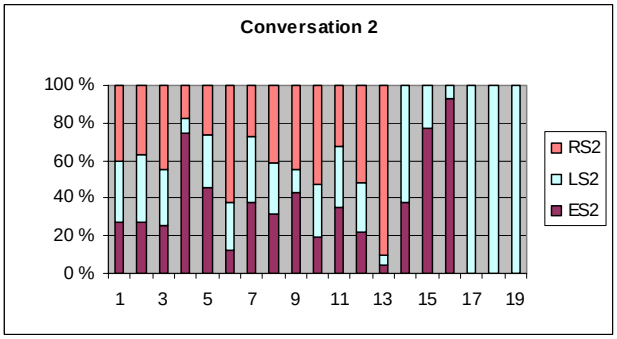

Figure 10 Percentage of speech activity in conversation 2

5. CONCLUSIONS AND FUTURE WORK

In this paper, conversational activity has been studied in casual conversational multi-party settings, and a special focus has been on the speech, turn-taking and eye-gaze as cues to measure such activity. Eye-gaze shows the speaker's focus of attention, and thus helps the partners in creating joint interaction space within which the shared context and joint visual attention can be constructed. The notion of utterance density was introduced to measure the participants' contribution to the conversational activity and their engagement in the conversation in general. It describes conversational activity on the basis of the interlocutors' relative speech density.

The different measures of turn-taking activity, utterance density, and speech visualisation were used to measure conversational activity. An important result from these investigations is that, in multiparty conversations, the interlocutors' turn-taking activity, utterance and gaze density are not directly correlated. For instance, one of the participants may be less active in turn-taking while the speaking activity in the conversation as a whole is large and may show interesting and lively conversation between the other participants. We conclude that the density notions can be a useful measure for the conversational activity analysis, and in the general studies concerning the participants' engagement. Moreover, interlocutors' cooperation on the building of the shared context can be coordinated by non-verbal signals which provide an unobtrusive means to deal with conversation management. In further studies of conversational activity and coordination, all modalities can be effectively used in measuring the participants' involvement in conversation.

The future plan is to work with the same method with respect to the interlocutors' hand and body movement. It is also possible to compare the length of the eye-gazing with respect to conversation activity. We also plan to compare interlocutors' activity in other language communities. For instance, in the NOMCO framework [23], the similar type and similarly coded corpora will be used for this kind of comparison.

6. ACKNOWLEDGMENTS

The author would like to thank the colleagues and students at Doshisha University for the collection and annotation of the eye-gaze data, the members of the NOMCO-project for collaboration on the analysis of multimodal conversational data, and the anonymous reviewers for their comments.

7. REFERENCES

[1] Kendon, A. 1967. Some Functions of Gaze Direction in Social Interaction. *Acta Psychologica* 26, 22–63

[2] Kendon, A. 1990. *Spatial organization in social encounters: the F-formation system, Conducting Interaction: Patterns of behavior in focused encounters*. Studies in International Sociolinguistics, Cambridge University Press

[3] Argyle, M., Cook, M. 1976. *Gaze and Mutual Gaze*. Cambridge University Press, Cambridge

[4] Bavelas, J. B. 2005. Appreciating face-to-face dialogue. In *AVSP-2005*, 1.

[5] Jokinen, K. 2009. Gestures in Alignment and Conversation Activity. *Proceedings of the PACLING Conference*. Sapporo, Japan, pp. 141-146

[6] Clark, H. H., Schaefer, E. F. 1989. Contributing to Discourse. *Cog Sci*. 13, 259–94

[7] Jokinen, K. 2009. *Constructive Dialogue Modelling:Rational Agents and Speech Interfaces*. Chichester: John Wiley.

[8] Gullberg, M. Holmqvist, K. 2006. What speakers do and what addressees look at: Visual attention to gestures in human interaction live and on video. *Pragmatics & Cognition*, Volume 14, Number 1, 2006 , pp. 53-82

[9] Jokinen, K., Nishida, M., Yamamoto, S. 2009. Eye-gaze Experiments for Conversation Monitoring. *Proceedings of the IUCS'09 conference*, ACM, Tokyo

[10] Jokinen, K., Harada, K., Nishida, M., Yamamoto, S. 2010. Turn-alignment using eye-gaze and speech in conversational interaction. *Proceedings of Interspeech*. Makuhari, Japan.

[11] Lee, J., Marsella, T., Traum, D., Gratch, J., Lance, B. 2007. The Rickel Gaze Model: A Window on the Mind of a Virtual Human. In C. Pelachaud et al. (Eds.) *Proceedings of the 7th International Conference on Intelligent Virtual Agents (IVA-2007)*. Springer Lecture Notes in Artificial Intelligence 4722, pp. 296-303. Springer-Verlag Berlin Heidelberg.

[12] Sidner, C.L., Lee, C., Kidd, C., Lesh, N., Rich, C. 2005. Explorations in engagement for humans and robots. *Artificial Intelligence* 166(1-2), 140–164

[13] Ishii, R., Nakano, Y. 2008. Estimating User's Conversational Engagement Based on Gaze Behaviors. In H. Prendinger, J. Lester, Ishizuka, M. (Eds.): *IVA 2008*, LNAI 5208. Berlin Heidelberg: Springer-Verlag. pp. 200–207.

[14] Endrass, B., Rehm, M., André, E. 2009. Culture-specific Communication Management for Virtual Agents. *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. Budapest, Hungary, pp. 281–288

[15] Jokinen, K., Navarretta, C., Paggio, P. 2008. Distinguishing the communicative functions of gestures. *Proceedings of the*

5th Joint Workshop on Machine Learning and Multimodal Interaction, 8-10 September 2008, Utrecht, The Netherlands.

[16] Goodwin, C. 1981. *Conversational Organization: Interaction between Speakers and Hearers*. Academic Press, New York NY, USA.

[17] Battersby, S. 2011. Moving Together: the organization of Non-verbal cues during multiparty conversation. PhD Thesis, Queen Mary, University of London.

[18] Jokinen, K., Nishida, M., Yamamoto, S. 2010. Collecting and Annotating Conversational Eye-Gaze Data. *Proceedings of Workshop "Multimodal Corpora:Advances in Capturing, Coding and Analyzing Multimodality*", Language Resources and Evaluation Conference (LREC), Malta.

[19] Kipp, M. 2001. Anvil – A generic annotation tool for multimodal dialogue. *Proceedings of the Seventh European Conference on Speech Communication and Technology*, pp. 1367–1370.

[20] Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., Paggio, P. 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. In J.-C. Martin, P. Paggio, P. Kuehnlein, R. Stiefelhagen, and F. Pianesi (Eds.), *Multimodal Corpora for Modelling Human Multimodal Behaviour*, Special Issue of the International Journal of Language Resources and Evaluation, pp. 273–287.

[21] Cavicchio, F., Poesio, M. 2009. Multimodal Corpora Annotation: Validation Methods to Assess Coding Scheme Reliability. In *Multimodal Corpora: From Models of Natural Interaction to Systems and Applications*. Springer, Lecture Notes in Computer Science /Lecture Notes in Artificial Intelligence.

[22] Campbell, N., Scherer, S. 2010. Comparing Measures of Synchrony and Alignment in Dialogue Speech Timing with respect to Turn-taking Activity. *Proceedings of Interspeech*. Makuhari, Japan

[23] P. Paggio, J. Allwood, E. Ahlsén, K. Jokinen, C. Navarretta 2010. The NOMCO multimodal Nordic resource - goals and characteristics. *Proceedings of LREC 2010*, pp. 2968- 297.