Multimodal Signals and Holistic Interaction Structuring

Kristiina JOKINEN   Graham WILCOCK
UNIVERSITY OF HELSINKI, Finland
kristiina.jokinen@helsinki.fi, graham.wilcock@helsinki.fi

Abstract

This paper focusses on multimodal activity and its functions, especially as a communicative means to structure the discourse among the interlocutors: to give feedback and indicate turn-takings and mutual agreement. Starting from the assumption that natural language communication is a holistic process which aims at creating shared understanding, and requires interpretation of vocal and visual signals as part of successful interaction, the paper aims to form a coherent picture of the participants' multimodal communication strategies and their engagement in the conversation. It presents observations on the conversational feedback and turn-taking functions especially related to head movement, hand gesturing, and body posture. The main claim concerns the meta-discursive function of visual signals, related to their use as unobtrusive means to control the interaction and to construct shared understanding. The paper deals with synchrony between head movements, hand gesturing and body posture, and builds models for the coordination of communication for intelligent and situated autonomous agents.

Title and Abstract in Finnish

Multimodaaliset signaalit ja holistinen vuorovaikutuksen jäsennys

Tämä artikkeli keskittyy multimodaalisen aktiivisuuden ja sen viestinnällisten tehtävien tutkimiseen, erityisesti sen käyttöön diskurssin jäsentämisessä keskustelijoiden kesken: palautteen antamiseen, vuoronvaihtojen osoittamiseen sekä yksimielisyyden ilmaisemiseen. Lähtien oletuksesta että kielellinen kommunikointi on holistinen prosessi, joka tähtää yhteisen ymmärryksen luomiseen ja vaatii vokaalisten ja visuaalisten signaalien tulkitsemista osana onnistunutta vuorovaikutusta, artikkeli pyrkii muodostamaan koherentin kuvan puhujien multimodaalisista viestintästrategioista ja keskusteluun osallistumisesta. Se esittelee huomioita, jotka käsittelevät keskustelupalautteen antoa ja vuoronvaihtelua erityisesti pään, käsien, ja vartalon liikkeiden avulla. Keskeinen väite koskee multimodaalisten signaalien meta-diskursiivista funktiota, joka liittyy niiden käyttöön ei-häiritsevinä keinoina keskustelun hallinnassa ja yhteisen ymmärryksen luonnissa. Artikkelissa kuvataan pään, käsien, ja vartalon liikkeiden synkroniaa, sekä kehitetään malleja, joita voidaan käyttää älykkäiden ja tilanteisten autonomisten agenttien viestinnän koordinoimiseksi.

Keywords: discourse structuring, multimodal dialogue management, gesturing, synchrony, feedback, turn-taking.

Keywords in Finnish: diskurssin jäsennys, multimodaalinen keskustelun hallinta, elehdintä, synkronia, palaute, vuoronvaihto.

Proceedings of COLING 2012: Posters, pages 527–538,
COLING 2012, Mumbai, December 2012.

Yhteenveto (Summary in Finnish)

Artikkelissa tarkastellaan multimodaalista viestintää ja erityisesti pään, käsien, ja vartalon liikkeitä osana kielellistä kommunikaatiota. Artikkelin päämäärä on kaksitahoinen: toisaalta se tukee kokonaisvaltaista "gestalt"-näkemystä inhimillisestä kommunikaatiokyvystä ja havainnollistaa tätä käytännön esimerkein, toisaalta se kehittelee malleja ja korrelaatioita puhujien keskustelupalautteen ja vuoronvaihtostrategioiden kuvaamiseksi, joita malleja voidaan käyttää autonomisten agenttien viestinnän koordinoinnin pohjana.

Aineisto on kerätty pohjoismaisessa NOMCO-projektissa (Navarretta et al., 2012), ja se koostuu 16:sta noin 6-10 minuutin ensitapaamiskeskustelusta. Puhujat eivät ole tavanneet toisiaan aikaisemmin, ja heidän ainoa tehtävänsä on tutustua toisiina. Keskustelut on translitteroitu ja käännetty englanniksi, ja ne on annotoitu multimodaalisten elementtien suhteen käyttäen muokattua MUMIN annotointiskeemaa (Allwood et al., 2007). Multimodaalisten elementtien jakauma on esitelty englanninkielisen osuuden taulukossa 1, ja eri elementteihin liittyien piirrearvojen jakauma yksityiskohtaisemmin taulukoissa 2– 4. Korrelaatiotulokset on esitetty alla olevissa kuvioissa Figures 1– 5 ja suhteutettu kokonaismäärään.

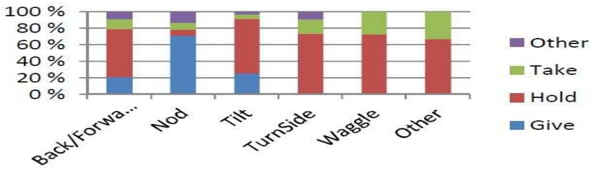

Figure 1: Head movement and turn-taking.

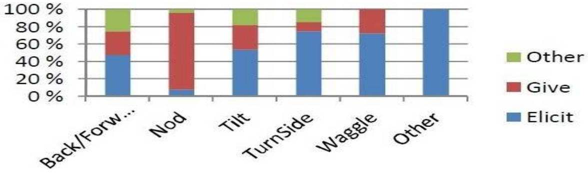

Figure 2: Head movement and feedback.

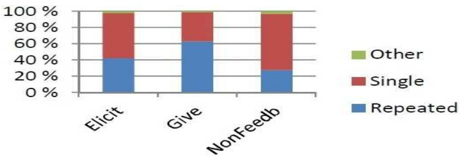

Figure 3: Single stroke vs. repeated hand movement and feedback.

Aineistosta laskettiin korrelaatioita ja yhteisesiintymiä sen selvittämiseksi miten pään, käsien, ja vartalon liikkeet suhteutuvat palautteen antamiseen ja vuoronvaihtoon. Vuoronvaihto oli jaettu kolmeen luokkaan (vuoron ottaminen, pitäminen, ja antaminen) kun taas palaute oli binäärinen luokka (palautteen antaminen vs. saaminen). Kokeelliset hypoteesit olivat seuraavat:

1. pään liikkeet ja palautteen antaminen korreloivat (vrt. Boholm and Allwood (2010))
2. käsieleet ja palautteen saaminen korreloivat (vrt. Battersby (2011))
3. kehon liikkeet ja vuoronvaihto korreloivat (vrt. Kendon (2010) ja f-formaatio)

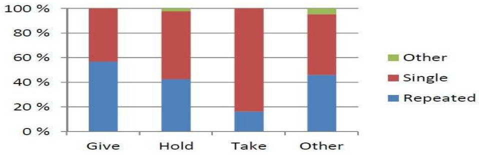

Figure 4: Single stroke vs. repeated hand movement and turn taking.

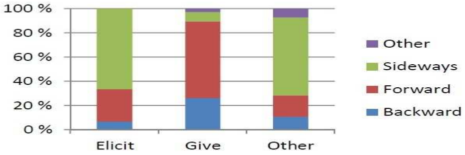

Figure 5: Body movement and feedback.

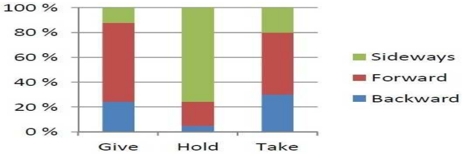

Figure 6: Body movement and turn-taking.

Keskustelijoiden kommunikatiivista aktiivisuutta verrattiin myös heidän itsearviointiinsa keskustelun onnistumisesta. Oletuksena oli, että puhujien itsearviointi liittyy heidän osallisuuteensa vuorovaikutustilanteessa: mitä osallistuvampi ja aktiivisempi puhuja on, sitä positiivisempi on hänen arviointinsa vuorovaikutuksesta. Lisäksi oletettiin, että vuoronvaihtojen määrää voidaan käyttää kriteerinä arvioitaessa keskustelijoiden aktiivisuutta ja osallisuutta keskusteluun: mitä enemmän vuoronvaihtoja, sitä aktiivisemmin puhujat osallistuvat keskusteluun koska he pyrkivät nopeasti koordinoimaan viestejään. Yksittäisten kuvauspiirteiden ja puhujien itsearvioinnin korreloinnissa löytyi statistisesti merkittävä korrelaatio ($p<0.05$) kommunikatiivisen aktiivisuuden ja keskustelukokemuksen onnellisuuden välillä (0.688).

1 Introduction

The paper starts from the assumption that natural language communication is a holistic process which aims at creating shared understanding, and requires perception and interpretation of a wide variety of vocal and visual signals as part of successful interaction. For instance, (Crystal, 1975) has talked about paralanguage, i.e. the 'tone of voice' that bridges non-linguistic forms of communicative behaviour and the core linguistic areas of grammar, vocabulary and pronunciation, whereas (Allwood, 2002) describes on a general level how all body movements that influence the partner can be considered communicative.

We study the use of visual signals, i.e. head movements, hand gestures, and body posture in interaction coordination, and use the term *multimodality* to emphasise the multiple modalities involved in communication. Our approach is holistic: we aim at a coherent picture of the participants' multimodal communication strategies and engagement in interaction. The goal of the paper is two-fold: on one hand, it contributes to the holistic 'gestalt' view of human communication capability, and exemplifies this in practise by examining the interlocutors' multimodal feedback and turn-taking strategies. On the other hand, it studies correlations between head movement, hand gesturing, and body posture so as to develop models for their synergy and synchrony in relation to the giving feedback and taking turns, in order to enable of automatic coordination of communicative functions for autonomous agents.

The main claim concerns the functions of multimodal signals related to their use as unobtrusive means to control the interaction and to construct shared understanding. Following Kendon (2004) we use the term *metadiscursive* to describe gestures that regulate the flow of information rather than express semantic content. Although multimodal aspects have been the subject of many previous studies, the correlation of these particular aspects has not been much discussed in computational linguistics. On the basis of our corpus we argue that body posture is an iconic way to hold the conversational turn, hand gesturing signals turn-taking, and nodding is an effective means to give feedback. Moreover, we hypothesize that the interlocutors' communicative activity gives evidence for their engagement in the interaction, and that engagement positively correlates with the interlocutors' self-assessment of the success of the interaction.

The paper is exploratory in nature, and combines observations of the data into a multimodal interaction model. It is structured as follows. Section 2 discusses previous studies that are confirmed and expanded by our work. Section 3 presents the data and corpus examples used. Section 4 provides results concerning the function and correlation of different visual modalities, and discusses conversational engagement in terms of multimodal activity in the conversation in general. Section 5 describes the interlocutors' own assessment of the interaction. Conclusions are drawn in Section 6.

2 Multimodal aspects of communication

Much of the information related to the basic enablements of communication (being in contact with, and being able to perceive the partner) are conveyed by multimodal means such as eye-gaze, head nods, facial expressions, etc. Multimodal signals also carry emotional and physical feelings, moods, interest, reactions, etc., and they convey social functions of communication: they help the interlocutors to share understanding, bond with their partners, and create social identity (Feldman and Rim, 1991). They also serve to control and coordinate the information flow in interactions. For instance, Kendon (2004) talks about meta-discursive function of hand gestures, while gaze is important in turn-taking (Argyle and Cook, 1976; Goodwin, 2000;

Lee et al., 2007; Jokinen et al., 2010). Body posture can be used to control the interaction: leaning forward often means interest whereas leaning backward signals withdrawing from the conversation (Jokinen and Scherer, 2012). Jokinen and Pärkson (2011) notice that some body movements are used to fill pauses in conversation if the speaker may not want to take the turn or is unable to take the turn. In group conversations, participants create a joint transactional space by forming spatial patterns, f-formations (Kendon, 2010), and they signal contact and availability to take part in the conversation by multimodal activity (Battersby, 2011).

Besides the functions of multimodal signals, there is substantial literature on the signals themselves. Correlation, for instance, between gaze and gesturing has been studied by Gullberg and Holmqvist (1999), and between speech and head movement by Boholm and Allwood (2010). Synchrony between acoustic cues (pitch accents) and visual cues (beat gestures, head nods, and eyebrow movements) is studied in detail by Krahmer and Swerts (2007), while various others have looked at alignment (Pickering and Garrod, 2004) and mimicry (Chartrand and Bargh, 1999).

Computational modelling of multimodal communication has focussed on Embodied Conversational Agents; see an overview in André and Pelachaud (2010). For instance, Nakano and Nishida (2007) experimented with an eye-gaze model to ground information in interactions with an embodied conversational agent, while (Swartout et al., 2010) focused on building realistic and engaging virtual agents for various practical applications. Recently, robots have become an important application domain. Bennewitz et al. (2007) developed a robot companion that can recognize gestures and become engaged in interaction. Jokinen and Wilcock (2012) and Csapo et al. (2012) describe multimodal interaction in conversations with the Nao robot. Non-verbal aspects can also be important in computational and cognitive linguistics: Koller et al. (2012) used eye-tracking in monitoring the hearer's reference resolution process, and Qu and Chai (2009) showed that the coupling of speech and gaze streams in a word acquisition task can improve performance significantly.

3 Data and hypotheses

The data consists of 9 (out of the 16) Finnish first encounter dialogues collected in the NOMCO project (Navarretta et al., 2012): 5 female and 4 male participants, aged 21-40, all native speakers of Finnish. Each participant took part in two conversations with different partners they did not know in advance. The participants were not given any particular topics to discuss but were asked to make acquaintance with the partner they had never met before. After the recordings the participants filled out a web-based questionnaire concerning how they felt about the conversations. Instructions were kept minimal: the participants were only advised to stand in a specific spot, so they would remain in frame throughout the recording. The participants were standing, since recording started when they entered the interaction space through the door, and it was thus a natural posture. It also provided a model of real first encounter situations which often take place standing in a party, lecture hall, etc.

The recordings were made in ambient lighting with three static cameras. Two cameras recorded the participants individually while one camera shot both participants at once. A separate audio recorder was used to obtain a clearer audio track. The encounters are about 6–10 minutes long. They are transcribed and translated into English, and the frontal views were annotated following a modified MUMIN annotation scheme (Allwood et al., 2007). Multimodal features concern the function and form of head, hand, and body movement. Interaction features are related to turn-taking (take, hold, and give the turn) and feedback (give vs. elicit feedback).

Annotation was done by two annotators independently and checked by an expert annotator. Inter-coder agreement between the annotators was checked by calculating kappa co-efficient on two dialogues. The corrected Kappa values varied between 0.71 on head movement to 0.41 on facial display and 0.36 on hand gesturing, while category agreements were 80.7%, 58.1%, and 56.5%, respectively. The main disagreements concern feedback direction (give vs. elicit) of hand gesturing and facial display, as well as trajectory type (complex vs. up vs. side vs. other) on hand gesturing. The body posture annotation was done only by one annotator whose annotations were then selected to be used in the experiments throughout.

Feedback (fb)	**Head**	**Hand**	**Body**	**Turn-taking (tt)**	**Head**	**Hand**	**Body**
Feedback Elicit	201	205	30	TurnGive	295	72	33
Feedback Give	375	90	38	TurnHold	220	228	41
Unclassified fb	69	107	28	TurnTake	64	37	10
Total	**645**	**402**	**96**	Unclassified tt	66	65	12
Total of all	**56** %	**35** %	**9** %	**Total**	**645**	**402**	**96**
Total of all fb	**61** %	**32** %	**7** %	**Total of all tt**	**58 %**	**34 %**	**8 %**

Table 1: Communicative functions (feedback vs. turn-taking) related to head, hand, and body movements. Unclassified movements are not annotated with respect to the given communicative feature, e.g. of the 645 head movements, 69 are not related to feedback.

Statistics of the multimodal elements (n = 1143) with respect to the communicative functions are given in Table 1, and the statistics of the different annotation features in each category are detailed in Tables 2– 4. Slightly more than half (56%) of all the communicative movements are produced by head, 35% by hand, and only 9% by body. Distributions with respect to feedback and turn-taking have a similar tendency.

Co-occurrences and correlations were calculated on the basis of the data, to find out how hand, head, and body movement were related to feedback and turn-taking functions. The experimental hypotheses were as follows:

1. Correlations can be found with respect to head movements (nods) and feedback giving (cf. Boholm and Allwood (2010)).
2. Correlations can be found with respect to hand gesturing and feedback elicitation (cf. Bavelas and Chovil (2000), Battersby (2011)).
3. Correlations can be found with respect to body movements and turn-taking (cf. f-formation in Kendon (2010)).

4 Experimental results

4.1 Head movement

The majority of head movements are nods (Table 2). We previously showed (Toivio and Jokinen, 2012) that there is a statistically significant difference between up-nods and down-nods, and the difference correlates with different semantics: down-nods are used in situations where common ground is already established, while up-nods are used if the presented information is surprising in the given context. Co-ocurrences normalised with respect to the total numbers and time show that different head movements correlate with turn-taking and feedback functions (Figures 1

Head movements	**Count**	**Percent**	**Head movements**	**Count**	**Percent**
Backward	38	6 %	TurnSide	76	12 %
Forward	77	12 %	Waggle	11	2 %
Nod	345	53 %	Other	3	0 %
Tilt	95	15 %	**Total**	**645**	**100** %

Table 2: Feature values and their frequency in head movements.

and 2 in the Finnish summary). In particular, nodding is a partner oriented signal, used to give feedback, and to signal turn giving. Nodding indicates cooperation: the speaker is engaged in interaction and willing to listen to the partner. The other head gestures, back/forward movements, tilt, sideways turning, and waggle are mostly used to elicit feedback, and also co-occur with turn holding events. We may assume that the speakers use them for regulating the reception of their own speech rather than backchannelling what the partner has said.

In summary, head gestures seem to convey two significantly different visual signalling patterns: nodding is a partner-oriented signal and almost exclusively used to give feedback or the turn to the partner, while other head movements are related to holding one's own turn, during which the speaker's visual signals are interpreted as feedback eliciting signals.

4.2 Hand gesturing

Table 3 shows that hand gestures usually employ both hands, they contain slightly more often single stroke than repeated ones, and they indicate rhythm of the speech (beating). From Table 1 we also notice that 68% of the turn related gestures occur with turn holding, 20% with turn giving, and 10% with turn taking events. Almost 70% of hand gestures are used for eliciting rather than giving feedback, which agrees with Battersby (2011) who demonstrated that the speakers gesture more than the listeners.

Hand movements	**Count**	**Percent**	**Hand movements**	**Count**	**Percent**
Handedness			**Interpretation**		
Both hands	265	66 %	Deixis	8	2 %
One hand	137	34 %	Emphasis	19	5 %
Repetition			Rhythm	185	46 %
Repeated	174	43 %	Other	188	47 %
Single stroke	220	55 %			
Other	8	2 %	**Total movements**	**402**	**100** %

Table 3: Feature values and their frequency in the three hand movement types.

Figures 3 and 4 in the Finnish summary show gesturing patterns with respect to feedback and turn-taking. Most communicative gestures can be one-stroke or repeated, but there is a tendency to give feedback and give turns with repeated gesturing. Single stroke hand gestures co-occur slightly more with feedback elicitation, but the difference is not significant. It is interesting, however, that when taking the turn, 85% of gestures are one-stroke. This suggests that the next speaker prepares for their turn by moving their hands into a kind of "speaking position" which would allow them to gesture in rhythm with their speaking. It has been shown that the gesture peak co-occurs with the speech stress (Krahmer and Swerts, 2007; Kendon, 2004), and thus the non-repetitive turn taking gestures may actually be part of the embodied speech production:

through such gesturing the partners indicate that they are ready to speak. Single stroke gestures signal turn taking and thus effectively prevent the speaker from continuing their turn.

The speaker's turn-holding gestures seem to be rhythmic movements that accompany the speech rather than intend to catch the partner's attention. Gullberg and Holmqvist (1999) showed that the listeners do not look at the speaker's gesturing (but at their face), and that the listener's gaze follows the speaker's hand movement only if the speaker has focussed their attention on their hand, too. Bavelas and Chovil (2000) talk about the meaning of gestures with respect to the degree of redundancy between a gesture and the co-occurring utterance, and notice that reference to the common ground deploys smaller and less explicit gestures, whereas gesturing associated with novel referents is larger and explicit. We can hypothesize that the speaker's continuous gesturing is a behaviour pattern associated with their turn holding and simultaneous feedback elicitation: the speaker can unobtrusively refer to the shared information and elicit feedback by small gesturing, without needing to put the intention into explicit words. We can also speculate that the partner's single stroke gestures are attention catchers that invite the speaker to give the turn, without explicit verbal confrontation or competition for the floor. More gesturing can indicate that the interlocutors are excited and thus active in taking turns and eliciting feedback from their partner.

To summarise, hand gestures in our data are used for feedback eliciting and turn holding. We conform to the assumption that speech and gesturing are closely linked in the production process, and presented an observation concerning gestures and turn-taking to support this view: most turn-taking gestures are single stroke gestures related to the speakers adjusting themselves into a speaking position where beat gesturing is easy to produce.

4.3 Body movement

Only about 9% of the communicative multimodal signalling is assigned to body movements (Table 1), most of them forward or sideways orientations (Table 4). An interesting, novel observation in our data is that sideways orientation relates to turn holding, while the body facing the partner (possibly moving forward and backward) opens up a turn exchange (Tables 5-6). Also, body posture sideways elicits feedback, whereas body movement backward and forward gives feedback.

Body movements	**Count**	**Percent**
Backwards	15	16 %
Forward	37	38 %
LeaningSideways	41	43 %
Other	3	3 %
Total	**96**	**100** %

Table 4: Feature values and their frequency in body movements.

Body posture seems to have an iconic function in interaction. The posture where the speaker is squarely towards the partner is potentially challenging, but a sideways posture avoids direct face-to-face interaction and gives the speaker a wider transactional space to plan contributions and to hold the turn. Standing sideways locks the space and direct back/forward movement, and consecuently prevents the partner from entering the space. Kendon (2010) describes spatial organization of the speakers in group conversations with the notion of f-formation ('facing formation'). In two-party conversations, we can say that the speakers tend to control turn-taking by similar spatial orientation: by facing or turning away from the partner.

5 Interlocutors' own assessments and communicative activity

We also compared the interlocutors own assessments of the interactions with the observed communicative activity in the same interaction. It is assumed that multimodal activity in giving feedback and taking turns can be used to estimate the interlocutors' communicative activity and engagement in the conversation in general: the more multimodal activity, the more engaged the interlocutors are in the activity. We also hypothesize that the interlocutor's self-assessment of the interaction is related to the amount of their communicative activity: the more engaged (i.e. the more active) the interlocutor is, the more positive impression she has about the interaction.

	Mean	**Min**	**Max**		**Mean**	**Min**	**Max**
Enjoyable	3.7	3	4	Anxious	1.9	1	4
Friendly	2.5	1	3	Natural	2.6	1	4
Impressive	2.1	1	4	Happy	2.6	1	4
Nice	4.0	3	5	Tense	1.9	1	4
Interesting	3.8	2	5	Awkward	2.1	1	4
Relaxed	3.0	2	4	Angry	1.3	1	1
				Average	**2.6**		

Table 5: Statistics of the self assessment questionnaire.

Self-assessments were based on a questionnaire where the users rated their interaction with respect to a set of descriptive adjectives on a Likert-scale 1 – 5, with 1 meaning 'I disagree, the interaction was not like this at all' and 5 meaning 'I agree, the interaction was very much like this'. Table 5 shows the mean, minimum and maximum values for each adjective. In general, participants found the interactions enjoyable, nice, and interesting (mean values of 3.5, 3.9, and 3.6). Ratings for the negative impressions, angry, tense and anxious, are clearly lower.

Interlocutor	**Gender**	**Activity**	**Assessment**	**Turn-taking**
1	F	3.9	2.42	103
2	F	5.5	2.58	106
3	M	1.7	2.58	99
4	F	2.6	2.67	107
5	F	4.8	2.58	123
6	F	7.5	2.92	107
7	F	2.6	2.25	95
8	M	1.7	2.08	96
9	F	1.5	-	93
10	M	0.9	2.17	74
11	M	0.6	2.50	74
Mean		3.0	2.4	

Table 6: The interlocutors' gender, average activity, self assessment mean score and turn-takings.

Table 6 lists the interlocutors' self-assessment mean score, multimodal activity with respect to the length of the interaction (Activity), and the number of turn-takings in the interaction (self assessment from participant 9 is missing). The table shows that the normalized multimodal activity has a rather large variation (mean = 2.7, standard deviation = 2.1), contrary to the interlocutors' self-assessments (mean = 2.2, standard deviation = 0.25). We can also notice that the speakers with most turn-taking activity (the speakers 4, 5, and 6 who all have more than 100 turn-takings in their interactions) have self-assessment values which are above the mean values, i.e. they have positive impressions of the interactions. This allows us to hypothesise that communicative activity and positive evaluation are related. However, we cannot conclude which way the causal relationship goes: maybe the positive impressions are due to the interlocutor's communicative activity, or maybe the large activity is due to the interlocutor's positive attitude. It is likely that the relation is not either-or, since impressions can change during the encounters, and the participants' predisposition may also affect their activity.

Considering the correlations between individual descriptors in the interlocutors' self-assessment and their communicative activity, we found a statistically significant correlation ($p<0.05$) between the activity and 'happiness' (0.688).

6 Conclusion

In this paper we have looked at non-verbal activity from a holistic point of view related to interaction control and feedback. We studied the participants' multimodal behaviour patterns, and correlated them with their engagement in the interaction and their self-reported impressions of the interaction. We found a positive dependence between the objective measures of communicative activity and the speakers' own impressions of the interaction, although the direction of the relation cannot be concluded. Considering the hypotheses set in Section 3, we identified the relation between head movements and feedback to concern nodding, while the other head movements correlated with turn holding. The hypothesis about hand gesturing and feedback elicitation seems to hold, but we also further specified single stroke hand gestures to be used to coordinate the interaction and turn-taking. This is also supported by the fact that motor activity accompanies speech: listener's gestures are related to their intention to take the turn while the speaker's gestures coincide with the stress of their utterances. Finally, correlations were found with respect to body movements and turn-taking, with an observation of the iconic function of body posture: the sideways posture seems to indicate turn-holding. In general, the results are interesting and unique, requiring further investigations.

Future studies can also answer the questions concerning the relative contribution of visual and vocal communication to multimodal interaction in general. Moreover, it is useful to investigate what are the optimal units for information exchange, and what is the role of context in the interpretation of these signals. It is necessary to use a larger corpus (e.g. all 16 dialogues, and even more) to draw more comprehensive conclusions. We are also in the process of exploring automatic analysis techniques for the recognition of visual signals.

Acknowledgments

Thanks to NOMCO partners in Gothenburg and Copenhagen for data collection and comparison, and to Emmi Toivio, Joonas Toivio, and Johanna Reinikainen for analysis of the Finnish data.

References

Allwood, J. (2002). Bodily Communication – Dimensions of Expression and Content. In Granström, B., House, D., and Karlsson, I., editors, *Multimodality in Language and Speech Systems*, pages 7–26. Kluwer Academic Publishers, Dordrecht.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., and Paggio, P. (2007). The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing. *Language Resources and Evaluation*, 41(3–4):273–287.

André, E. and Pelachaud, C. (2010). Interacting with embodied conversational agents. In Cheng, F. and Jokinen, K., editors, *Speech Technology: Theory and Applications*, pages 123–150. Springer.

Argyle, M. and Cook, M. (1976). *Gaze and Mutual Gaze*. Cambridge University Press.

Battersby, S. (2011). *Moving Together: the organization of Non-verbal cues during multiparty conversation*. PhD thesis, Queen Mary, University of London.

Bavelas, J. and Chovil, N. (2000). Visible Acts of Meaning. An Integrated Message Model of Language in Face-to-Face Dialogue. *Journal of Language and Social Psychology*, 19(2):163–194.

Bennewitz, M., Faber, F., Joho, D., and Behnke, S. (2007). Fritz - a humanoid communication robot. In *Proceedings of the 16th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*.

Boholm, M. and Allwood, J. (2010). Repeated head movements, their function and relation to speech. In *Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*.

Chartrand, T. L. and Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, 76:893–910.

Crystal, D. (1975). Paralinguistics. In Benthall, J. and Polhemus, T., editors, *The body as a medium of expression*, pages 162–174. London: Institute of Conteporary Arts.

Csapo, A., Gilmartin, E., Grizou, J., Han, J., Meena, R., Anastasiou, D., Jokinen, K., and Wilcock, G. (2012). Multimodal conversational interaction with a humanoid robot. In *Proceedings of 3rd IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2012)*, Kosice.

Feldman, R. and Rim, B. (1991). *Fundamentals of Nonverbal Behavior*. Cambridge University Press, Cambridge.

Goodwin, C. (2000). Action and embodiment within situated human interaction. *Journal of Pragmatics*, 32:1489–1522.

Gullberg, M. and Holmqvist, K. (1999). What speakers do and what listeners look at. A comment on visual deixis and mimesis. *Pragmatics and Cognition*, 7:35–63.

Jokinen, K. (2010). Pointing gestures and synchronous communication management. In Esposito, A., Campbell, N., Vogel, C., Hussein, A., and Nijholt, A., editors, *Development of Multimodal Interfaces: Active Listening and Synchrony*, pages 33–49. Springer.

Jokinen, K., Harada, K., Nishida, M., and Yamamoto, S. (2010). Turn-alignment using eye-gaze and speech in conversational interaction. In *Proceedings of 11th International Conference on Spoken Language Processing (Interspeech 2010)*, Makuhari, Japan.

Jokinen, K. and Pärkson, S. (2011). Synchrony and copying in conversational interactions. In *The 3rd Nordic Symposium on Multimodal Interaction*, Helsinki.

Jokinen, K. and Scherer, S. (2012). Embodied communicative activity in cooperative conversational interactions - studies in visual interaction management. In *Acta Polytechnica. Journal of Advanced Engineering*.

Jokinen, K. and Wilcock, G. (2012). Multimodal open-domain conversations with the Nao robot. In *Fourth International Workshop on Spoken Dialogue Systems (IWSDS 2012)*, Paris.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

Kendon, A. (2010). Spacing and orientation in co-present interaction. In *Lecture Notes in Computer Science, 5967*, pages 1–15. Springer.

Koller, A., Garoufi, K., Staudte, M., and Crocker, M. (2012). Enhancing referential success by tracking hearer gaze. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, Seoul, South Korea.

Krahmer, E. and Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, 57(3):396–414.

Lee, J., Marsella, T., Traum, D., Gratch, J., and Lance, B. (2007). The Rickel Gaze Model: A Window on the Mind of a Virtual Human. In *Proceedings of the 7th International Conference on Intelligent Virtual Agents (IVA-2007). Springer Lecture Notes in Artificial Intelligence 4722*, pages 296–303, SpringerVerlag Berlin Heidelberg.

Nakano, Y. and Nishida, T. (2007). Attentional behaviours as nonverbal communicative signals in situated interactions with conversational agents. In Nishida, T., editor, *Engineering Approaches to Conversational Informatics*. John Wiley.

Navarretta, C., Ahlsén, E., Allwood, J., Jokinen, K., and Paggio, P. (2012). Feedback in Nordic first-encounters: a comparative study. In *Proceedings of Eighth International Conference on Language Resources and Evaluation (LREC 2010)*, Istanbul.

Paggio, P., Allwood, J., Ahlsén, E., Jokinen, K., and Navarretta, C. (2010). The NOMCO Multimodal Nordic Resource - Goals and Characteristics. In *Proceedings of Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pages 2968–2973, Valletta, Malta. European Language Resources Association (ELRA).

Pickering, M. and Garrod (2004). Towards a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169–226.

Qu, S. and Chai, J. (2009). The role of interactivity in human-machine conversation for automatic word acquisition. In *Proceedings of the SIGDIAL Conference 2009*, pages 188–195.

Swartout, W., Traum, D., Artstein, R., Noren, D., Debevec, P., Bronnenkant, K., Williams, J., Leuski, A., Narayanan, S., and Piepol, D. (2010). Ada and Grace: Toward realistic and engaging virtual museum guides. In *International Conference on Intelligent Virtual Agents (IVA)*, pages 286–300.

Toivio, E. and Jokinen, K. (2012). Multimodal Feedback Signaling in Finnish. In *Proceedings of the Human Language Technologies - The Baltic Perspective*.