Contents

Abstract

This research presents a multimodal non-verbal conversation analysis performed on a typical institutionalized political TV interview. The focus is on facial, hand, and body gestures and their role in carrying out communicative functions such as feedback and how speakers know when it is their turn to speak. What we wanted to know is what non-verbal gesturers speakers of institutionalized interviews use and also, importantly now these gestures compare between cultures. This work is based on previous studies done in Greece and similar investigations in Europe for comparison between different cultures. In this research there was a comparison made between the Greek study and this one. The conclusion was that they have similar frequencies of non-verbal expressions. The tools and the coding scheme that were used in the current study will be described. We discuss how different interview settings, institutional vs. casual, can affect behaviours of the participants. The conclusion of this research was that when speakers are asking for feedback in the institutional setting they most commonly use hand gestures. Finally, we elaborate on how this research can be used and built upon for further studies.

1. Introduction

The distinct modalities present in natural interaction have been studied (Allwood, Cerrato, Jokinen, Navarretta, & Paggio, 2007) to deepen the understanding regarding how humans communicate face-to-face on multiple levels, i.e. when communicating a speaker might move his hands in a certain way, gaze towards the listener while speaking to get a better connection to the listener and to make sure that he is being listened too and understood by the listener. The reason for researching this is to help understand how humans behave while speaking, and what non-verbal expressions they use in communication.

To be able to describe what persons use when they communicate it is necessary to annotate all modalities that are present in a face-to-face communication. This is done so that we can predict how humans would behave in different circumstances. It is also necessary to gather this information to improve spoken human – computer interaction and there is a need for a corpus that can capture human multimodal behaviour not only to guide the development of dialogue systems but also to develop their visual interfaces (Pastra & Wilks, 2004). The reason for doing so is to be able to provide a much richer set of communication channels between the agent and the human (Cassell, Bickmore, Campbell, Vilhjálmsson, & Yan, 2001).

The motivation for this research is to identify and interpret gestures, facial expressions, body posture and speech, which are features that critically contribute to the conversational interaction in an attempt to find evidence about their potential systematic roles. The main contributions of this thesis are:

- Working towards the description and annotation of a multimodal corpus of Icelandic TV interview available for further development and exploitation
- Using the Icelandic corpus to compare multimodality annotated data with similar corpora from a Greek study

The setting in which the interview takes place, as well as the social and discursive roles of the speakers, are features that formulate the discourse structure and further

influence the interactants' conversational behavior in all its expressive dimensions such as speech, hand and facial gestures. There are three main characteristics which have been studied related to TV interviews, depending on different settings of the interview: institutionalized discourse, which is a setting where two speakers, an interviewer and interviewee are talking about a specific topic e.g. political interview under a very formal setting, semi-institutionalized discourse which could be a setting where two speakers are talking more casually about specific matters, e.g. an interview about someone's life or thoughts, and casual conversation (Heritage, 2005) (Ilie, 2001).

In this study we focus on the institutionalized setting, and gather multimodal data from a political TV interview on Icelandic television. This data will be compared to a similar study on a similar interview on Greek television. The initial hypothesis was that we would find a great difference in the frequency of non-verbal expression, based on popular knowledge of these two cultures, but our analysis shows that the frequency of non-verbal expressions is very similar so the hypothesis was not verified. One particular communicative function, which is *feedback elicitation,* was analyzed in depth in the Icelandic data, and in particular we wanted to know what kinds of non-verbal behavior accompany this function. A strong result is that hand gestures are very likely to occur when the speaker is asking the listener to either give him a response or when the speaker is giving the listener a chance to speak.

This thesis is structured as follows: In the next chapter we go over previous work that has been done in this field. In chapter 3 we discuss the Icelandic corpus, the annotation process and the annotation output. In chapter 4 we look into a comparative research which was done between the Icelandic corpus and a Greek corpus and the results from that study. Finally, conclusions and further work are covered in chapter 5.

2. Background

There is a growing demand for intelligent multimodal systems that can seamlessly integrate visual and linguistic data in natural, intuitive communication between humans and machines (Pastra & Wilks, 2004).

In conversation there are two processes that are particularly vital for successful exchange between the participants: Turn taking and feedback. To make sure that there are no overlaps when speakers communicate they must know when their turn is up and not be speaking all at the same time. Turns are requested, taken, held and given, using non-verbal behaviours that are often parallel to speech. These signals can be for example gaze, intonation and gestures (Vilhjálmsson, 2009).

For successful exchange between two persons they need to know whether or not they are being listened to or if the listener is receiving the information and understanding it. The speaker may need to look for or request signs of understanding. If there is a lack of interest from the listener the speaker will discontinue speaking. This calls for attentive dynamic feedback from the listener (Vilhjálmsson, 2009).

Through non-verbal communication the speakers cast their feelings and beliefs towards either their own statements (e.g. confidence about what they say) or to their speaker's statements (MacNeill, 1992).

In a previous study that was done in Chicago the emphasis was on rules for taking speaking turns in conversations. Two interviews were videotaped between an applicant for therapy and a therapist-interviewer. The first 19 minutes were taped and then transcribed for speech and body motion (Duncan, 1972).

According to this study there are six discrete turn-yielding behavioral cues that speakers can pick up on when speaking. Those signals can be displayed together or they may occur either simultaneously or in a tight sequence. To further elaborate what other studies have found out we wanted to see how their results compare to this research. So these six turn-yielding cues are (Duncan, 1972):

- Intonation: the use of any pitch level terminal junction combination at the end of a phonemic clause.
- Paralanguage (prosody) drawl: drawl on the final syllable or on the stressed syllable of a terminal clause.
- Body motion: the termination of any hand gesticulation used during a speaking turn or the relaxation of a tensed hand position during a turn.
- Sociocentric sequences: the appearance of one of several stereotyped expressions, typically following a substantive statement. Examples are “but uh,” “or something“.
- Paralanguage (prosody) pitch/loudness: a drop in pitch and/or loudness in conjunction with one of the sociocentric sequences described above.
- Syntax: the completion of a grammatical clause, involving a subject-predicate combination.

According to Kendon (Kendon, 1967) shifts of gaze are systematically coordinated with the timing of speech and help with synchronizing. If speaker does not look up at the end of an utterance there is a longer pause before the other replies.

From these, we can see that how we perceive and interpret communication incorporates multiple modalities involving the speech level (what is being verbalized) how things are said (intonation, pitch) and the various movements that accompany speech which are i.e. hand gestures, facial displays, and movement of the torso.

For annotation of multimodality we need a specific coding scheme for the different verbal and non-verbal expressions and their communicative functions. In this research we followed the MUMIN coding scheme (Allwood, Cerrato, Jokinen, Navarretta, & Paggio, 2007) which is a multimodal annotation scheme that concentrates on gestures in interpersonal communication, with particular regard to the role played by multimodal expressions for feedback, turn management and sequencing. Their scheme includes descriptions of different attributes, *descriptive* and *functional.* The functional attribute interprets the meaning of the behavior that is being described and they illustrate processes that are crucial for successful communication (Vilhjálmsson, 2009).

The descriptive attributes relate to hand gestures, facial expressions and body posture that each speaker uses when communicating (Vilhjálmsson, 2009).

The MUMIN coding scheme was made by a Nordic Network for Multimodal Interfaces (MUMIN, 2002) and has been tested on the analysis of multimodal behaviour in short video clips in Swedish, Finnish and Danish (Allwood, Cerrato, Jokinen, Navarretta, & Paggio, 2007). It has also been used in a study where two annotated corpora (Danish and Estonian) were used in a promising machine learning experiment (Jokinen, Navarretta, & Paggio, 2008).

Descriptive attributes are features that relate to hand gestures, facial expressions and body posture for each speaker as well as the semiotic type while the functional attributes are features that concern the annotation of multimodal feedback and turn management as well as the relations between speech and non-verbal expressions. Functional attributes are annotated at a higher level of interpretation than the descriptive attributes (Vilhjálmsson, 2009). A table of descriptive and functional attributes can be seen in **Table 1**.

Descriptive attributes			Functional attributes		
Facial display	Hand gesture	Semiotic type	Feedback	Turn management	Multimodal relations
Gaze	Handedness	Deictic	Give	Turn Gain	Repetition
Eyes	Trajectory	Non-deictic	Elicit	Turn End	Addition
Eyebrows		Iconic		Turn Hold	Substitution
Mouth		Symbolic		Sequencing	Contradiction
Lips					Neutral
Head					

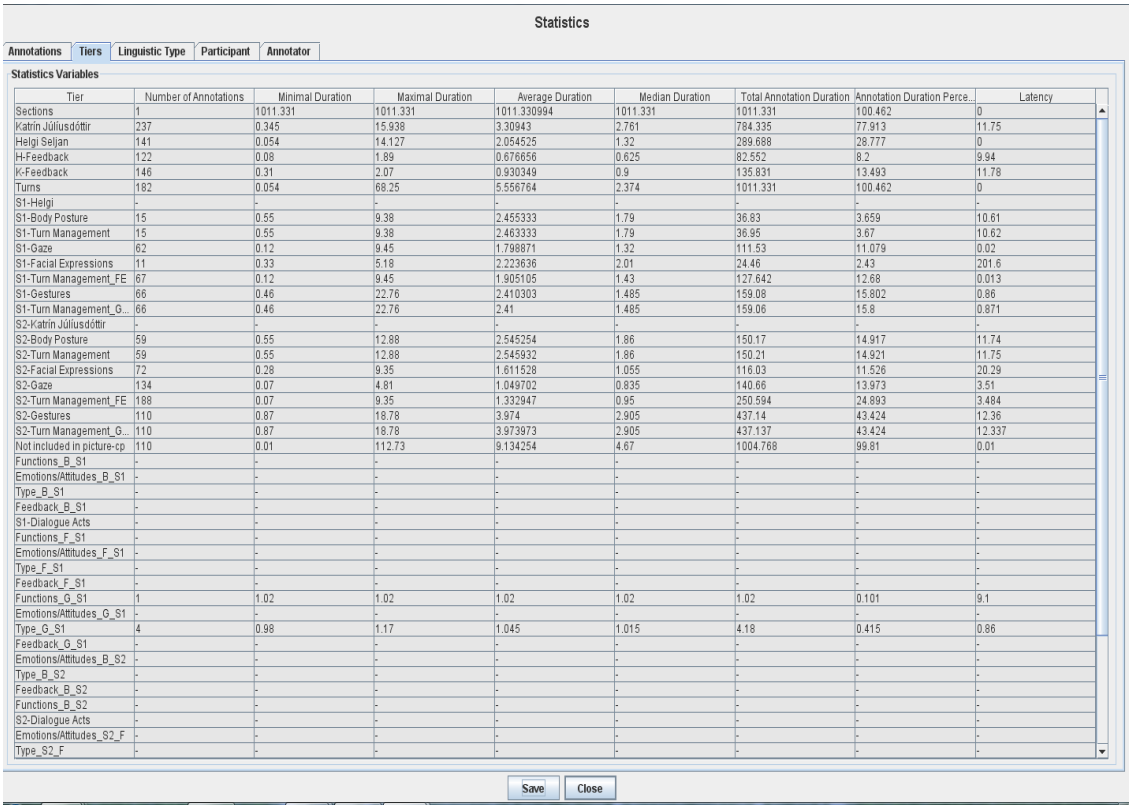
Table 1. MUMIN overall coding scheme

Descriptive attributes are connected directly to what speakers do when they are speaking, how their face, hand and torso move. For facial displays you have gaze, eyes, eyebrows, mouth, lips and head. For hand gestures you have the handedness and trajectory. A more detailed table can be found in **Appendix A**.

Functional attributes describe how the conversation flows for speakers in terms of what they accomplish in the interaction, this includes feedback, whether or not the

speaker elicits feedback and if the listener gives him that feedback, and turn management, which has to do with who holds the turn and how the turn is gained, held or ended.

For the annotation process it was decided to use the tool Elan (Brugman & Russel, 2004). The reason for doing so was that Elan seemed to be user friendly during the annotation process as it offered many functionalities such as copying or importing annotations, merging annotations of two different files into a single and merging annotations from two files in order to check their agreement as well as the *Undo* functionality if a mistake was made. There is also a very good search engine for drawing statistics from either a single or multiple files, as can be seen in **Figure 1**



Tier	Number of Annotations	Minimal Duration	Maximal Duration	Average Duration	Median Duration	Total Annotation Duration	Annotation Duration Perce...	Latency
Sections	1	1011.331	1011.331	1011.330994	1011.331	1011.331	100.462	0
Katrín Júlíusdóttir	237	0.345	15.938	3.30943	2.761	784.335	77.913	11.75
Helgi Seljan	141	0.054	14.127	2.054525	1.32	289.688	28.777	0
H-Feedback	122	0.08	1.89	0.676656	0.625	82.552	8.2	9.94
K-Feedback	146	0.31	2.07	0.930349	0.9	135.831	13.493	11.78
Turns	182	0.054	68.25	5.556764	2.374	1011.331	100.462	0
S1-Helgi	-	-	-	-	-	-	-	-
S1-Body Posture	15	0.55	9.38	2.455333	1.79	36.83	3.659	10.61
S1-Turn Management	15	0.55	9.38	2.463333	1.79	36.95	3.67	10.62
S1-Gaze	62	0.12	9.45	1.798871	1.32	111.53	11.079	0.02
S1-Facial Expressions	11	0.33	5.18	2.223636	2.01	24.46	2.43	201.6
S1-Turn Management_FE	67	0.12	9.45	1.905105	1.43	127.642	12.68	0.013
S1-Gestures	66	0.46	22.76	2.410303	1.485	159.08	15.802	0.86
S1-Turn Management_G...	66	0.46	22.76	2.41	1.485	159.06	15.8	0.871
S2-Katrín Júlíusdóttir	-	-	-	-	-	-	-	-
S2-Body Posture	59	0.55	12.88	2.545254	1.86	150.17	14.917	11.74
S2-Turn Management	59	0.55	12.88	2.545932	1.86	150.21	14.921	11.75
S2-Facial Expressions	72	0.28	9.35	1.611528	1.055	116.03	11.526	20.29
S2-Gaze	134	0.07	4.81	1.049702	0.835	140.66	13.973	3.51
S2-Turn Management_FE	188	0.07	9.35	1.332947	0.95	250.594	24.893	3.484
S2-Gestures	110	0.87	18.78	3.974	2.905	437.14	43.424	12.36
S2-Turn Management_G...	110	0.87	18.78	3.973973	2.905	437.137	43.424	12.337
Not included in picture-cp	110	0.01	112.73	9.134254	4.67	1004.768	99.81	0.01
Functions_B_S1	-	-	-	-	-	-	-	-
Emotions/Attitudes_B_S1	-	-	-	-	-	-	-	-
Type_B_S1	-	-	-	-	-	-	-	-
Feedback_B_S1	-	-	-	-	-	-	-	-
S1-Dialogue Acts	-	-	-	-	-	-	-	-
Functions_F_S1	-	-	-	-	-	-	-	-
Emotions/Attitudes_F_S1	-	-	-	-	-	-	-	-
Type_F_S1	-	-	-	-	-	-	-	-
Feedback_F_S1	-	-	-	-	-	-	-	-
Functions_G_S1	1	1.02	1.02	1.02	1.02	1.02	0.101	9.1
Emotions/Attitudes_G_S1	-	-	-	-	-	-	-	-
Type_G_S1	4	0.98	1.17	1.045	1.015	4.18	0.415	0.86
Feedback_G_S1	-	-	-	-	-	-	-	-
Emotions/Attitudes_B_S2	-	-	-	-	-	-	-	-
Type_B_S2	-	-	-	-	-	-	-	-
Feedback_B_S2	-	-	-	-	-	-	-	-
Functions_B_S2	-	-	-	-	-	-	-	-
S2-Dialogue Acts	-	-	-	-	-	-	-	-
Emotions/Attitudes_S2_F	-	-	-	-	-	-	-	-
Type_S2_F	-	-	-	-	-	-	-	-

Figure 1. A statistical summary provided by Elan after multimodal annotation of video data

The output from Elan is in XML format which is easy to import into other tools. Modifications can be made along the way, adding or deleting new elements, new values etc. directly from the interface in Elan. For future work there is a possibility in Elan to extend the annotation scheme by adding linguistic information such as part-of-

speech tagging and syntactic parsing and synchronizing it with other components within either time or unit boundaries.

The entire coding scheme with the acronyms for each annotation that was used for this research can be seen in **Appendix A.**

2.1 Descriptive annotation categories

The annotation of facial displays involves looking into the timed changes in eyebrow position, movements of the mouth, head and eyes (Cassell, 2000) as seen in **Table 2**. Facial displays can be characterized by the movement of muscles or part of the body that is shifting or the amount of time they last but they can also be characterized by their function in the conversation (Allwood, Cerrato, Jokinen, Navarretta, & Paggio, 2007). The reason for having two separate features for eyes and gaze is because gaze refers to the eye movement we make in the general direction of another's face and is not relevant if the speaker is just looking at nothing in particular (Knapp & Hall, 2002). Gaze as a behavior may be used to manage turns and for receiving feedback or giving feedback so it regulates the flow of the conversation. For example when persons look into each other eye area mutual gaze occurs (Knapp & Hall, 2002).

Most studies of facial expressions have concerned themselves with various emotional states such as anger, sadness, surprise, happiness, fear and disgust. Since facial expressions also function as regulatory gestures they provide feedback and manage the flow of the interaction. Many researchers believe that the primary function of the face is to communicate, not to express emotions (Knapp & Hall, 2002).

For facial display features see **Table 2**.

Facial display feature	Form of expression	
Tier	**Value**	**Annotation in Elan**
General face	smile	Smile
	Laughter	Laugh
	Scowl	Scowl
Eyebrows	Frowning	B_F
	Raising	B_R
Eyes	Exaggerated Opening	Eye_EO
	Closing-both	Eye_CB
	Closing-one	Eye_CO
	Closing-repeated	Eye_RC
Gaze	Towards interlocutor	Gaze_Tow
	Up	Gaze_Up
	Down	Gaze_Down
	Sideways	Gaze_Side
Mouth - Openness	Open mouth	Mouth_O
	Closed mouth	Mouth_C
Mouth - Lips	Corners up	
	Corners down	
	Protruded	Lips_Pro
	Retracted	Lips_Ret
Head	Single Nod (Down)	Head_N
	Repeated Nods (Down)	Head_RN
	Single Jerk (Backwards Up)	Head_SJ
	Repeated Jerks (Backwards Up)	Head_RJ
	Single Slow Backwards Up	
	Move Forward	Head_MF
	Move Backward	Head_MB
	Single Tilt (Sideways)	Head_ST
	Repeated Tilts (Sideways)	Head_RT
	Side-turn	Head_Turn
	Shake (repeated)	Head_Shake
	Waggle	
Unidentified		Eye_SC
		Gaze_Un

Table 2. MUMIN facial display coding scheme

As with facial displays hand gestures can be feedback related and have turn managements functions as well as giving meaning and contribute to the

communication. There are two dimensions that are looked at when it comes to the shape of the gesture, *Handedness* and *Trajectory*. See **Table 3**.

Gestures	**Shape of gesture**			
Tier		**Short tag**	**Annotation in Elan**	**Description**
Hand gestures	**Handedness**	Both-H both hands	BH	
		Single-H single hand	SH	
		fingers_pointing	F_P	
		Fingers tapping	F_T	
		Shoulders	Shoulders_UP	
	Trajectory	Up	BH_U / SH_U	Both hands up/ Single hand up
		Down	BH_D / SH_D	Both hands down/ Single hand down
		Sideways	BH_Side/SH_Side	Both hands sideways/Single hand sideways
		Complex	BH_complex/SH_complex	Both hands complex/Single hand complex
		Repeated	SH_R_U_D	Single hand repeated up and down
			BH_R_U_D	Both hands repeated up and down
		Other	S_H_O/B_H_O	Single hand other/both hands other

Table 3. Gesture coding scheme

Handedness refers to whether or not one or both hands are being used and the trajectory to the movement itself i.e. whether the hands are moving up, down, sideways etc.

In the MUMIN coding scheme the only types of gestures that are taken into consideration are hand gestures and facial displays but as they state in their research the body posture is also relevant and important when annotating multimodal

communication behaviour (Allwood, Cerrato, Jokinen, Navarretta, & Paggio, 2007). In this research the body posture was annotated according to **Table 4**.

Body Posture	Shape of gesture
Torso	Torso Bend Forward
	Torso Bend Backwards
	Torso Turn Right
	Torso Turn Left
	Torso Lean Left
	Torso Lean Right

Table 4. Body posture coding scheme

When annotating hand movements it is important to know what these movements stand for. If the movement doesn't have intended communicative meaning i.e. speaker scratching his nose then that movement was annotated as an adaptor. Most speakers' hand movement contributes to the communication, but they have different roles. According to David McNeil (MacNeill, 1992) the communicative hand movements fall into several categories. A closer look can be seen in **Figure 2**.

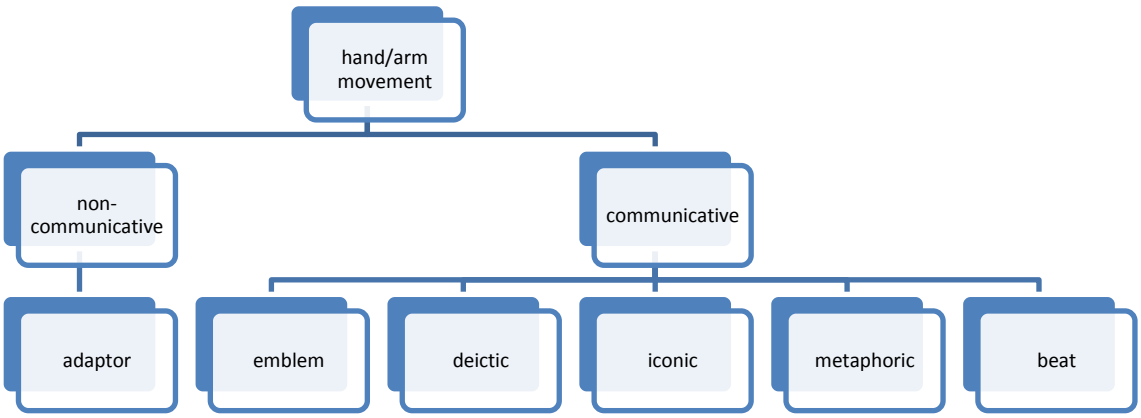


Figure 2. Role of gesture

Here below is a detailed description of these different types of hand movements and what their role in the communication is. These descriptions are based on Michel Kipp's

work on automated gesture generation for virtual humans based on imitation (Kipp, 2004).

Adaptors

Adaptors are movements that have no meaning like scratching your neck or playing with a pen (Kipp, 2004). They can however indicate the speaker's state of mind, if someone is fidgeting a lot with papers in front of them it indicates that the speaker is feeling uncomfortable and that can have an effect on the conversation.

Emblems

Signals that have a meaning all by themselves and can be used in the absence of speech to convey the meaning are called emblems. These signals can be thumbs up or the OK sign. They are often used when speakers need to communicate over loud sounds or when the channel is constricted in some way. Emblems are most often culture specific (Kipp, 2004).

Deictics

Deictics are pointing movements that speakers use to point at a specific item, person, location or direction or it can be pointing at an abstract thing, imaginary thing or a concept (Kipp, 2004).

Iconics

Iconics can be seen in a conversation when the speaker makes some kind of a movement of the hand which illustrates what is being said by making some kind of form or a thing. An example of this is when a speaker is describing a box and he makes a box shape in the air to illustrate what the box looked like. They cannot be emblems because they are made at the spur of the moment (Kipp, 2004).

Metaphorics

Metaphorics are gestures that are similar to iconics in the sense that they are also gestures to illustrate what is being expressed. The difference between an iconic

gesture and a metaphoric gesture is that they represent an abstract feature concurrently spoken about. A typical metaphoric hand movement describes an abstract concept i.e. love, happiness etc. (MacNeill, 1992).

Beats

Beats are movements that are rhythmical and accompany speech and have correlation to the meaning of what is being said as well as intonation. The connection to speech is first and foremost regarding the timing of the beat movement not the form of the movement itself. They can be used to emphasize what is being said (Kipp, 2004).

2.2 Functional annotation categories

In the MUMIN scheme, three types of communicative functions are being annotated: feedback, turn management and sequencing (Allwood, Cerrato, Jokinen, Navarretta, & Paggio, 2007). Of those functions, we focus on feedback and turn management. Feedback is something that is always apparent in human communication. Conversation partners exchange feedback constantly to provide a way of showing if the interaction is a success or failure (Allwood, Cerrato, Jokinen, Navarretta, & Paggio, 2007). Successful interaction could be if the speaker is asking for feedback on something he said and the listener nods or agrees with him. In case were the listener fails to hear the speaker or doesn´t understand what the speaker is talking about the speaker doesn´t get the feedback he asked for and can therefore repeat what he said or rephrase his sentence. Giving feedback shows that you have heard and understood what is being said and eliciting feedback means that you want to see if the listener has understood what you just said.

Turn management controls how the interaction takes place between speakers and minimizes overlapping speech and pauses. *Turn take* is when a speaker takes over turn that wasn´t offered to him, like for example he interrupts the other speaker, *turn accept* is when the turn is offered to the listener and he accepts it. *Turn end* is annotated if the speaker was interrupted and he decides to release the turn and *turn yield* if the speaker releases the turn under pressure. *Turn offer* is when the speaker offers the turn to the listener and *turn complete* if the speaker signals completion of

the turn and end of the dialogue at the same time (Allwood, Cerrato, Jokinen, Navarretta, & Paggio, 2007)

As seen in **Table 5** there are three general turn management functions coded in this research, *turn gain*, *turn keep* and *turn end,* that further break down into specific functions. For *turn gain* the function value is *turn grab* and *turn accept*. The function type *turn keep* has the function value *turn keep.* For *turn end* it is *turn yield, turn give and turn complete.*

Turn Management	Function
Turn Gain	Turn Grab Turn Accept
Turn Keep	Turn Keep
Turn End	Turn Yield Turn Give Turn Complete

Table 5. Turn management coding scheme

In the picture **Figure 3** below the interface for Elan is shown. To annotate e.g. turn management there needs to be a main level or tier and underneath that a more precise function. To show how the tiers work the tier *turn management* was opened and it can be seen that it includes TT which is *turn taking,* TA which is *turn accept* etc. The entire table can be seen in **Appendix A**.

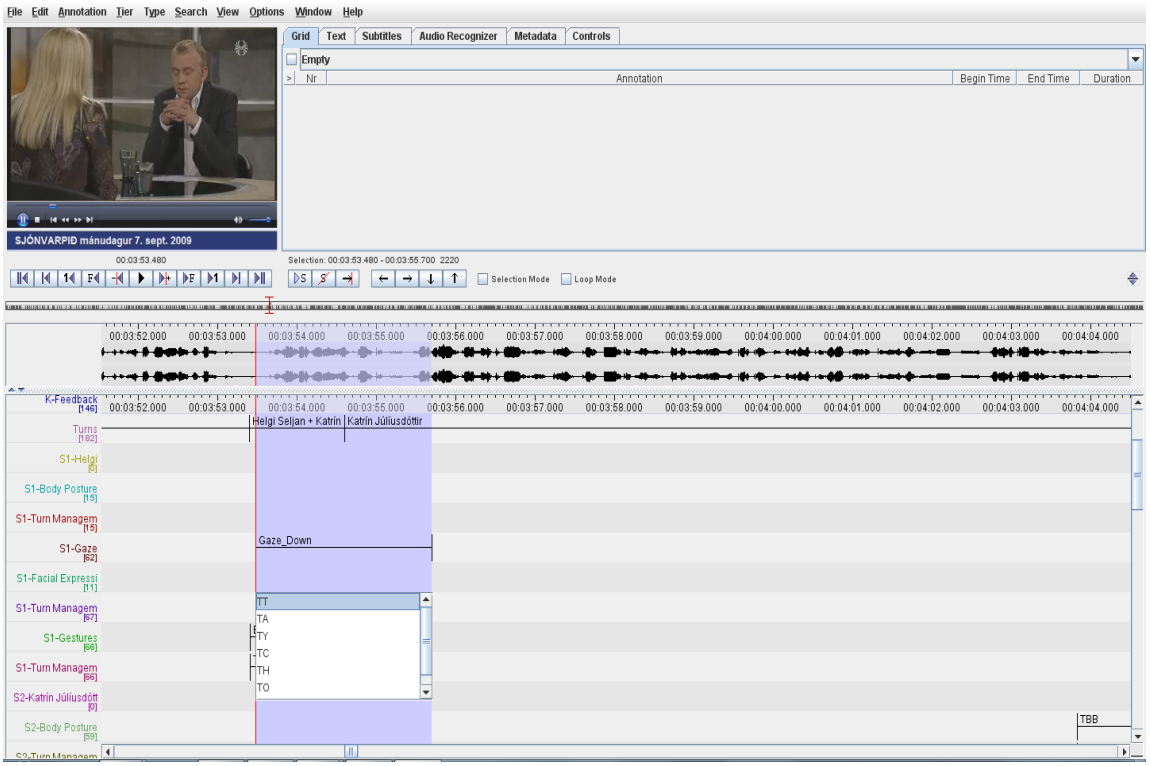


Figure 3. Turn management functions are annotated by marking a time interval, going to the right tier and selecting from a list of available functions.

2.3 Summary

For fully describing what occurs during face-to-face conversation, one has to annotate multimodal behaviour, both in terms of the surface form (descriptive attribute) and in terms of what the behaviour accomplishes (functional attribute) (Vilhjálmsson, 2009). Descriptive attributes are facial displays, hand gestures and body posture while the functional attributes are feedback and turn management.

For the audio transcriptions we used *Transcriber* and for the non-verbal expressions the tool *Elan* was used. The coding scheme which was mostly followed was MUMIN v.3.3.

3. Icelandic Corpus Description

In this research special focus was on turn taking since the turn taking mechanism allows participants to manage the smooth and appropriate exchange of speaking turns in face-to-face interaction. Turn taking is expressed actively through content (words),

intonation, prosody, gaze and gestures. The multimodal analysis of turn-taking provides significant information and more accurate observations than if one were to study the speech by itself. Non-verbal expressions do not simply accompany speech but they are indicators of the degree of success of the speakers´ intentions and projections and shed light on the strategies used for the accomplishments of the interaction. They are signals by which each participant indicated his state with regard to the speaking turn (Duncan, 1972).

The corpus consists of 16 minutes and 46 seconds of a political interview where the host is a male journalist and the guest is the Minister of Industry, a female. The topic is regarding aluminum smelters which are a highly debated topic at that current moment.

The interview is extracted from the website www.visir.is[1] where the interview was available online but initially it comes from the Icelandic TV station RÚV. The interview was recorded using Camtasia[2], which is a tool used for recording straight from the computer screen. The structure of the interview consists of a host controlled question-answer sequences.

We chose the TV interview so that we could compare our annotations with research done in Greece on similar data. It is important to realize that the data we are using is not from casual conversation but from an institutionalized situation. This situation in interviews means that the interview is expected to take place in a particular setting which is monitored by a whole team of professional people with the help of technical devices. An institutionalized interview has the following constraints:

- *Time restriction* ''the discussion is monitored and periodically interrupted by the host for commercial breaks'' (Heritage, 2005).
- *Speaker selection restrictions* ''the speaker has restricted time to answer'' (Heritage, 2005).

[1] http://vefmidlar.visir.is/VefTV/
[2] http://www.techsmith.com/camtasia/

- *Turn taking restrictions* ''the host is primarily responsible for selecting the next speaker and for orchestrating the turn taking sequences'' (Heritage, 2005).

Therefore, there is an asymmetrical role distribution where the host has a role that they self assume, which is to lead the interview, whereas the guest assumes a role that has been assigned to them. There is asymmetry in the speakers´ interactional rights and obligations such as asking questions, making statements, interrupting and addressing each other, and so on (Heritage, 2005).

The interview can be regarded as an institutionalized interaction, as it appears to be standardized in role distribution and turn management predictability because it appears to be constrained by institutional role distribution and turn pre-allocation and is not prone to spontaneous interventions. Our results therefore pertain to this kind of situation and should not be generalized across all kinds of conversations.

The recorded video, along with its audio, was imported into Elan and annotated.

3.1 Annotation Process

First step was to get acquainted with the tools that can be used to annotate multimodality in corpora. These tools are for example Praat[3] and Transcriber[4] (Boudahmane, Manta, Antoine, Galliano, & Barras, 1998) for annotation of speech. For annotation of descriptive and functional attributes of video recorded data there are two tools which were looked at: Elan (Brugman & Russel, 2004) and Anvil[5].

One of the goals of this study was to be able to compare corpora between different cultures and because of that the decision was made to use the same tools as in a previous study which was done in Greece (Koutsombogera & Papageorgiou, 2009), mostly to simplify the comparison between those two studies. Those tools are Elan and Transcriber. The video file was imported from the tool Camtasia in the Windows Media Audio/Video file (.wmv) format over to Elan, the tool that was used for the entire video annotation, and then the audio signal was extracted from the relevant

[3] http://www.fon.hum.uva.nl/praat/
[4] http://trans.sourceforge.net/en/presentation.php
[5] http://www.anvil-software.de/

video and loaded into Transcriber. There it was orthographically transcribed but filled pauses, which are sounds like “mmm”, “aha” and so on, were also transcribed, along with speech mistakes. Care was taken to ensure that the speech transcript was kept synchronized with the video.

Finally the triplet [.wav, .mpeg, .trs files] were imported to the ELAN video annotation tool in order to start the annotation. The coding scheme was created directly from Elan by inputting the features and the values we wanted. In whole there are 1995 annotations in the corpus which do not include the transcription of the speech.

Facial, hand and body gestures of interest were identified marking their start and end points and then annotated according to their characteristics (e.g. head nod, single hand up etc.) The levels and labels used in the annotation scheme are mainly inspired by the MUMIN coding scheme notation which can be seen in **Appendix A**.

In the MUMIN scheme the tier for Facial display feature included gaze a long with general face, eyebrows, eyes, mouth and head. It became clear that gaze often co-occurred with eyes, eyebrows or head so it was decided to have a separate tier for gaze. An example of this can be seen in **Figure 4** where a head movement, single tilt to the side, collides with gaze to the side.

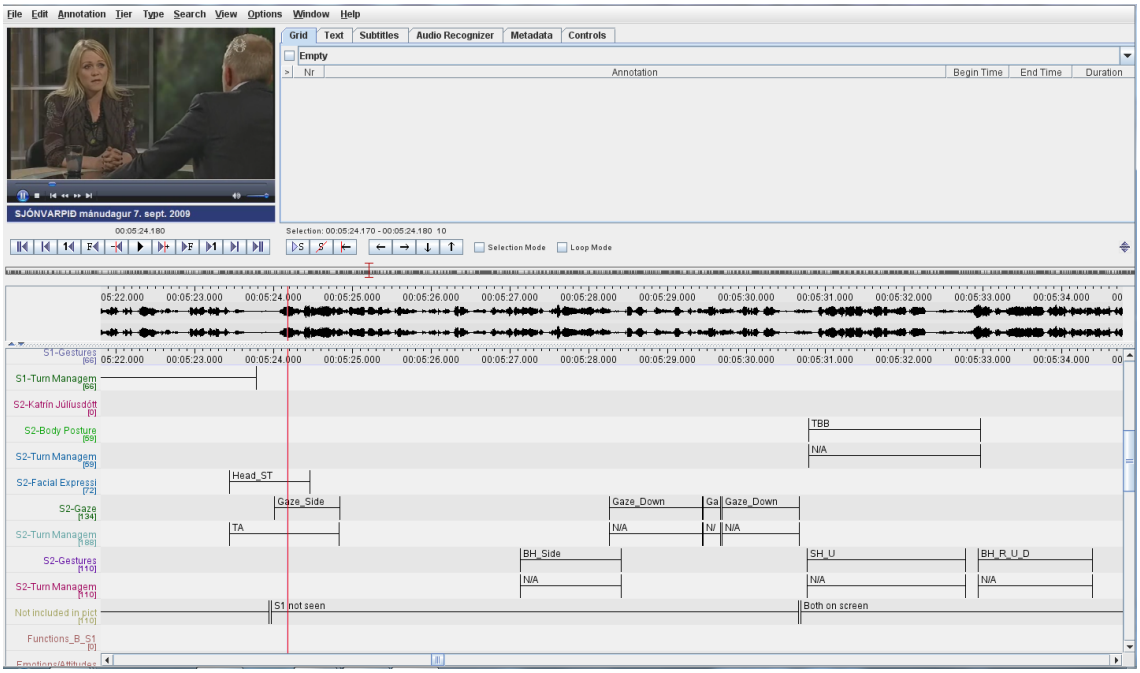


Figure 4. Gaze seperate from facial display

Every movement of both the interviewer and interviewee was annotated and the rule was to start marking the place where the action started and make it end when the movement had ended.

The annotation process was done in these steps:

1. It was decided to start the annotation by marking the descriptive features. First step was to view the video without any sound, and annotate every hand movement, whether or not only one hand was moved or both and how the trajectory was. The annotation started from the beginning of the movement until the hand was back into original position or grounded. There was a problem regarding who was seen on the screen so it was annotated who was seen at each time.
2. Then all head movements were annotated, whether they were nods or just the head turning to the sides.
3. All movements of the eyebrows were annotated, when they were raised so the annotation started at the beginning of the movement until they had been lowered back into their original position.

4. At the end of annotating the descriptive features it was decided to focus only on the transcription of the speech. The audio file was extracted from the video into a *wav* file and then imported into the tool Transcriber (Boudahmane, Manta, Antoine, Galliano, & Barras, 1998) and then orthographically transcribed as well as all hesitations, feedback and overlaps in speech. It was ensured to identify all pauses by marking it in the Transcriber by making a new line in the text, this was to make it easier to identify feedback elicit, pauses, filled pauses or feedback given by only listening to the audio and seeing how it would look like when being imported into Elan (Brugman & Russel, 2004).
5. Next step was then to annotate the functional attributes which are the forms of turn management. They include *turn take* and *turn accept*, *turn yield*, *turn offer*, *turn complete* and *turn hold*. Those were annotated with the descriptive attributes which are i.e. movements of hands, facial displays and body posture in the same timeframe.
6. The last step was to import the audio file and the transcription without any other annotations into Elan to annotate feedback elicitation, the giving of feedback, pauses and filled pauses. Annotating these in isolation from other non-verbal annotation was done to avoid possible bias in favor of correlation. If correlations are found, we can trust that they naturally arise from the data and are not engineered by the annotator. Feedback elicit was annotated if the pitch or intonation at the end of a sentence changed, feedback give was annotated if the listener gave feedback by saying either “mmm”, “aha” or started speaking. Pause was annotated if the speaker stopped for a short while to think or to breathe and the listener gave feedback or took over the conversation and filled pauses were annotated if the speaker started hesitating but kept talking or said “uuuuhhh” or repeated the same word and the listener started to speak.

After this annotation all the information that had been gathered and annotated was finally combined and imported into completed single Elan file for further analysis on correlations.

Both audio and visual signals as well as the annotations are perfectly synchronized: the overall set of annotation levels is distributed according to each speaker, and all the information is integrated into a single XML file for future use and to be able to continue comparisons with other countries.

3.2 Research on the annotation

The study of the annotated corpus reveals the multiple functions of non-verbal communication and how it interacts with verbal signals. To be able to interpret the non-verbal behavior we have to look at the situation that the speakers find themselves in, the context and the role of the participants during interaction. Since non-verbal expressions coordinate interaction they convey information such as when the speaker should take or give turn (Ekman, 1999).

In this interview the setting is institutionalized and the speakers only have the floor for certain amount of time. It is therefore likely that they are more prone to taking the turn rather than offering it, unless the host is asking a direct question that he wants an answer to.

In the interview there were verified 511 non-verbal expressions. The speakers assign multimodal expressions which can be simple or more complex by using their facial characteristics (gaze, eyebrows, nods etc.), hand gestures (single or both hands, fingers, shoulders etc.) and upper part of the body (leaning forward and backward) in order to manage the interaction process and ensure proper uptake of information.

In this corpus **Appendix B** there are visible signals which can be related to different forms of turn management. In *turn taking* the most preferred form for the host is *finger pointing,* where the host is pointing in the direction of the guest, in 16,1% of gesture expressions or 5 out of 31 gestures that occurred with turn taking. For the guest it is gaze down in 44,4% or 4 out of 9 instances of facial expressions. For *turn accept* it is *finger pointing,* where the host is pointing in the direction of the guest again, in 75% or 6 out of 8 gesture expressions and *gazing down* for the guest in 42,8% or 9 out of 21 facial expressions. For *turn yield* it is *gaze down* and *gazing towards* interlocutor for the host, which are the only face expressions he uses but *eyebrow*

raise for the guest in 60% or in 3 out of 5 facial expressions. For *turn offer* it is *single hand repeated up and down* and *both hands sideways* for the host in 57% or 4 out of 7 gestures, but *eyebrow raise* for the guest in 62,5% or 5 out of 8 facial expressions. For *turn hold* it is *gaze down*, *gaze towards interlocuto*r and *gaze to the side* for the host which are the only face expressions he uses but *gaze down* for the guest in 55,5% or in 20 out of 36 facial expressions as according to **Table 6**.

Icelandic Interview		
	Host	Guest
Turn Take	Finger pointing	Gaze Down
Turn Accept	Finger Pointing	Gaze Down
Turn Yield	Gaze Down, Gaze Towards	Eyebrow raise
Turn Offer	Single hand repeated up and down, Both hands sideways	Eyebrow raise
Turn Complete	-	-
Turn Hold	Gaze Down, Gaze Towards, Gaze Side	Gaze Down

Table 6. Turn management of the host and the guest compaired.

It is interesting to see that neither the host nor the guest are able to complete the turn, they are fighting to keep their turn and are either successful or they yield. For fully detailed description see **Appendix B**.

Speakers that do not want to be distracted or disturbed particularly look away at the beginning of an utterance while planning it (Kendon, 1967). As seen in our data much of the turn management is accomplished by gaze, so this reinforces the claim made by Kendon.

Figure 5 shows an example of the guest offering the turn to the host and as she gazes down she raises her eyebrows. When he takes the turn he starts making gestures and gazing down. S1 stands for the host and S2 for the guest.

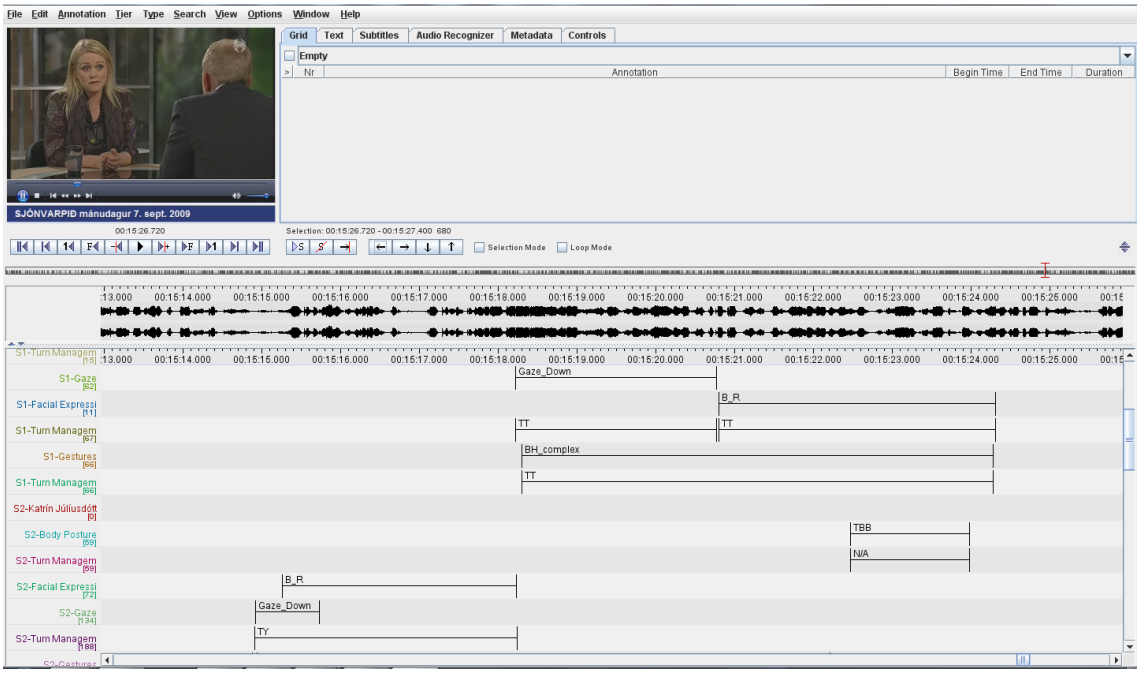


Figure 5. Turn management for S1 (the host) and S2 (the guest). The guest yields her turn and gazes down and raises her eyebrows while the host takes the turn and starts gazing down and moving both his hands.

Table 7 presents the distribution of the non-verbal expression involved in the turn management communicative function in the interview. The overall number of non-verbal expressions performed by each speaker is shown in the row labeled *NVEs*. The *N/A* row contains non-verbal expressions that are not related to turn management but to other communicative functions such as feedback expression, emotions or attitude expression, content-related NVEs (e.g. iconic gestures) and so on. The *TM NVEs* row depicts the number of NVEs related to Turn Management.

According to these numbers the host´s preferred modality of non-verbal turn management are gestures (52) while the guest prefers the facial expressions (81). It is also obvious that the guest is more productive in terms of non-verbal expressions with 139 TM NVEs while the host has 112 TM NVEs.

Another interesting point is that the host shows most non-verbal expressions for turn take (46,4%) while the guest shows them to hold the turn (58,3%).

Number of non-verbal expressions attested in each interview and distribution between host and guest can be seen in **Table 7**.

Icelandic Interview										
	Host					Guest				
	gesture	face	body	Total	%	gesture	face	body	Total	%
total NVEs	66	73	15	154		*110*	*188*	*59*	357	
N/A	14	20	3	37		*75*	*106*	*35*	216	
TM NVEs	52	48	12	112		*35*	*82*	*24*	141	
Turn Take	31	29	8	68	60,7%	4	11	3	18	12,7%
Turn Accept	8	6	1	15	13,4%	7	21	6	34	24,1%
Turn Yield	3	4	1	8	7,1%	3	5	1	9	6,4%
Turn Offer	7	6	2	15	13,4%	3	8	3	14	9,9%
Turn Complete				0	0,0%				0	0,0%
Turn Hold	3	3		6	5,4%	18	37	11	66	46,8%
	46,4%	42,9%	10,7%			*24,8%*	*58,1%*	*17,0%*		

Table 7. Distribution of non-verbal expressions. In the table the largest numver in a category is highlighted in red.

When finding the correlation between *feedback elicit* and *gestures* there were many interesting findings as seen in **Figure 6**.

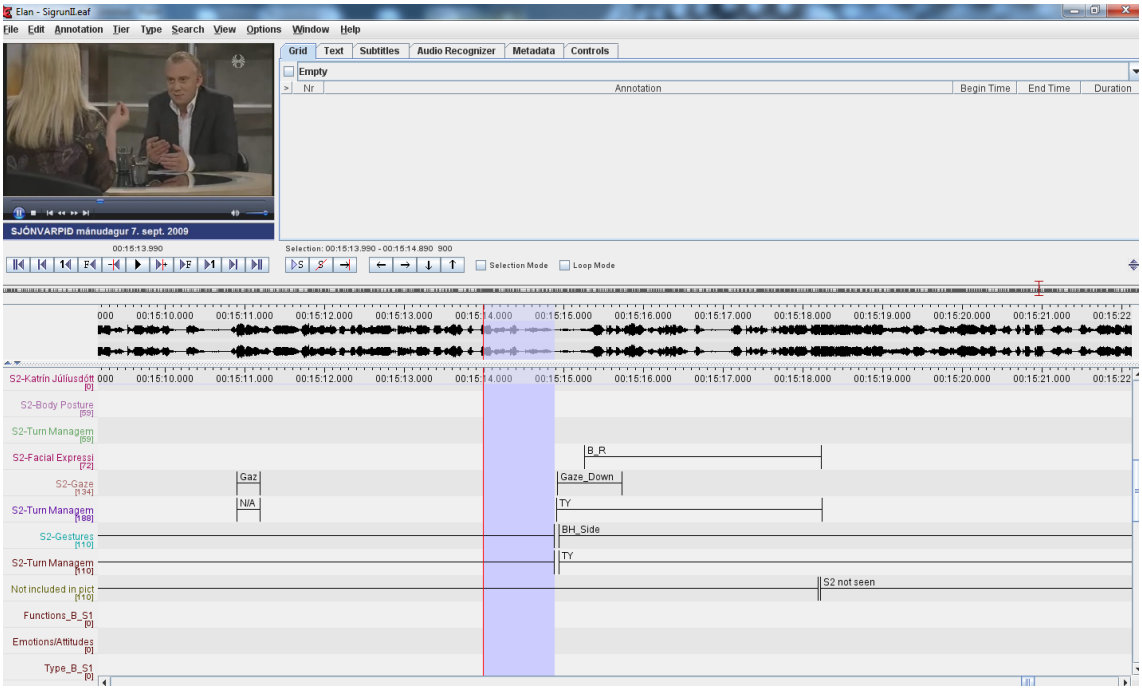


Figure 6. Feedback elicit. The tier feedback elicit is highlighted and it shows that at the same time S2 (the guest) stops her gestures, she waits for a reply from the host and then continues with new gestures and gazes down.

For example, the guest would ask for feedback from the host and as can be seen in **Figure 6**, she stopped making the hand gesture, waited for the response and then continued with more gestures when continuing with her speech.

To get a better view of which non-verbal behavior are most commonly used for eliciting feedback the feedback function tier was highlighted and all descriptive attributes for behaviors that fell into that highlighting were counted but only for the speaker that was asking for the feedback.

The conclusion was that the host used gestures to convey that he wanted feedback. The same thing was obvious for the guest. Closer look can be seen in **Table 8**.

	FBE H	FBE G	Pause H	Pause G	FillH	Fill G
Gesture	21	38	2	21	4	3
Body	5	16		3	1	2
Gaze	10	12	1	6	3	1
Face	7	21		5	2	1
N/A	14	11		5	1	1

Table 8. Feedback elicit, pauses and filled pauses for the guest and the host.

FBE H stands for feedback elicit for the host, FBE G is feedback elicit for guest. Fill is filled pause. According to this gestures are the most frequent non-verbal expression when speakers are asking for feedback on what is being said. It should be noted though that most of the observed non-verbal expressions during the FBE is a part of ongoing non-verbal expressions activity of the current speaker.

When looking in more detail at the gestures that were being used, hand movements, single hand up, down, complex etc. and both hands sideways, complex and repeated up and down, were the most frequent ones. Often the hand movement was long and did not seem to be in correlation with the feedback elicit, pause or filled pause but there were hand movements that were in the same time slot as feedback elicit.

To have more detailed information about how often the non-verbal expressions correlated in the same timeslot as the functions, the non-verbal expressions that correlated with the same timeslot as feedback elicit, filled pause or pause were counted. These numbers can be seen in **Table 9**. A more detailed table can be seen in **Appendix C**.

	FBE H	FBE G	Pause H	Pause G	FillH	Fill G
Gesture	11	14	0	4	1	3
Body	2	4		0	0	2
Gaze	8	10	0	3	2	1
Face	3	13		1	0	0

Table 9. Descriptive attributes that correlate precisely with the functions.

What was interesting was that for the host there were eight instances of single hand movements that were most often used for feedback elicit but there were also four instances of finger pointing also when asking for feedback. Those gestures were in most cases beats, where the host was emphasizing what he was saying. When he paused there were two gestures that occurred at the same time but those were single hand sideways and both hands sideways.

When looking at K, or the guest, she only had one instance of single hand movement connected directly with feedback elicit. The most frequent directly connected gesture for feedback elicit for the guest is moving both her hands in a complex manner or in 5 instances but in whole we counted eleven *both hand* gestures. There are three instances of filled pause with both hands either complex or repeated up and down. This can all be seen in detail in **Appendix D**.

3.3 Summary

The corpus consists of 16 minutes and 46 seconds of a political interview where the host is a journalist and the guest is the Minister of Industry. The topic is regarding Aluminum Smelters which is a highly debated topic at that current moment. The setting of the interview is in a studio and the broadcast is live. In the data preprocessing the face-to-face interviews were multimodally annotated. There were 1995 annotations in whole for the corpus. In the interview the audio signal is transcribed. The subsequent video annotation deals with the labeling of the non-verbal expressions with special attention to turn management and feedback elicitation and feedback giving.

Turn-taking is expressed actively through words, prosody and non-verbal expressions. Since this is an institutionalized interaction there are certain restrictions on the speakers such as time restriction, speaker selection restriction and also turn taking restriction. The interview is host-controlled and there is predictability in turn management. The speakers also have asymmetrical roles whereas the host has the power to assign turns while the guest has to obey to the rules of the interview. This will impact the kind of turn-taking behaviour seen in our data. After careful annotation

and analysis, we can see that neither the host or the speaker are able to complete their turn, they are always fighting to keep the turn but the host showed most non-verbal expressions when taking the turn while the guest showed most non-verbal expressions to hold the turn.

The preferred modality of non-verbal turn management attributes are gestures for the host but facial expressions for the guest. The guest is more productive when it comes to non-verbal expressions.

According to the non-verbal expressions that are being used when eliciting feedback, gestures are the most frequent non-verbal expressions. For the host there were eight instances of single hand movements that correlated exactly in the same timeslot as feedback elicit. For the guest the most frequent gesture was using both her hands in the same timeslot for feedback elicit.

The conclusions from this research are not to be taken as generalization regarding gaze or the turn taking functions but they do reinforce previous studies regarding these functions. It is vital to get more data to increase the corpus so that the forms could be compared between different types of people and to be able to compare our culture with other cultures and last but not least to be able to have a human-agent conversation with and Icelandic agent.

4. Icelandic and Greek corpora

The motivation for this research was to study the multimodal behavior in interview data in different languages and cultures, and compare the findings in order to investigate similarities or differences. For the comparison, we focused on the non-verbal expression of turn taking i.e. regulatory non-verbal expressions as attested in institutionalized political interviews.

Our initial hypothesis was that multimodal behaviors between the two languages would be different because of cultural differences. Comparative studies of non-verbal expressions have been carried out in relation to speakers belonging to different ethnic

groups, social groups and speakers whose language is of different structure (Kendon, 2004) (McNeill & Duncan, 2000) (Kita & Özyürek, 2003)

This study goes further because multimodality between two countries Greece and Iceland is compared within institutionalized interviews for a more precise comparison.

The following chapters deal with the comparison of the Icelandic and Greek data and then our interpretation of the comparison.

4.1 Comparing the Icelandic and Greek data

In our study our data share a common basis with regards to the communicative dimension i.e. turn–taking. The settings of both are live TV political interviews in a studio. As previously stated those are institutional types of interaction. The identity of the participants is identical: journalist versus politician. The topics of discussion are completely different.

The data from both studies were preprocessed in the same manner to be able to compare the two different countries. The duration of the Icelandic interview is 16 minutes and 46 seconds while the Greek one is 16 minutes and 35 seconds.

Table 10 presents the distribution of the non-verbal expressions employed for the turn management communicative function in both interviews. The total non-verbal expressions row (labeled total NVEs) contains the overall non-verbal expressions performed by each speaker. The *N/A* row contains non-verbal expressions that are not related to turn management but to other communicative functions such as feedback expression, emotions or attitude expression, content-related *NVEs* (e.g. iconic gestures) and so on. The *TM NVEs* row depicts the number of NVEs related to Turn Management.

As explained above *gestures* are hand movements, *face* includes eye movements, head movements, gaze, mouth and general face. A closer look can be seen in **Appendix A**. Main conclusions can be seen in **Table 10.**

	Host									
	Greek					Icelandic				
	gesture	face	body	total	%	gesture	face	body	total	%
Total NVEs	133	166	41	*340*		66	73	15	*154*	
N/A	82	92	19	*193*		14	25	3	*42*	
TM NVEs	51	74	22	*147*		52	48	12	*112*	
Turn Take	23	22	6	*51*	*34.7*	31	29	8	*68*	*60.7*
Turn Accept	3	11	4	*18*	*12.2*	8	6	1	*15*	*13.4*
Turn Yield	-	1	-	*1*	*0.7*	3	4	1	*8*	*7.1*
Turn Offer	19	27	9	*55*	*37.4*	7	6	2	*15*	*13.4*
Turn Complete	3	12	3	*18*	*12.3*	-	-	-	*0*	*0*
Turn Hold	3	1	-	*4*	*2.7*	3	3	-	*6*	*5.4*
%	*34.7*	*50.3*	*15*			*46.4*	*42.9*	*10.7*		
	Guest									
	Greek					Icelandic				
	gesture	face	body	total	%	gesture	face	body	total	%
Total NVEs	158	421	63	*642*		110	188	59	*357*	
N/A	137	320	48	*505*		75	106	35	*216*	
TM NVEs	21	101	15	*137*		35	82	24	*141*	
Turn Take	3	4	-	*7*	*5.1*	4	11	3	*18*	*12,7*
Turn Accept	6	52	5	*63*	*46*	7	21	6	*34*	*24.1*
Turn Yield	-	2	-	*2*	*1.5*	3	5	1	*9*	*6.4*
Turn Offer	-	-	-	*0*	*0*	3	8	3	*14*	*9.9*
Turn Complete	4	18	9	*31*	*22.6*	-	-	-	*0*	*0*
Turn Hold	8	25	1	*34*	*24.8*	18	37	11	*66*	*46.8*
%	*15.3*	*73.7*	*11*			*24.8*	*58.1*	*17.0*		

Table 10. Icelandic and Greek corpus. In the table the largest numver in a category is highlighted in red

The most preferred modalities of non-verbal Turn Management are *face* for the Greek Host but *gestures* for the Icelandic one. The guests prefer the same modality, which is in the face while there is difference for the hosts.

The Greek host produces more turn management NVEs (147) compared to the Icelandic one (112). However the guests have a comparable non-verbal behavior, producing 137 and 141 TM NVEs respectively.

The majority of TM NVEs are employed by the Greek host to express turn offer (37.4%) but the Icelandic host expresses turn take (60.7%).

The Greek guest expresses *turn accept* (46%) and the Icelandic guest expresses *turn hold* (46.8%).

4.2 Interpretation of the comparison

There seems to be a cause and effect relation where in the Greek data there is the correlation between the host expressing turn offer and the guest expressing turn accept while the Icelandic host expresses turn take while the Icelandic guest expresses turn hold.

As our initial hypothesis was that multimodal behaviors would likely be different between those two languages we discovered that participants in both interviews shared roughly the same rate of turn management non-verbal expressions so our hypothesis was not verified however, when we examined the most prominent labels for each turn management function we realized that there are some differences in forms of turn management between the two interviews as **Table 11** shows:

	Greek Interview			Icelandic Interview	
	Interviewer	**Interviewee**		**Interviewer**	**Interviewee**
Turn Take	Head Tilt	Eyebrow raise		Finger Pointing	Gaze Down
Turn Accept	Smile	Head Tilt, Gaze Towards		Finger Pointing	Gaze Down
Turn Yield	Head Nod	Smile		Gaze Down, Gaze Towards	Eyebrow raise
Turn Offer	Head Tilt	-		Single hand repeated up and down, Both hands sideways	Eyebrow raise
Turn Complete	Smile, Gaze Towards	Head Nod, Smile		–	–
Turn Hold	-	Head Tilt, Gaze Down		Gaze Down, Gaze Towards Gaze Side	Gaze Down

Table 11. Comparison of the forms of turn management behaviour in the Greek and the Icelandic interview.

In a more detailed level the most preferred non-verbal expressions for the Greek host are *head tilt* for turn offering, *smile* for turn accept, while for the Icelandic host it is *gaze down* and *gaze towards* the guest for turn offering, and finger pointing for *turn accept*.

4.2 Summary

We have shown that speakers in both interviews share the same frequencies of non-verbal expressions to a high degree. A possible reason for that is that the political interview is a strictly framed discourse setting, institutional and highly conventionalized, which affects the production of non-verbal expressions so it is different from casual conversation, where speakers can express themselves in a more spontaneous and unrestricted manner. There is a possibility that if we moved to another interactional domain or setting, perhaps some cultural differences would be more evident.

However, there are subtle differences in the distribution and the forms of non-verbal expressions, as shown in the previous tables.

5. Discussion and Future Work

This thesis has presented the first systematic study of multimodal turn-taking and feedback behavior in an Icelandic interview setting. Overall, the data so far has matched results from previous research into these behaviors in other European and North-American cultures, while there may be some slight differences in form.

Since the main focus on this data has been on the quantitative aspect there would be an opportunity for further work to better focus on the qualitative aspects such as duration, intensity and complexity of the expressions to get a more comprehensive understanding of the multimodal behaviour seen in conversations.

There also needs to be more work done to enrich the corpus with different kinds of interviews that would show respectively semi-institutionalized interactions and casual interactions to research the difference in multimodal behaviour between those different settings to provide more generalizable descriptions.

This research shows some indication that since both interviews in Greek and Icelandic were institutionalized interactions it is likely that personalities of the speakers and guests did not show the anticipated cultural difference. However there is still a need to analyze more institutional interviews since they provide a good restrictive setting for comparative studies between cultures. The results can be used to deepening our understanding of human multimodal communication, and ultimately allow us to build computational face-to-face models for believable interactive agents employed in a variety of natural language applications.

Bibliography

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., & Paggio, P. (2007). The MUMIN Coding Scheme for the Annotation of Feedback, Turn Managements and Sequencing Phenomena. Multimodal Corpora for Modeling Human Multimodal Behaviour. *Journal on Language, Resources and Evaluation , 41 (3-4)*, 273-287.

Boudahmane, K., Manta, M., Antoine, F., Galliano, S., & Barras, C. (1998, June 26). *Transcriber a tool for segmenting, labeling and transcribing speech*. Retrieved April 8, 2007, from Sourceforge.net: http://trans.sourceforge.net/

Brugman, H., & Russel, A. (2004). *Annotating Multi-media / Multi-modal resources with ELAN.* Retrieved May 17, 2011, from Language Archiving Technology: http://www.lat-mpi.eu/papers/papers-2004/Brugman-ELAN.pdf

Cassell, J. (2000). Nudge nudge wink wink: Elements of face-to-face conversatin for embodied conversational agents. *Embodied conversational agents* , 1-27.

Cassell, J., Bickmore, T., Campbell, L., Vilhjálmsson, H., & Yan, H. (2001). More Than Just a Pretty Face: Conversational Protocols and the Affordances of Embodiment. *Knowledge Based Systems 14* , 55-64.

Duncan, S. (1972). Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology* , 283-292.

Ekman, P. (1999). Emotional and Conversational Nonverbal Signals. (L. C. Messing, Ed.) *Gesture, Speech and Sign* , 45-55.

Heritage, J. (2005). Conversation Analysis and Institutional Talk. In R. F. Sanders, *Handbook of Language and Social Interaction* (pp. 103-146). New Jersey: Lawrence Erlbaum.

Ilie, C. (2001). Semi-institutional Discourse: The Case of Talk Shows. *Journal of Pragmatics 33* , 209-254.

Jokinen, K., Navarretta, C., & Paggio, P. (2008). Distinguishing the Communicative Functions of Gestures. An Experiment with Annotated Gesture Data. *LNCS 5237* , p. 38.

Kendon, A. (2004). *Gesture: Visible Action as Utterance.* Cambridge: Cambridge University Press.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica 26* , 22-63.

Kipp, M. (2004). *Gesture Generation by Imitation - From Human Behavior to Computer Character Animation.* Boca Raton, Florida: Dissertation.com.

Kita, S., & Özyürek, A. (2003). What Does Cross-Linguistic Variation in Semantic Coordination of Speech and Gesture Reveal? Evidence for an Interface Representation of Spatial Thinking and Speaking. *Journal of Memory and Language 48* , 16-32.

Knapp, M., & Hall, J. (2002). *Nonverbal Communication in Human Interaction.* Boston: Wadsworth.

Koutsombogera, M., & Papageorgiou, H. (2009). Multimodality Issues in Conversation Analysis of Greek TV Interviews. (A. E. al., Ed.) *Multimodal Signals: Cognitive and Algorithmic Issues LNAI , 5398*, 40-46.

MacNeill, D. (1992). *Hand and Mind: What Gestures Reveal About Thought.* Chicago: University of Chicago Press.

McNeill, D., & Duncan, S. (2000). Growth points in thinking-for-speaking. (D. McNeill, Ed.) *Language and Gesture* , 141-161.

MUMIN. (2002). *Home*. Retrieved February 11, 2011, from A Nordic Network for Multimodal Interfaces: http://www.cst.dk/mumin/index.html

Pastra, K., & Wilks, Y. (2004). Image-Language Multimodal Corpora: needs, lacunae and an AI synergy for annotation. *In Proceedings of the Language Resources and Evaluation Conference* (pp. 767-770). Athens: Institute for Language and Speech Processing.

Vilhjálmsson, H. H. (2009). Representing Communicative Function and Behavior in Multimodal Communication. *Multimodal Signals: Cognitive and Algorithmic Issues Lecture Notes in Artificial Intelligence 5398* (pp. 47-59). Berlin Heidelberg: Springer-Verlag.

Appendix A: The annotation scheme

Facial display feature	Form of expression		
	Value	**Short tag**	**Elan**
General face	smile	smile	Smile
	Laughter	Laugh	Laugh
	Scowl	Scowl	Scowl
Eyebrows	Frowning	Frown	B_F
	Raising	Raise	B_R
Eyes	Exaggerated Opening	X-Open	Eye_EO
	Closing-both	Close-BE	Eye_CB
	Closing-one	Close-E	Eye_CO
	Closing-repeated	Close-R	Eye_RC
Gaze	Towards interlocutor	Interlocutor	Gaze_Tow
	Up	Up	Gaze_Up
	Down	Down	Gaze_Down
	Sideways	Side	Gaze_Side
Mouth - Openness	Open mouth	Open-M	Mouth_O
	Closed mouth	Close-M	Mouth_C
Mouth - Lips	Corners up	Up-C	
	Corners down	Down-C	
	Protruded	Protruded	Lips_Pro
	Retracted	Retracted	Lips_Ret
Head	Single Nod (Down)	Down	Head_N
	Repeated Nods (Down)	Down-R	Head_RN
	Single Jerk (Backwards Up)	BackUp	Head_SJ
	Repeated Jerks (Backwards Up)	BackUp-R	Head_RJ
	Single Slow Backwards Up	BackUp-Slow	
	Move Forward	Forward	Head_MF
	Move Backward	Back	Head_MB
	Single Tilt (Sideways)	Side-Tilt	Head_ST
	Repeated Tilts (Sideways)	Side-Tilt-R	Head_RT
	Side-turn	Side-Turn	Head_Turn
	Shake (repeated)	Side-Turn-R	Head_Shake
	Waggle	Waggle	
Unidentified		Eyes Semi Closed	Eye_SC
		Gaze Unfocused	Gaze_Un

Gestures	Shape of gesture			
		Short tag	**Elan**	
Hand gestures	**Handedness**	Both-H both hands	BH	
		Single-H single hand	SH	
		fingers_pointing	F_P	
		Fingers tapping	F_T	
		Shoulders	Shoulders_UP	
	Trajectory	Up	BH_U / SH_U	Both hands up/ Single hand up
		Down	BH_D / SH_D	Both hands down/ Single hand down
		Sideways	BH_Side/SH_Side	Both hands sideways/Single hand sideways
		Complex	BH_complex/SH_complex	Both hands complex/Single hand complex
		Repeated	SH_R_U_D	Single hand repeated up and down
			BH_R_U_D	Both hands repeated up and down
		Other	S_H_O/B_H_O	Single hand other/both hands other

Body Posture	Shape of gesture	Elan
	Torso Bend Forward	TBF
	Torso Bend Backwards	TBB
	Torso Turn Right	TTR
	Torso Turn Left	TTL
	Torso Lean Left	TLL
	Torso Lean Right	TLR

Turn Management	Function	Short tag	Elan
Turn Gain	Turn Grab	Turn-T	TGr
	Turn Accept	Turn-A	TA
Turn Keep	Turn Keep	Turn-C	TK
Turn End	Turn Yield	Turn-Y	TY
	Turn Give		TGi
	Turn Complete		TC

Role of gestures	
Deictic	Deictic
Beats+Butterworths	Beat
Iconic	Iconic
Emblem	Emblem
Adaptor	Adaptor

Feedback	Specific function value	Short tag	Elan
Give	Contact/ Perception/ Understanding Contact/ Perception Accept Non-Accept		FG_CPU FG_CP FG_A FG_NA
Elicit	Contact/ Perception/ Understanding Contact/ Perception Accept Non-Accept		FE_CPU FE_CP FE-A FE-NA
Emotions/Attitudes		Happy Sad Surprised Angry Disgust Frightened Certain Uncertain Interested Uninterested Disappointment Satisfaction	

Multimodal Relations	Elan
Repetition	Rep
Addition	Add
Substitution	Sub
Contradiction	Con
Neutral	Ne
Target	Target

Appendix B: Icelandic corpus

Here are all the turn management annotations with the functional attributes. In this first table it is the facial expressions and which expressions accompanied which turn management function. TT stands for turn taking, TA, for turn accept, TH for turn hold, TY for turn yield and TO for turn offer.

		Host							Guest					
		tt	ta	th	tc	ty	to		tt	ta	th	tc	ty	to
fe	Head Single Tilt									2				
	Head Nod	1					1			1		2	2	
	Smile													
	Brow Raise	4					1		1	2	8		3	5
	Brow Frown													
	Head Single Jerk													
	Head Move Forward													
	Head Move Back													
	Gaze Down	14	4	1		2	2		4	9	20			1
	Gaze Side			1					2	3				
	Gaze Towards Interlocker	8	2	1		2	2		2	2	6			
	Head Repeated Nods	2								2	1			2
	Head Repeated Tilt										1			
	Eye Closing Repeated													
	Mouth Open													
	other													
	Head Turn													
	Gaze Up													
	Head Shake													
	Gaze Unidentified													

G stands for gestures.

		Host							Guest					
		tt	ta	th	tc	ty	to		tt	ta	th	tc	ty	to
g	Finger Pointing	5	6			1	1				1			
	Single Hand Up	1					1				2			
	Single Hand Down	1	1											
	Single Hand Other	1												
	Single Hand Sideways	3												
	Both Hands Repeated Up Down	4		1						2	6			
	Single Hand Repeated Up Down	4		1			2		1		1			
	Both Hands Up								1	1				
	Both Hands Other	1												
	Finger Tapping	1							1		1			
	Both Hands complex	4		1		2			1	2	4		2	1
	Single Hand complex	2	1				1							2
	Both Hands Down	2												
	Both Hands Sideways	2					2			2	2		1	

BP is body posture

		Host							Guest					
		tt	ta	th	tc	ty	to		tt	ta	th	tc	ty	to
bp	Torso Lean Left									2	1			1
	Torso Lean Right	1									1			
	Torso Bend Backwards	4	1			1			2	3	7		1	1
	Torso Turn Right	1												
	Torso Bend Forward	2					2		1	1	2			1

Appendix C: Feedback elicite

Here is a table where every feedback elicit, pauses and filled pauses and the functions that were associated with them were counted. They are counted in the time order that they appeared on the video.

Where only the letter H or K appears and no functions it only means that there was some feedback elicit, pause or filled pause but no functions associated with them.

H – Host

K – Guest

Feedback elicit	Pause	Filled pause
H.Body, Gestures TO		
		K.
K.Face, Gesture		
K. Gaze, Gesture		
K. Body, Face, Gesture		
K. Face,		
K. Face,		
H.		
	H. Gesture	
H.		
	K. Face,	
H. Body, Gaze, TY, Gesture		
		K. Gesture,
H. Gaze, Gesture, TH,		
		H. Gesture, TY
K. Body, TH, Gesture,		
	K. Gesture, TH,	
K. Body, Gesture,		
K. Gesture,		
K. Face, Gesture,		
	K. Face, Gesture	
		K. Gesture, TH
K. Face,		
K. Body, Gesture,		
K. Face, TO,		
		H. Body, TA, Gaze, Gesture,
H.		
K. Face,		

K. Body, Gaze, Gesture,		
K. Body, Face,		
K. Face,		
K. Face, TO,		
H. Gaze, TO, Gesture,		
K. Gaze,		
K.		
		K. Body, Gesture, Gaze
K. Face, TO,		
H.		
K. Gaze,		
	K. Gesture	
K. Gaze, TT,		
K. Gesture,		
	K. Face, TY	
H. Gaze, TO,		
H.		
H.		
K.		
K. Face,		
K. Face		
K.		
K. Gesture,		
K. Gesture,		
K. Gesture, TH		
H. Gesture, TT		
H.		
H.		
K. Face, TA, Gesture,		
K. Face,		
H.		
K. Body, Gesture,		
K. Body, Gesture,		
K. Body, Gesture,		
K. Face,		
		K. Face,
K. Gaze, Gesture,		
K. Gesture		
K. Gesture,		
K. Gesture,		
K. Face, Gesture,		
		K. Body, TO,
K.		

K. Gaze, Gesture TO		
K. Body, TO, Gesture,		
		H.
H. Gaze, Gesture,		
	H.	
H. Face, TO		
	K.	
	K.	
	H. Gesture, TO	
H. Gesture, TO		
K. Gesture, TA		
	K. Gesture, TA	
H. Gesture, TT		
K. Face, Gaze, TH		
	K. Face,	
K.		
	K. Gaze,	
K.		
H. Face, TT		
	K. Gaze, TH, Gesture	
H.Gesture, TT,		
	K. Body, TH, Gesture,	
	K. Gesture, TH	
H. Gaze, TT, Gesture		
K. Body, Gaze, Gesture,		
	K.	
	K. Gesture,	
K. Body, TY, Face, Gesture,		
H. Gesture, TO,		
H. Face, TT,		
K. Gaze, Gesture		
K. Gesture,		
K.		
K. Body, TT, Gaze,		
H.		
K.		
K.		
H. Gesture, TT		
K. Gesture,		
K. Gesture,		
K. Gesture,		
	K. Body, Gaze, Gesture	
	K. Gaze, Gesture, TH	

	K. Gaze, TH, Gesture	
	K. Gesture, TH	
H. Body, TT, Gesture,		
H. Body, TT, Gaze, Face, Gesture,		
H. Gaze, Gesture,		
H. Gaze, Gesture,		
		H. Gaze, Face, Gesture,
		H. Gaze, Face, Gesture,
H. Face, TO, Gesture,		
	K.	
	K.	
H.		
K.		
	K. Gesture,	
H.		
K. Gesture, TH		
K. Gesture,		
K. Gesture,		
H. Body, TT, Gaze, Face, Gesture,		
H. Face, TT, Gesture,		
H.		
H. Gesture, TO		
H. Gestures, TT,		
H.		
K. Body, Gesture,		
	K. Gesture	
	K. Gesture,	
	K. Gesture,	
	K. Gesture,	
K. Gesture,		
K. Body, Gaze, Gesture,		
	K. Body, Gaze, Gesture	
	K. Gesture	
K. Body, TH, Face,		
	K. Gesture TH	
	K. Face, TH, Gesture	
K.		

Appendix D Gestures for host and guest

H-Host

Gestures H	H:Correlation with feedback elicit	H:Correlation with filled pause	H:Correlation with pause
Single hand up	Same timeframe		
Single hand sideways			Movement starts there
Both hands complex	Begins a little earlier and ends later		
Both hands repeated up and down	Same timeframe		
Both hands complex		Movement starts there	
Single hand down		Same timeframe	
Single hand repeated up and down	Movement ends there		
Finger pointing	Starts much earlier and continues		
Finger tapping	Starts much earlier but ends there		
Both hands sideways			Starts much earlier and continues
Same gesture as above	Ends here		
Single hand sideways	Begins a little earlier and ends soon		
Single hand complex	Begins a little earlier and ends soon		
Single hand repeated up and down	Begins a little earlier and ends soon		
Single hand complex	Begins a little earlier and ends soon		
Finger pointing	Begins a little earlier and ends there		
Finger pointing	Same timeframe		
Single hand complex	Starts much earlier and continues		
Both hands complex	Starts much earlier and continues		
Same gesture as above	Starts much earlier and continues		
Same gesture as above		Starts much earlier and continues	

Same gesture as above		Starts much earlier and continues
Finger pointing	Begins a little earlier and ends soon	
Both hands complex	Starts much earlier and continues	
Same gesture as above	Starts much earlier and continues	
Single hand repeated up and down	Begins a little earlier and ends soon	
Finger pointing	Same timeframe	

K- Guest.

Gestures K	K: Correlation with feedback elicit	K: Correlation with filled pause	K: Correlation with pause
Both hands sideways	Starts there ends soon after		
Both hands sideways	Same timeframe		
Single hand complex			Started sooner ended there
Both hands complex		Begins a little earlier and ends later	
Finger tapping	Starts much earlier and continues		
Same gesture as above			Starts much earlier and continues
Same gesture as above	Starts much earlier and stops there		
Single hand complex	Starts there and ends later		
Both hands sideways	Starts there and ends later		
Both hands sideways			Starts earlier and ends there
Both hands repeated up and down		Starts earlier and ends there	
Both hands repeated up and down	Starts much earlier and continues		
Both hands repeated up and down	Starts earlier and stops soon after		
Both hands sideways	Starts much earlier and continues		
Both hands complex		Starts earlier and ends there	
Single hand sideways			Starts earlier and ends there
Both hands repeated up and down	Starts there ends later		
Both hands complex	Starts much earlier and continues		
Same gesture as above	Starts earlier and stops soon after		

Both hands repeated up and down	Starts earlier and stops soon after		
Both hands complex	Starts earlier and stops soon after		
Both hands complex	Starts much earlier and continues		
Same gesture as above	Starts earlier and stops soon after		
Both hands complex	Starts much earlier and continues		
Both hands repeated up and down	Movement starts there and continues		
Same gesture as above	Starts much earlier and continues		
Same gesture as above	Starts much earlier and continues		
Same gesture as above	Starts much earlier and continues		
Finger tapping	Starts much earlier and continues		
Both hands complex	Starts there and continues		
Same gesture as above	Starts earlier and stops there		
Both hands repeated up and down	Starts much earlier and continues		
Same gesture as above			Starts much earlier and continues
Both hands repeated up and down	Started earlier and stops there		
Both hands repeated up and down			Starts much earlier and continues
Single hand repeated up and down			Starts much earlier and continues
Both hands complex			Starts much earlier and continues
Both hands up	Started earlier and stops there		
Both hands repeated up and down			Starts much earlier and

			continues
Both hands complex	Same timeframe		
Both hands repeated up and down	Starts a little earlier and ends later		
Both hands repeated up and down	Starts earlier and ends there		
Both hands repeated up and down	Starts much earlier and continues		
Same gesture as above	Starts much earlier and continues		
Same gesture as above	Starts much earlier but ends there		
Both hands complex			Starts much earlier and continues
Single hand sideways			Starts earlier but ends there
Both hands complex			Starts much earlier and continues
Same gesture as above			Starts much earlier and continues
Single hand repeated up and down			Starts much earlier and continues
Both hands repeated up and down	Starts much earlier but ends there		
Both hands complex	Starts much earlier and continues		
Same gesture as above	Starts much earlier but ends there		
Both hands sideways	Starts much earlier and continues		
Both hands sideways			Starts there but ends later
Same gesture as above			Starts much earlier and continues
Same gesture as above			Starts much earlier and continues
Both hands complex			Starts much earlier and continues

Same gesture as above	Starts much earlier and continues		
Finger tapping	Starts there but ends later		
Both hands complex			Starts much earlier and continues
Same gesture as above			Starts much earlier and continues
Both hands complex			Starts much earlier and continues
Same gesture as above			Starts much earlier and continues