Ágnes Abuczki:
A multimodal analysis of the sequential organization of verbal and nonverbal interaction
Argumentum 7 (2011), 261-279
Debreceni Egyetemi Kiadó

Working paper

Ágnes Abuczki

A multimodal analysis of the sequential organization of verbal and nonverbal interaction

Abstract

The paper addresses the study of the formal features of conversation from both a theoretical and an empirical perspective. In particular, following the framework and terminology of Sacks, Schlegoff and Jefferson (1974), turn-taking and sequencing are designated 'formal' features of interaction. Furthermore, the study adopts and fits into the framework of the theoretical modelling of human–computer interaction (HCI), as outlined in Hunyadi (2011). After reviewing the theoretical background of the multimodal analysis of conversation, a communicative act-based pragmatic annotation system will be outlined which is to contribute to a better understanding of the sequential organization of interaction. Finally, a case study on the sequential organization of nonverbal cues for turn regulation as well as their alignment with verbal content will be described.
Keywords: multimodal analysis of conversation, pragmatic annotation system, turn management, nonverbal communication, human–computer interaction

1 Introduction

One may ask why the study of the basic structure of interpersonal communication and turn management is necessary for the description and development of human–computer interaction. My answer is that the way people take turns when talking in a synchronized manner seems to be the most salient feature of interpersonal communication, therefore, natural, human-like man–machine interaction must include the phenomenon of turn-taking with accurate timing. In general, any form of joint activity requires the coordination and management of the participants' activities. A further reason for analysing the operation of the turn-taking system is the fact that it is an invariant feature of any conversation, whatever the number, age, gender or occupation of the participants. As Sacks (1992: 34) puts it: "They hold across types of conversations – arguments, business talks, whatever else. They hold across the parts of a conversation – beginnings, middles, ends. They hold across topics." In this sense, the system of turn-taking is basic and abstract; therefore, it should form an essential part of a communication model. Moreover, as we know from our personal judgements of everyday conversations in general, the structure of conversations is just as important as the content itself. Needless to say, people can understand each other because linguistic organization allows them to understand novel sentences as well.

Concerning its methodology and material, the study also aims at taking a novel multimodal approach to the analysis of the above mentioned phenomena by taking into account the nonverbal behaviour of the interlocutors as well. Why is a multimodal analysis of turn management strategies necessary? Basically, it is because the mechanisms by which people take turns in discourse are not just verbally regulated; these processes are both spoken and nonverbal, as well as open and subconscious (Wiemann & Knapp 1975).

With a close observation of the recorded multimodal database, the goal is to find patterns across instances. Taking a data-driven inductive approach in the turn management case study, I first observe the interviews and carry out a qualitative microanalysis. Secondly, I analyse their multi-level annotations both horizontally – to reveal the simultaneous occurrence of certain signals – and vertically – for a sequential analysis of the conversational units –, and then I describe typical human behaviour during interaction. With the help of the correlation of certain labels at the discourse level and the video level, I strive to find the nonverbal cues for the detection of different turn/floor control types. However, the communication model of the project has a higher objective as it intends to establish the generic structure of all possible communicative events (including even those never before observed), after having uncovered the relevant general features of actual communicative events contained in the database. As Hunyadi (2011) puts it:

> […] our task is to provide technology with such a general communicative structure that meets the minimum requirement: it should have significance beyond the unique pragmatic differences of arbitrary communicative events. This way, (during bottom-up processing), built on the universal building blocks of communication, we move from the general towards the individual, creating a possible set of event structures by such a mechanism that is able to distinguish well-formed communicative structures based on their content.

Turn-taking organization for conversation fits into this model since the major aspects of turn-taking are insensitive to contextual parameters (cf. Sacks et al. 1974). In short, the proposed communication model of Hunyadi (2011) is a generative, rule-based model of human–human as well as human–computer interaction, taking various modalities (verbal, visual, nonverbal acoustic) into consideration, which is necessary since human communication is inherently multimodal and sequential, and ideally, naturalistic human–computer interaction should be in many ways similar to interpersonal communication. Hunyadi's model is inspired by generative linguistics and computational pragmatics (cf. Bunt & Black 2000). The Hunyadi-model (2011) supposes that communication is a code-guided behavior and has a general, abstract, meaningless underlying structure. Moreover, it seeks to take into account both directions of communication: production and interpretation, i.e., synthesis and analysis. The Hunyadi-model (2011) itself is composed of two main parts: the invariant structure and the representation of this invariant structure. The invariant structure is further made up of two modules: the invariant formal basic structure and the invariant functional extension. The representation of these is realized at the level of the pragmatic extension. The pragmatic extension involves the pragmatic functions, those turning the skeleton of an event into an actual surface realization. For example, the coincidence of eye contact with falling pitch and falling intensity ending in silence signs turn-giving pragmatic information.

According to Németh T.'s (2011) evaluation of the proposed Hunyadi-model (2011), the basic idea of Hunyadi is that during interaction the speakers take into consideration not only their own intentions but also the general underlying structure of communication, the possible communicative strategies and the multimodal tools which can be applied in order to achieve

their communicative goals in the context of the interaction. Conversely, the significance of the model lies in the fact that it takes into account analysis as well, i.e., while a communicative partner interprets the communicator's multimodal communicative behavior he/she relies not only on the surface clues available in the discourses but also his/her knowledge about the underlying structure of communication, context as well as his/her beliefs about the communicator's intentions, beliefs and desires. In sum, the advantage of such a structural view of communication is that it enables technology to treat the more or less different communicative events in a uniform manner, creating uniform methods and template-like realizations.

One of the basic research questions of communication studies in general is whether language, gesturing and ostensive stimuli are orchestrated signals, and if so, to what extent and how? My goal is to find out how language and nonverbal behaviour act together in conversation – with special regard to turn-taking mechanisms – by tying together their observation and annotation in spontaneous speech. I attempt to outline the rules and signals which govern the structuring of natural conversation and find the nonverbal cues (especially gaze, hand gestures and posture changes) for floor control and different turn types. The paper tries to identify the visually transmitted nonverbal cues and the cues carried in speech-wave form, which are responsible for the regulation of conversational exchanges.

It must be pointed out that most gesture research is based on the assumption that body movements during speech are not random. However, gesture researchers are split in their views of whether gestures are just by-products of speech production processes or their purpose is inherently communicative (cf. Kendon 1994). I would add that gestures are not inherently communicative but rather they are driven by the communicative intention of the speaker. The HuComTech multimodal corpus provides numerous examples in which eye gaze together with head and hand movements are both communicative components of language expression and also serve to coordinate interaction.

One of the long-term prospects of the present study is to contribute to the detection of the turn-giving and turn-grabbing intention of the other speaker together with the help of computational tools and software based on silence detection (the typical length of SIL indicating speaker change), hand gesture recognition and eye-gaze tracking.

2 Conversation Analysis

In order to observe and model how people sequence their conversation requires empirical investigation of how people actually talk. The basic ideas and methodology of the present study come from the traditions of Conversation Analysis (CA) – which is an empirically oriented research activity (Silverman 1998: 153) dealing with the interconnectedness of practices in talk-in-interaction (Schlegoff 2006: 472) –, and compatible with the Hunyadi-model of communication (Hunyadi 2011). Regarding turn-taking, Goodwin and Heritage (1990: 290) point out that CA intends to provide a rule set that orders possible turn transition and speaker-selection techniques, including self-selection and next speaker or next action selection as well.

These options for actions can be grasped most clearly by the concept of adjacency pairs which project probable and expected second turns and may lead to various exchanges of

questions and answers, as it can be seen in the simulated job interviews of our HuComTech database (DB).

Sacks and his colleagues highlight the essential features of the sequential organization of conversation (Sacks 1992: 32, Sacks et al. 1974: 95–9, 621–3):

People talk one at a time. However, overlapping speech sometimes occurs in the database (DB) and can be seen as a signal of involvement in the conversation rather than a sign of competitive or aggressive behaviour.

Speaker change recurs. In the job interviews of the HuComTech corpus, it is mostly the interviewer who is the initiator of the topic by asking questions, to which the interviewee cooperatively provides answers. It must be highlighted that discourse structure largely depends on the genre of the discourse. That is why it is easier to predict the structure of job interviews than that of informal conversations.

Sequences that are two utterances long and are adjacently placed may be 'paired' activities: The organization of two consecutive utterances provides the concept of 'adjacency pairs' (Sacks 1992: 188-196). APs are 'relatively ordered' because one always goes before the other. Sacks (1992) suggests that adjacency relationship is the most powerful tool for relating utterances since it is a way of organizing a relationship between a current utterance and prior and next utterances (Sacks 1992: 555).

Certain activities are 'chained'. For instance, as I have already pointed out, in the job interview scenario long strings of talk may be organized in the Q-A-Q-A format. In the Hunyadi-framework of human–computer interaction (HCI) (Hunyadi 2011), it is also allowed in the system of discourse to embed sub-dialogs within these adjacency exchanges. Mey (1993: 223) provides the following short dialogue as a typical example of inserted sequences:

– Are you doing anything tonight?
– Why are you asking?
– I thought we might see a movie.
– Well, no, nothing in particular. What do you want to see?

Of course, it must also be noted that in an exchange different conversational goals may be pursued at the same time: greetings, informing, exchange of info about the weather, invitations, etc. At the level of technology, these different goals can be grasped and identified with their corresponding visual, nonverbal acoustic and verbal acoustic cues. The goal of the pragmatic annotation system is to reveal these cues and signals called markers in the Hunyadi-model (Hunyadi 2011).

Activities can be required to occur at 'appropriate' places. For instance, in the Hunyadi-model of HCI (Hunyadi 2011), the dynamic and sensitive system of communication confers upon interactions a distinctive temporal organization: it provides a metric within which a given length of silence may be understood as the marker of the interlocutor(s) not speaking. Of course, it must be added that silence may perform several functions, from cognitive planning to hesitation, therefore, the meaning of a particular pause is judged by the co-occurrence of other acoustic and visual markers.

Turning back to the main line of argumentation, the major intriguing question of the present study is how people go about allocating turns to each other or themselves, and how it can be applied in HCI technologies. The answer lies in the above mentioned 'turn-taking mechanisms' (Mey 1993) which involve the identification of the so-called transition-relevance places (TRPs in the ethno-methodological parlance) where a natural transition may

occur (e.g. during a pause for breath or when people run out of things to say). The agents constantly monitor each other and each other's activities in the communicative situation and react to the situation. One of the objectives of this study is to collect and organize the cues for the automatic detection of these TRPs for use in HCI since a basic feature of human–human interaction is that speakers collaboratively locate transition points and allocate turns to the next speaker. In order to achieve this goal, I must take into account tempo, volume, pitch of speech, intonation and word-choice patterns, pre-sequences, pre-closing signals, and all nonverbal cues such as gaze direction, posture, and the absence or presence of hand gesturing. Supposing a TRP is identified, the next question is how a TRP can be exploited. Sacks and his colleagues (Sacks et al., 1974, Sacks 1992) provide four possible options for what the interlocutors may do at a TRP:

- the right to speak may be allocated to another speaker
- more indirectly, the floor might be left wide open to whoever wishes to take it
- alternatively, a speaker might just ignore the TRP and continue past it
- speakers may employ the technique of 'masking' the TRP by emitting sounds (such as 'Öööhhm' in the HuComTech corpus) at potential transition points, thus warning other speakers of their intention to continue past the TRP as soon as they have got their breath (Mey 1993).

On the other hand, conversation analysts have observed that non-floor-holders in a conversation are not mere silent bystanders; instead, they constantly monitor the speaker and provide backchannels (e.g. by nodding and saying 'I see', 'Right') which provide support for the speaker. Bystanders might also directly intervene in the conversation by taking the floor, preferably at a TRP. One clearly identifiable case of predictability is that of adjacency pairs (Levinson 1983). It is important to note that spoken utterances of the HuComTech corpus occur in dialogues (and basically all dyadic interactions of course), where utterances not only encode intentions, they also respond to previous utterances and invoke further responses (Bunt & Black 2000).

'Recurrent and reliable patterns' of correlation are often found between the two domains, that of 'preference' and that of 'the kind of action performed'. In Levinson's view (1983) reference is a FORMAT of utterances. Here is a table of correlations that Levinson has in mind:

FIRST PARTS		request	offer/ invitation	assessment	question	blame
SECOND PARTS	Preferred	acceptance	acceptance	agreement	expected answer	denial
	Dispreferred	refusal	refusal	diasagreement	unexpected (or no) answer	admission

Table 1: Typical patterns of correlation between first parts and second parts of adjacency pairs (based on Levinson 1983)

Given one part of the adjacency pair, the other is normally predictable (e.g. question-answer, greeting-greeting, invitation-accepting/refusal, etc.). As a technological application of this system, graphical user interfaces (Bogdan et al. 2006) work similarly: a question is given by the computer in the opening part of the adjacency and then it must be followed by an answer, supposing the cooperative behaviour of the human user. The same exchanges apply in the

conversations in the HuComTech DB as well, especially long strings of question-answer exchanges together with some exchanges of assessment and agreement. If the computer has the plan (based on adjacency relations) and drives the interactions according to it, then the dialogue management might work more effectively.

Where the expected next action does not occur, it is treated as noticeably absent (Goodwin & Heritage 1990: 287), and that absence can be the object of remedial efforts and justifiably negative inferences. Adapted to the framework of the Hunyadi-model of HCI, the absent second part of the pair can be seen as a zero marker of a pragmatic primitive (Hunyadi 2011). Even if it is absent, its place is there, and its absence is very noticeable indeed, marking some significant shift in the flow of the conversation.

Marked behaviours such as producing dispreferred second pair parts (such as making a turn –as in Duncan's traffic metaphor of interaction (Duncan 1972)) have to be announced, preferably before they happen (e.g. by turning on one's signals). Furthermore, marked behaviours are dispreferred because they require more effort on the part of other users since the speakers need to give an account for their acts. Dispreferred answers are of 'No-plus' form (Sacks 1992: 414) as they provide an account of the circumstances of the negative reply. On the other hand, preferred answers are shorter because they meet the inferred expectation.

A similar marked behaviour is when the speaker brings in a new topic to the flow of conversation. When one wants to say something unrelated to the prior utterances, they must do something special to raise attention to their utterance, e.g. by using discourse markers or linking items (e.g. "otherwise", "you know", "by the way").

More recent action-based approaches to language assume that communication is a form of rational activity, where the participants use communicative acts to achieve communicative as well as non-communicative goals in a rational way (Bunt & Black 2000). Similarly, the pragmatic annotation of the HuComTech corpus outlined below offers a communicative act-based approach to understand the mechanisms of interaction.

3 Pragmatic annotation in the HuComTech project

Besides labelling the turn types (T (turn-take), K (turn-give), G (turn-give), BC (backchannel) and SIL (silence>250 ms)) in the discourse level (Abuczki 2011) of the audio annotation (Pápay 2010), the HuComTech project strives to contribute to a better understanding of the above described multimodal phenomena and the underlying structure and relations of communicative events with the help of designing and completing the pragmatic annotation of the corpus.

Annotation in general involves the segmentation and labelling of the discourse, while pragmatic annotation is the process of marking up stretches of talk with information about the communicative acts performed (c.f. Bunt et al. 2010). As Bunt & Black (2000) point it out, spoken utterances – just like the interviews and conversations of the HuComTech corpus – most often occur in dialogues, where utterances not only encode intentions but also respond to previous utterances and invoke further responses.

Many utterances are best viewed as speech acts since during interaction people perform speech acts, such as making statements, giving commands, and asking questions (Searle 1969). Since speech act theory provides a formal and clean view of communication, computer scientists have found them useful for describing communication and designing the interfaces

for human-computer interaction (HCI) and dialogue management systems (Bogdan et al. 2008).

Bunt & Black (2000) goes a step further and proposes that communication can be best modelled using the concept of dialogue acts. They define dialogue acts (DA) as regular patterns of utterances in dialogues, a combination of the speech act and the semantic force of an utterance.

Communicative acts (listed in Table 3) as basic units of pragmatic annotation of the HuComTech corpus were chosen since they construct coherent discourses. Moreover, a communicative act contains stimuli from various modalities (not only from the verbal modality as speech act theory): visual, verbal, non–verbal acoustic and ostensive behavior without code-using.

The notion of communicative act plays a significant role in spoken discourse studies. As Bunt et al. (2010) sum up, dialogue act annotation (which can be seen as an example of or a subtype of pragmatic annotation) has a key role in (1) the interpretation of communicative behaviour of dialogue participants, (2) building annotated corpora, and (3) in the design of dialogue management systems for spoken human–computer interaction.

Therefore, the final goal of the pragmatic annotation of the HuComTech corpus is the recognition of communicative acts, and then, in response, the automatic generation of appropriate replies.

Allwood's approach to pragmatics, called *Communicative Activity Analysis* has been influential in the HuComTech project. Communicative Activity Analysis views communication as action, which is seen as constituted by a combination of behavioral form, intention, context, and result. "Every utterance in a conversation is assumed to have a functional structure with three components: (1) functions obligated by the preceding discourse; (2) functions obligating for the succeeding discourse; (3) 'optional' functions, which are neither obligating nor obligated" (Bunt & Black 2000: 26). It is claimed that communicative intentionality has an expressive aspect, which is to express a certain attitude, and an evocative aspect, which is to evoke a certain reaction from the addressee. In sum, Allwood's approach offers a rich conceptual framework for the analysis of human dialogue, and has been influential in the HuComTech project as well.

The design of pragmatic annotation has been planned and developed by the members of the Communication Modelling subproject of the HuComTech research group (Abuczki, Á., Bársony, O., Bódog, A., Hunyadi, L., Ivaskó, L., Németh T., E., Tóth, Cs.)

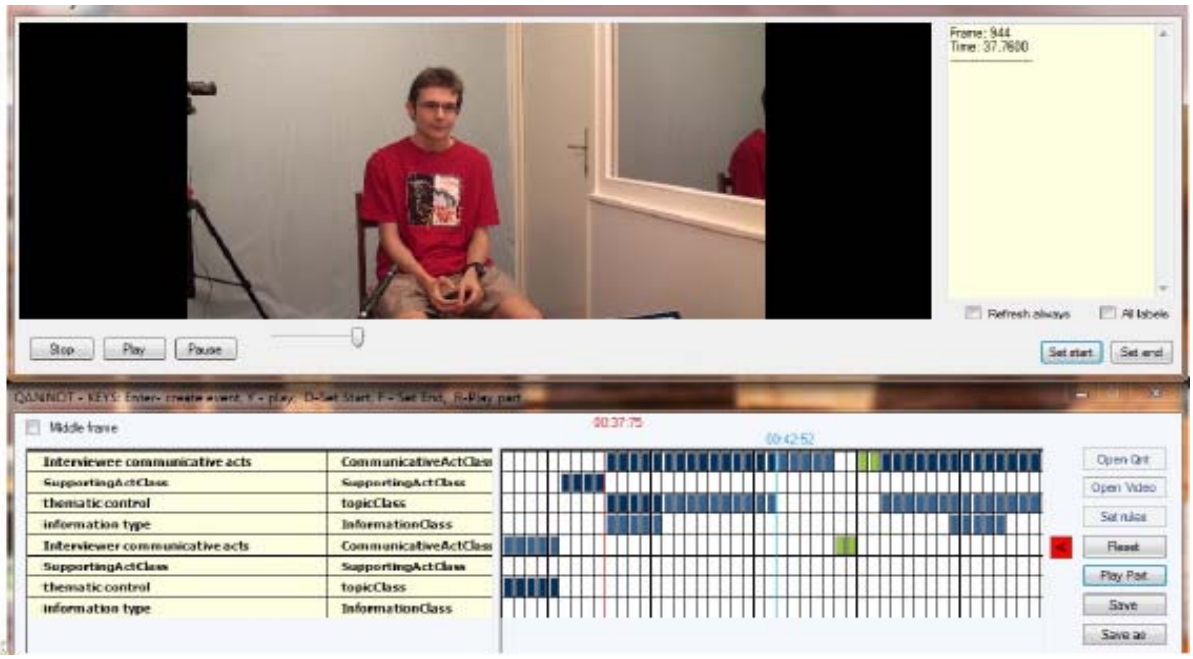

Figure 2: Screenshot of the pragmatic annotation of the HuComTech corpus in Qannot

The annotation tool of the pragmatic annotation of the HuComTech corpus is Qannot (same software used for video annotation) since the visual and nonverbal acoustic markers are also significant and Qannot displays the labels of video annotation as well. Besides, Qannot (presented in Figure 2) is also able to read and import Praat textgrids of the audio annotation, thus enabling the cross-modal statistical analysis of the audio, video and pragmatic levels and labels.

In the pragmatic annotation scheme of the HuComTech corpus annotations are directly attached to the stretches of dialogue defined by begin- and end points. However, in some other similar annotations, such as in Dialogue Act Markup Language (DiAML) (Bunt et al. 2010), annotations are attached to structures at other levels of analysis, for instance, at the output of a tokenizer.

The levels and labels of the pragmatic annotation are as follows (Table 3):

1st level: communicative acts: (Searle 1969, Austin 1962, Németh T. 1996, 2005, Abuczki et al. 2011)

contribution to the discourse ~ turn	communicative act type	CON: answering, confirming, informing, predicting, recalling
		DIR: asking, requesting, suggesting
		COM: agreeing, offering, promising
		ACK: greeting, accepting
		none
	supporting act type	backchannel
		politeness markers (please, I know...but, etc)
		repair
		none

Table 3: The level of communicative acts (Abuczki et al. 2011)

The first level involves the labelling of the communicative acts (listed in the table above). In addition to the four basic types of communicative acts, repaired acts and indirect acts are also marked in the annotations. It must be emphasized that the labels of communicative act types are not interpreted but they are behavioural labels (Bódog 2011). It means that an act of informing or answering should be 'purely' determined by the perception of the ongoing communication and not by the interpretation of the conversation. Labelling is based on illocution (act) and not on perlocution (effect).

2nd level: thematic control:

The purpose of the second level of the pragmatic annotation is the distinction of the visual and nonverbal acoustic markers of:

– topic initiation
– topic elaboration
– topic change.

3rd level: given vs. new information:

It has been proposed that (semantically) new information – new compared to what has been said previously in the context of the conversation – should also be marked since in my experience, lexically-semantically most prominent information is often accompanied by heavy gesturing. The present hypothesis can be tested with the introduction of this level of annotation. Of course, all the levels are annotated on the sides of both the interviewer and the interviewee.

In sum, the purpose of pragmatic annotation is to find the underlying structure of communicative behaviour as well as the visual, acoustic and verbal correlates (in the further audio and video levels of the annotation design) of different communicative acts. I am going to examine and distinguish lexical, prosodic, syntactic, discourse and visual surface features that are probabilistically associated with particular communicative acts. For instance, it might be proposed that declarative structure and word order is a possible cue for identifying commands. Similarly, prosody, including intonation, is another cue for communicative act identity. For example, information-checking questions often have rising tone. However, intonation alone is not a guarantee for communicative act identity, visual cues, such as facial expressions must also be taken into account.

Generally, as Bunt et al. (2010) also highlight, successful communication depends on the addressee's understanding of the communicative functions of the speaker's utterances in the way intended by the speaker. Therefore, the disambiguation and identification of the communicative acts will definitely contribute to the technological implementation and improvement of dialogue management systems.

4 Theoretical background of gesture analysis: the notion of composite utterances

First, Sacks and his colleagues (Sacks et al. 1974) introduced the idea that the concept of utterances should be used in CA instead of sentences. Later on, in the multimodal framework of the notion of "composite utterances" (Kendon 2004, Enfield 2009) it has been proposed that the nonverbal behaviour of the speaker and the verbal content are orchestrated together, gesture completes the utterance's meaning, therefore, meaning is a composite notion. As Kendon (2004) puts it: "The gestural component and the spoken component interact with one another to create a precise and vivid understanding". Moreover, gestures often express something complementary to what is being expressed verbally.

One might ask what exactly a gesture is, how it can be defined and what sorts of communicative functions it fulfils. Kendon (2004) provides a broad but still straightforward definition of a gesture: "It is a form of human expression, an activity that is significant for the understanding of a speaker's expression" (Kendon 2004: 107). Also, gestures form a nonverbal sign and code system, so they are easier to teach to machines.

However, it must be kept in mind during analyses that gesture is a multifunctional 'device' as it is a medium which can be employed in many different ways. Therefore, after recognizing and identifying a gesture, people must be careful while disambiguating its meaning.

In the traditional classification of communicative gestures (Ekman 1998, Kendon 2004), one can distinguish two basic functions which gestures are used for:

Firstly, gestures often represent something (by some sort of depiction or enactment to exemplify objects or actions) that is relevant to the referential content of what is being said. For instance, descriptive gestures are very common in the HuComTech DB, especially in informal conversations, and they are mostly used to describe the size and form of objects or events, as well as to show spatial relations. Similarly, somewhat metaphorically, when the speakers refer to or recall past events, they tend to use hand movements to their left side; in contrast, when they refer to the future – which is much less frequently the case in the HuComTech corpus as recalls are more common – they rather gesture to their right side which can be seen as a result of their literacy traditions: writing from left to right, progressing from left to right.

Secondly, speakers may display the logical structure of their discourse, especially by hand gestures. Moreover, in some cases, when head and hand movements were coupled with this parsing function, temporally synchronized head and hand movements were produced.

Besides these two above mentioned basic functions, Kendon (2004: 225) distinguishes four major functions of gestures:

- pragmatic function: as they contribute to or constitute the acts accomplished by utterances

- modal function: if they operate on a given unit of verbal discourse and show how it is to be interpreted
- performative function: gestures showing what sort of speech act a speaker is engaging in
- parsing function: gestures marking the structure of spoken discourse

Based on a qualitative analysis, it is assumed that within the gestures of the HuComTech DB, those with parsing and pragmatic functions are the most common.

As far as the disambiguation of the meaning of the gestural component of communication is concerned, Enfield (2009) defines four sources of a gesture's meaning that must also be followed in gestural analyses:

- (coterminous) speech
- (prior) stimulus or cause of an utterance
- response to/effect of an utterance
- purely formal characteristics of the gesture.

Moreover, during the multimodal analysis of interactions, a gesture should not be analysed on its own as it occurs; rather, three 'contexts' of the gestural component of an utterance must be taken into account:

- stimulus/cause: what has just happened
- speech + gesture/movement together: what else is happening now
- response/effect: what happens next (Enfield 2009).

Completing the above theories, Hunyadi (2011) proposes that gestures might also have some sort of grammar or prosody. Therefore, they must have some definable and identifiable constituents. According to Hunyadi (2011), all gestures have a beginning (preparation), a peak (stroke), and an ending (recovery). Similarly, Enfield talks about and defines so-called 'gesture phrases'. In his framework, he calls an entire movement excursion a 'gesture unit' and the nucleus of the gesture phrase a 'stroke' (2009: 124–125).

In the interactional and gestural analysis of the recordings of the DB, I took into consideration all the above mentioned theories. Furthermore, I also want to test and validate for Hungarian speakers as well some hypotheses about gestures put forward earlier in the literature. For instance, Kendon (1972) was the first to observe that some head movements (= the alternate manifestations of the eye gaze pattern) relate to the discourse structure of an utterance. As Jokinen (2010) rightly points it out, body movements function interactively, for example, the 'speech-preparatory' repositioning of the head before the start of talk simultaneously signals the assumption of a turn or the intention to continue, so it is also a crucial part of turn management. In the HuComTech corpus, the same pre-positioning of the head occurs most of the time. Regarding the gaze behaviour of the listeners, Goodwin (1981) was the first to discuss how speakers can assess recipiency for their talk by inspecting the gaze of others. Concerning hand gesturing, if the listener makes hand gestures, such as raising hands, then he/she may be requesting control of the floor – that is called the "traffic signal" approach (Duncan 1972). During the analyses I am currently testing all the above mentioned phenomena in the corpus for a better understanding of the alignment of speech and gesturing.

5 Case study and research findings: The sequential analysis of the alignment of verbal and nonverbal cues in turn management

With the help of revealing the correlation of certain labels among the audio level, the discourse level and the video level, the study intends to find nonverbal cues for the detection of different floor control types. The material of the study is comprised of twenty formal simulated job interviews (8–10 minutes each) and twenty informal, natural dialogues (10-14 minutes each) from the HuComTech audio-visual database (in Hungarian), along with their multi-level multimodal annotation, carried out by Praat (audio annotation) and Qannot (video annotation) software. The reader can find detailed information on the annotation system and process of the HuComTech corpus in Pápay et al. 2011 and in Staudt & Pápay 2011.

Audio annotation (Pápay 2010) involves five levels: intonational phrase type, emotional and discourse levels as well as the verbatim transcripts of the utterances of the speakers. For the purposes of the analysis of turn management, the most important level is the discourse level. The labels of the discourse level (Abuczki 2011) are: T (turn-take), K (turn-give), G (turn-give), BC (backchannel) and SIL (silence>250 ms).

In the framework of the HuComTech project, for the development and training of the turn type detector software and the demo of the job interview dialog management system, the most important floor control types to detect are: 1. turn give with speaker change (so that the computer agent can start his turn), and 2. turn keep (so that the computer agent can produce backchannels).

In order to account for the nonverbal behaviour of the interlocutors, video annotation is just as important as audio annotation. The video annotation software, called Qannot, was developed by Szilvia Szeghalmy at the University of Debrecen. The annotation of the nonverbal behaviour of the speakers involves the following levels and labels presented in Table 4. Many thanks are due to the annotators for their diligence, patience and precision in segmenting and labelling the recordings.

Level	**Group**	**Event**	**Attribute**
Basic	Communication	start, end	begin, end
Physical	Facial expression	natural, happy, surprised, sad, recalling, tensed	begin, end, intensity
	Gaze	blink, orientation (up, down, left, etc.)	begin, end, intensity
	Eyebrows	up, scowl	begin, end, side
	Head movement	nod, shake, turn, sideways, etc.	begin, end, orientation – optional
	Hand shape	open, half-open, fist, index-out, thumb-out, spread	begin, end, side
	Touch motion	tap, scratch	begin, end, touched part of body
	Posture	upright, lean, rotate, crossing arm, holding head, shoulder up	begin, end, orientation – optional
	Deictic	addressee, self, shape, object, measure	begin, end, side
Functional	Emotions	natural, happy, surprise,d sad, recalling, tensed	begin, end, intensity
	Emblems	attention, agree, disagree, refusal, doubt, numbers, etc.	begin, end

Table 4: Levels and labels of video annotation (Pápay 2010)

On the one hand, the most significant and most easily detectable visual–nonverbal cues (by an eye gaze tracking and hand movement tracking software as well) are 1. gaze direction, 2. the presence or absence of hand gesturing and 3. posture shifts. On the other hand, the most significant and most easily detectable acoustic cues are silences. However, it must be noted that silence is assigned different significance by conversation analysts depending on such factors as (1) its length and (2) where it occurs in a conversation. For instance hesitation pauses occur within a turn and switching pauses between turns, while a great number of topic boundaries are preceded by numerous longer gaps (Mey 1993). It must be added to the above distinction that it is also frequent to hesitate at turn boundaries, not only within a turn.

My work methodology included a qualitative microanalysis of ten dyadic conversations, accompanied by a gesture typology and analysis, and the quantitative comparison of the twenty informal and twenty formal dialogues in terms of the number of occurrences of turn-takings, backchannels, silences and hesitations.

The results outline a systematic description of the typical features of the four different turn types (turn-give G, turn-keep K, turn-take T, backchannel BC) annotated in the HuComTech project. The nonverbal cues for the different categories will be described in detail, with special regard to the role of gaze and gesturing.

turn-give (G):

The most reliable nonverbal cue for turn-give is the gaze behaviour of the present/previous speaker which can best be described as a long glance. The current floor holder continuously (longer than before) looks at the other conversation partner seeking a reply, a future action or any reaction. Gaze shifts toward the listener frequently coincide with a shift in conversational turn – they can be seen as a signal that the floor is available (Abuczki 2011).

As far as the phenomenon of gesturing is concerned, no gesturing can be observed during this turn type; the cessation of manual gesturing is especially typical of turn-gives (Duncan 1972). If the current floor holder spends a considerable percentage of time gesturing at the end of an utterance, he/she may be more likely to continue controlling the floor after the end of the utterance.

Regarding acoustic features, the most typical intonation pattern of turn-give is falling question intonation with decreasing pitch or loudness, decreasing intensity, followed by silence. These markers belong to the behavioural markers at the technological level of the Hunyadi-model (Hunyadi 2011).

Regarding the technological implementation of a spoken dialogue system, machine detectable cues of turn-give are silence, a pauselength of average 500 msec (preceded by x msec speech), very often coupled with the eye contact of the speakers which can be detected with eye gaze tracking software.

turn-keep (K):

During turn-keep the direction of gaze is relatively quickly changing. Surprisingly, looking away from the interlocutor has been correlated with the beginnings of turns (as also proposed earlier in Cassel et al. 1999). A possible explanation of speakers' tendency to often look sideways while speaking is that they try to avoid information overload. On the other hand, during recalls the speakers tend to look upwards which might well be seen as a signal of cognitive activity or cognitive (pre-) planning. Looking at the conversational partner or looking away from the partner can provide indirect cues of the speaker's willingness to

continue interaction, and gazing at particular elements in the vision field can tell where the speaker's focus of attention is.

The heaviest gesturing of the speaker can be observed during turn-keep, depending of course on personality types and individual variations of the speakers as well. As noted above, when a speaker is still making gestures at the end of an utterance, he/she may still intend to continue speaking and therefore will maintain control of the floor after the end of the utterance. Gestures can also be understood as by-products of lexical pre-planning processes which often precede verbal content – especially during recalls or listing; on the other hand, the stroke of a gesture sometimes coincides with the semantically most prominent information (Abuczki 2011).

turn-take (T):

As for eye gaze cues, to grab control of the floor more smoothly, the subsequent floor holder often looks at the current floor holder in order to establish eye contact. Most of the times no eye contact can be observed between the interlocutors when the speaker begins the utterance, perhaps for cognitive reasons. However, the speaker always establishes eye contact when she/he gives the floor over at the end of an utterance.

As for the posture of speakers, shifts occur more frequently at discourse segment boundaries (often at silences) than within discourse segments. Speakers most often generate a posture shift when initiating a new topic and starting a new discourse segment, so shifts in posture always mark shifts in the conversation structure as well.

backchannel (BC):

BC nonverbal behaviour includes head nodding as a prototypical example without any form of hand gesturing. However, nodding is often accompanied with facial expressions such as eyebrow raise. With continuous nodding often accompanying iterative phrases and short utterances the hearer indicates that he/she is taking note of what the speaker is saying. It can be seen as a response to the speaker's nonverbal request for feedback. Nodding as a signal of continuing attention is very frequent in the HuComTech corpus (Abuczki 2011). It has been observed and shown in the HuComTech corpus as well that listener nods precede their vocalizations (as proposed in McClave 1998). In Yngve's understanding (Yngve 1970), vocal or gestural expressions of the listener's BC signal that he/she does not wish or intend to assume the floor.

In sum, results of the comparative study of turn management indicate (1) correlation between gaze shifts, posture shifts and discourse segment boundaries and turns; and (2) considerable differences between formal (F) and informal (IF) dialogues with regards to the frequency of turn-takings (IF >F), backchannels (IF >F), silences (F>IF), hesitations (F>IF), gaze and posture shifts, presented in Table 5 in detail.

	Formal (job) interview (occurrence/minute)	Informal conversation (occurrence/minute)
Turn-taking	3,19497	4,09554
Backchannel	0,12678	0,61065
Silence (> 250 ms)	2,71319	2,17486
Hesitation	3,20258	1,71612

Table 5: Comparison of 20 formal and 20 informal interviews (Abuczki 2011)

Ágnes Abuczki:
A multimodal analysis of the sequential organization of verbal and nonverbal interaction
Argumentum 7 (2011), 261-279
Debreceni Egyetemi Kiadó

6 Conclusions

As far as the sequential organization of interaction and its representation in the pragmatic annotation system is concerned, the potential significance of the application of communicative acts and adjacency pairs in multimodal communication modelling can be seen from the perspectives of analysis and synthesis as it might contribute to the recognition of communicative acts, and then, in response, the automatic generation of appropriate replies. Since an initiating act in a dialogue creates a strong normative pressure for certain types of response (Bunt & Black 2000), the results of the pragmatic annotation might bring out dialogue grammar rules regarding the organization of communicative acts. In the modelling of interpersonal as well as human–computer interaction, dialogue grammar rules might be applied in dialogue management systems as they might help the probabilistic automatic recognition of dialogue acts in spoken dialogues.

However, the first step before the identification of communicative acts performed is the automatic recognition of the turn-give intention of the speaker. My work methodology included a qualitative microanalysis of ten dyadic conversations accompanied by a gesture typology and analysis, and the quantitative comparison of the twenty informal and twenty formal dialogues in terms of the number of occurrences of turn-takings, backchannels, silences and hesitations. Having carried out the qualitative and quantitative analysis of the different floor control types and the corresponding visual and acoustic cues in twenty formal and informal dialogues of the HuComTech corpus, I identified the following tendencies of turn management strategies in the recordings: (a) the acoustic and suprasegmental features of turn-take involve the acoustic realizations of interruption, often accompanied by posture shifts and sometimes by nodding; (b) turn-keep is accompanied by the increase of fundamental frequency, higher pitch excursion and higher intensity while the speaker often looks at the interlocutor or looks up during recalls; (c) turn-give is often expressed by a question, accompanied by eyebrow raise and long gaze at the other speaker, followed by silence; and (d) backchannel is often accompanied by nodding and the acoustic realizations of iteration and humming. The results of this study on turn management markers and the most common end-of-turn features might be systematically organized into a decision tree and a learning algorithm might be developed for deciding whether a spoken dialogue system should take turn when a user is pausing (for a certain length of time).

The results of the quantitative analysis (in Table 5) confirm my work hypothesis that turn management strategies differ in formal job interviews and informal conversations in terms of the frequency of turn-takings (IF>F), backchannels (IF>F), silences (F>IF) and hesitations (F>IF), which shows a typical pattern for these two distinct genres of interaction. Consequently, as expected, we have found more instances of linguistic as well as nonverbal adaptation (Verschueren 1998) of the speakers to each other in the informal dialogues, expressed by the higher frequency of backchannels and posture mimicry.

In the course of the qualitative gestural analysis of the recordings I found that gestures not only structure the flow of the interaction but, naturally, they also reflect the emotions of the interlocutors. For instance, the position of the eyebrows of the listener often conveys information about the emotional state of the listener:

- (half) raised eyebrows often indicate surprise
- fully raised eyebrows may express disbelief

– (half) lowered eyebrows often signal puzzlement
– fully lowered eyebrows may indicate anger

As for the speaker's facial nonverbal behaviour, eyebrow movements made by the speaker tend to coincide with the most prominent and (semantically) new information, which is in accordance with what has been said before. Therefore, eyebrow tracking, the involvement of the automatic recognition and the analysis of eyebrow movements in facial feature detection software would be very useful for future developments in spoken dialogue management systems and naturalistic HCI.

In order to design a naturalistic technological implementation of human–machine interaction it is necessary to make explicit the joint occurrences of verbal codes, nonverbal codes and ostensive behavior without code-using with their surface markers (cf. Hunyadi 2011) which are suitable for making algorithms for computers.

7 Future plans

As soon as the annotation process is completed, I am going to carry out a detailed analysis of the results of the pragmatic annotation in order to find patterns in the sequential distribution of communicative acts and see how they are locally put together. This will be followed by a statistical analysis of the correlation of audio, video and pragmatic annotation labels in order to find out how and when the gestural and other nonverbal components interact with the verbal. Based on the results of the annotation, I would also like to find surface features that are probabilistically associated with certain communicative acts.

The long-term prospect of this study involves contribution to improvements in human-computer interaction technologies, especially the naturalness of dialogue management systems so that human-like conversational agents can predict the turn-giving and turn-taking intention of the other (human) speaker, as well as exhibit appropriate prosodic and nonverbal behaviour: silences, hesitations, eye-gaze movements, posture shifts and gestures orchestrated and harmonised at the appropriate points of the conversation.

References

Abuczki, Á. (2011): Multimodal Annotation and Analysis of Turn Management Strategies: A Comparative Study of Formal and Informal Dialogues of the HuComTech Corpus. In: Boda, I.K. & Mónos, K. (eds.): *MANYE XX. Az alkalmazott nyelvészet ma: innováció, technológia, tradíció.* Debrecen: Debreceni Egyetemi Kiadó, 282–288.

Abuczki, Á., Bársony, O., Bódog, A. & Németh T., E. (2011): *Pragmatic annotation guidelines of the HuComTech project*. Debrecen: IKUT/HuComTech documentation.

Abuczki, Á., Bódog, A. & Németh T., E. (2011): Theoretical Basis of Pragmatic Annotation in the Modelling of Human–Computer Communication. Presented at: *The Theoretical Fundamentals of Human–Computer Interaction Technologies Conference.* Aquaticum Thermal and Wellness Hotel, Debrecen, May 25–27.

Allwood, J. (2000): An Activity Based Approach to Pragmatics. In: Bunt, H. & Black, B. (eds.): *Abduction, Belief and Context in Dialogue; Studies in Computational Pragmatics.* Amsterdam: John Benjamins.

Austin, J.L. (1962): *How to Do Things with Words.* Boston: Harvard University Press.

Bódog, A. (2011): Pragmatic Annotation in the HuComTech Corpus: Methods and Preliminary Results. Presented at: *The Theoretical Fundamentals of Human–Computer Interaction Technologies Conference.* Aquaticum Thermal and Wellness Hotel, Debrecen, May 25–27.

Boersma P. & Weenink, D. (2007): *Praat: doing phonetics by computer 5.0.02.* University of Amsterdam: Institute of Phonetic Sciences. http://www.praat.org.

Bogdan, C., Kaindl, H., Falb, J. & Popp, R. (2008): Modeling of interaction design by end users through discourse modeling. In: *Proceedings of the 13th international conference on Intelligent user interfaces*. New York: ACM.

Bunt, H. & Black, W. (2000): The ABC of Computational Pragmatics. In: *Abduction, Belief and Context in Dialogue: Studies in Computational Pragmatics.* Amsterdam: John Benjamins.

Bunt, H., Alexandersson, J., Carletta, J., Choe, J., Fang, A.C., Hasida, K., Lee, K., Petukhova, V., Popescu-Belis, A., Romary, L., Soria, C. & Traum, D. (2010): Towards an ISO standard for dialogue act annotation. In: *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC '10, Valletta).*

Cassell, J., Torres, O. & Prevost, S. (1999): Turn Taking vs. Discourse Structure: How Best to Model Multimodal Conversation. In: Wilks, Y. (ed.): *Machine Conversations*. Hague: Kluwer, 143–154.

Duncan, S. (1972): Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology* 23.2, 283–292.

Ekman, P. & Rosenberg, E.L. (1998): *What the Face Reveals*. Oxford: Oxford University Press.

Enfield, N.J. (2009): *The Anatomy of Meaning. Speech, gesture, and composite utterances.* Cambridge: Cambridge University Press.

Goodwin, C. & Heritage, J. (1990): Conversation analysis. *Annual Review of Anthropology* 19, 283–307.

Goodwin, C. (1981): *Conversational Organization: Interaction Between Speakers and Hearers*. New York: Academic Press.

Hunyadi, L. (2011): Multimodal Human–Computer Interaction Technologies. *Argumentum* 7. (submitted for publication)

Have, P. ten (2007): *Doing Conversation Analysis. A Practical Guide.* London: Sage.

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

Kendon, A. (1972): Some relationships between body motion and speech. In: Seigman, A. & Pope, B. (eds.): *Studies in Dyadic Communication*. Elmsford, New York: Pergamon Press, 177–216.

McClave, E. (1998): Cognitive and Interactional Functions of Head Movements in Conversation. In: Santi, S., Guaitella, I., Cavé, C. & Konopczynski, G. (eds.): *Oralité et Gestualité.* Paris: L'Harmattan, 365–369.

McNeill, D. (1996): *Hand and Mind: What Gestures Reveal about Thought*. Chicago: University Of Chicago Press

Mey, J.L. (1993): *Pragmatics. An Introduction.* Oxford: Blackwell.

Németh T., E. (1996): A szóbeli diskurzusok megnyilatkozáspéldányokra tagolása. *Nyelvtudományi Értekezések* 142. Budapest, Akadémiai Kiadó.

Németh T., E. (2005): Az osztenzív-következtetéses kommunikációtól a verbális kommunikációig. In: Ivaskó, L. (ed.): *Érthető kommunikáció*. SZTER, Médiatudományi Tanszék, 77–87.

Németh T., E. (2011): On Pragmatic Models of Human Multimodal Communication. Presented at: *The Theoretical Fundamentals of Human–Computer Interaction Technologies Conference.* Aquaticum Thermal and Wellness Hotel, Debrecen, May 25–27.

Okrent, M. (2002): Equipment, World, and Language. *Inquiry* 45.2, 195–204.

Pápay, K., Szeghalmy, Sz. & Szekrényes, I. (2011): HuComTech Multimodal Database Annotation. *Argumentum* 7. (submitted for publication)

Pápay, K. (2010): Using the HuComTech Multimodal Database for Improving a Hungarian Prosodic Segmenter. Presented at: *The PINK COST 2102 International Conference on Analysis of Verbal and Nonverbal Communication and Enactment: The Processing Issues.* Budapesti Műszaki Egyetem, Budapest, September 7–10.

Sacks, H. (1992): *Lectures on Conversation.* Oxford: Blackwell.

Sacks, H, Schlegoff, E.A. & Jefferson, G. (1974): A simplest systematics for the organization of turn-taking in conversation. *Language,* 50.4, 696–735.

Schegloff, E.A. (2006): *Sequence organization in interaction: A primer in conversation analysis.* Cambridge: Cambridge University Press.

Searle, J. (1969): *Speech Acts: An Essay in the Philosophy of Language.* Cambridge: Cambridge University Press

Sidnell, J. (2010): *Conversation Analysis. An Introduction*. Chichester: Wiley-Blackwell. 36–57.

Silverman, D. (1998): *Harvey Sacks – Social Science and Conversation Analysis*. Cambridge: Polity Press, 97–210.

Staudt, A. & Pápay, K. (2011): The Annotation of the HuComTech Audio Database in Practice – Observations and Questions Arising. *Argumentum* 7. (submitted for publication)

Yngve, V. (1970): On getting a word in edgewise. In: *Papers from the Sixth Regional Meeting. Chicago:* Chicago Linguistic Society.

Verschueren, J. (1998): *Understanding Pragmatics.* London: Arnold Publishers.

Wiemann, J.M. & Knapp, M.L. (1975): Turn-taking in Conversation. *Journal of Communication,* Spring, 75–92.

Ágnes Abuczki
University of Debrecen
Department of General and Applied Linguistics
Pf. 24
H-4010 Debrecen
abuczki.agnes@gmail.com