Kinga Pápay, Szilvia Szeghalmy & István Szekrényes: HuComTech Multimodal Corpus Annotation
Argumentum 7 (2011), 330-347
Debreceni Egyetemi Kiadó

Working paper

Kinga Pápay, Szilvia Szeghalmy & István Szekrényes

HuComTech Multimodal Corpus Annotation[1]

Abstract

The Hungarian audio-visual corpus recording and annotation project is being carried out by the HuComTech (Hungarian Human-Computer Interaction Technologies) research group at the University of Debrecen and is a part of a comprehensive multimodal human-machine interaction modelling project. The research contributes to the exact knowledge of the overlaps between the verbal and nonverbal aspects of communicative events and prosodic features through the examination of spontaneous speech, with special regard to syntactical embeddings, insertions, iterations, hesitations and restarts, various kinds of emotions and discourse markers. The efficiency of speech recognition systems can also be increased by proper acoustic preprocessing and investigation of the suprasegmental characteristics of spontaneous speech. Concerning Hungarian, the lack of a multimodal, prosodically labelled, representative spontaneous speech corpus makes the development more difficult. The spontaneous multimodal corpus is being recorded via guided – formal and informal – conversations. During the conversations, several points are to be discussed in order to provoke longer monologues accompanied by gestures, facial expressions, and also including the above mentioned phenomena of spontaneous speech to be examined. The present paper focuses on the aspects of multimodal annotation, especially the details of the annotation of prosodic and suprasegmental features. The visual, nonverbal channel of these phenomena are also to be annotated thus we can examine and implement multimodal features and their overlaps as well.
Keywords: Hungarian audio-visual corpus recording, audio annotation, video annotation

1 Introduction

The Hungarian audio-visual corpus recording and annotation project is being carried out by the HuComTech (Hungarian Human-Computer Interaction Technologies) research group at the University of Debrecen and is a part of a comprehensive multimodal human-machine interaction modeling project. The research contributes to the exact knowledge of the overlaps between the verbal and nonverbal aspects of communicative events and prosodic features through the examination of spontaneous speech, with special regard to syntactical embeddings, insertions, iterations, hesitations and restarts, various kinds of emotions and discourse markers. One of the main goals of the corpus construction is to be able to examine the properties of spontaneous speech, for example the characteristics of hesitations, hummings (Markó 2005) and backchannels. Another aim of the project is to compare the characteristics of formal and informal communication. The efficiency of speech recognition

[1] The corpus construction is a part of the *Theoretical fundamentals of human-computer interaction technologies* project (TÁMOP-4.2.2-08/1/2008-0009).

systems can also be increased by proper acoustic preprocessing and investigation of the suprasegmental characteristics of spontaneous speech. Concerning Hungarian, because of the lack of a multimodal, prosodically labeled, representative spontaneous speech corpus, the first step of the project was the corpus design and collection, the second step is its annotation. Though it is important to note, that two spontaneous speech corpora do exist for Hungarian – one is BUSZI (Váradi 1998) and the other is BEA (see a short description here: http://www.nytud.hu/dbases/bea/index.html) –, but none of them is available for external researchers because of legal and ethical reasons, and the other problem is that they do not contain video recordings, so they can not be used for multimodal research, though multimodal corpora have been recorded in several other languages in the past years. From these, we are highlighting some of them, which are to some extent similar to our corpus.

1.1 Multimodal corpora

A well-known example of a large multimedia corpus is the AMI project's (2004) meeting corpus. The aim of the project is to develop meeting browsers and to help group members to be able to efficiently join the meeting even if they are late. The AMI meeting corpus includes 100 hours of meeting data, which was collected between 2004 and 2005. The majority of meetings were elicited using a scenario whereby groups of four participants played different roles in a corporate design team. The data was collected in three smart meeting rooms at IDIAP in Switzerland, the University of Edinburgh in Scotland and TNO Human Factors Institute in the Netherlands. They used 4 cameras, 24 microphones in each room and special tools to capture handwriting and slides. Regarding speech annotation, the utterances are segmented to sentences or phrases (breakpoints are inserted at different natural linguistic points) and similar rules were used for transcribing as our rules (see a short description in 3.4). The AMI dialog act annotation is about the type of intention the speaker has – each time a new intention is expressed, it is marked as a new segment; backchannels, stalls, fragments and different types of speech acts are labeled. As for the AMI affect annotation, the task of the annotators was to annotate the video recordings with information about the mental state of the participants. The following states were distinguished based on the recordings: curious, amused, distracted, bored, confused, uncertain, surprised, frustrated, decisive, disbelief, dominant, defensive and supportive.

Another good example for emotional corpus is EmoTV (2005), as part of the HUMAINE (Human-Machine Interaction Network on Emotions) project, which is an audiovisual corpus collected for studying everyday life contexts and emotions. The modeling of emotional behavior is needed for various applications in signal processing, such as emotion detection in a surveillance system or the design of animated interactive characters also called Embodied Conversational Agents (ECAs). The EmoTV corpus is in French and consists of 51 clips with 48 different subjects. The total duration of the corpus is 12 minutes (average length of 14 seconds per clip), its lexicon size is 800 words (the total number of words is 2500). They enabled the annotation of each segment with two emotional labels and proposed a topology of non-basic emotional patterns: blended emotions (two emotions are merged, and occur at the same time), masked acted emotions (like a smiling with a real disappointment behind), sequence of emotions (one occurring shortly after the other, in a single emotional segment), cause-effect conflict emotion (e.g. positive/negative conflict, cry for joy and relief) and ambiguity between two emotions in the same class (e.g. anger and irritation).

The MUMIN Multimodal Coding Scheme (Allwood et al. 2005) was originally created to experiment with annotation of multimodal communication is short clips from movies and in video clips of interviews taken from Swedish, Finnish and Danish television broadcasting. However, the coding scheme also intends to be a general instrument for the study of gestures and facial displays in interpersonal communication, in particular the role played by multimodal expressions for feedback, turn management and sequencing. Some parts and categories of this scheme were adopted to our video annotation scheme, e.g. handedness (both hands/single hands), trajectory (up, down, sideways), gesture types (e.g. deictic), the movements of the eyebrows, eyes, gaze and head. A main difference is that we are using emotional categories independently from feedback categories. Our turn categories are also similar, but the annotation is based only on the audio material and we do not distinguish the different types of them (e.g. whether the turn is taken with an interruption or is accepted after it was yielded by the other speaker, we only mark that it was a turn-gain, end or hold) and only the fact of the feedback is marked as backchannel – see 3.3).

The German SmartKom (2003) and TALK (2007) projects are examining new possibilities of the interaction between human and machine. Their data were collected in so-called Wizard-of-Oz experiments: the subjects had to solve certain tasks with the system. They were made believe that the system they interacted with was fully functional, but actually many functions were only simulated by humans who controlled the system from another room. In the SmartKom project 4 microphone arrays, a directional microphone and a headset or a clip-on microphone were used for the audio recordings. A digital camera was used to capture the facial expressions of the subjects and a second digital camera captured a side view of the speaker for the gestures, and an additional infrared camera captured the hand gestures. They also recorded the coordinates of pointing gestures as well as the inputs of a pen on the graph tablet. The recorded spontaneous speech (the dialog between user and machine) is labeled on the word level using a broad orthographic transliteration system. As for the video annotation, a simplified, practice-oriented system was used, two broad categories are labeled: head gestures and hand gestures. The hand gestures are defined functionally/intentionally (not morphologically), with regard to the intention of the user's assumed goal. The head gestures are coded with regard to three morphological categories: head rotation, head incline forward/backward, head incline sideward. The emotional facial expressions are labeled in six categories: anger, boredom, joy, surprise, neutral and face partly not visible. These are also similar to our categories and we have almost fully adopted the TALK project's speech transcription rules and symbols (see 3.4) , and the clause or clause-like units as markables or segments of the first audio layer (see 3 and 3.1).

2 Audio and video material

2.1 Speakers

The HuComTech corpus aims to represent Hungarian university students and workers, therefore the speakers of the corpus are the same kind of people, with appropriate gender, age and birth place distribution. Currently, we have a multimodal corpus of 121 speakers, as for gender distribution, 44.6% of the speakers are women and 55.4% are men, which corresponds to our previous plans. The age of the speakers is between 19 and 30, none of the speakers are over 30. Distribution according to age group can be seen in Figure 1. Most of the speakers (46%) is 20 or 21 years old.

Kinga Pápay, Szilvia Szeghalmy & István Szekrényes: HuComTech Multimodal Corpus Annotation
Argumentum 7 (2011), 330-347
Debreceni Egyetemi Kiadó

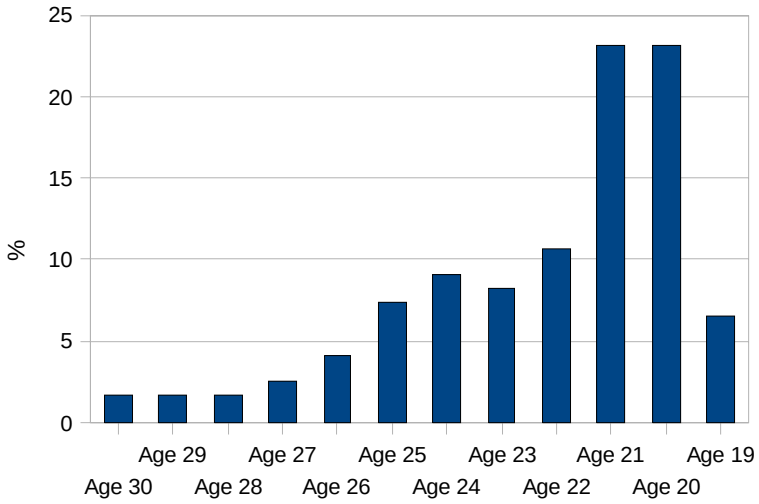

Figure 1: Age distribution

Regarding the origin distribution, Figure 2 shows that most of the speakers, namely 40.5% are from Debrecen, 10.7% are from Nyíregyháza, 5-5% are from Miskolc and Szolnok, 4.1% are from Berettyóújfalu, and 3.3%-3.3% are from Budapest and Eger. Others are from different parts of the North-Eastern part of the country and only a few of them are from the Western and from the Southern part. Thus the corpus rather represents the North-Eastern part of Hungary.

Figure 2: Origin distribution

As for the speakers' skin, hair and eye colour distribution (it is important to consider these rates when we use the video data for testing different video image processing techniques), most of the speakers, 74.4% have intermediate (between light and dark) skin, 22.3% have light skin, and the rest have dark or very light skin colour. The hair colour distribution is:

52.1% are brown, 28.1% are dark blond, 14% are blond and the rest are black, sienna or grizzled. Eye colour distribution is similar, 48.8% have brown, 20.7% have blue, 19% have green eyes and the rest have teal, hazel or grey eyes.

2.2 Studio equipment

The studio (see Figure 3) is equipped with an adequate PC, recording software (Sound Forge Pro 10) and 2 far-talk cardioid microphones (Shure 16 A). The ideal position of the microphones is next to the speakers, not too close and not too far from each speaker. We record stereo .wav files with 44,1 kHz sampling frequency and 16 bit quantization.

The studio is equipped also with 1 HD camera (Sony HDRXR520VE) recording the agent and 2 HD plus 2 web cameras (Logitech Webcam Pro 9000) directed to the speaker to record his/her face and hand gestures, and appropriate lights. All the cameras record sound as well, synchronization is managed using flash lights and beep sounds during the recordings. The file format of the HD cameras is .mts, while the file format of the web cameras is .jpg.

Figure 3: Studio equipment (left: agent side, right: speaker side)

2.3 Audio and video contents

Each speaker has 3 short tasks: First, the speaker reads out 20 phonetically rich sentences and 7 words (this is needed in case of a continuous SRE to cover all the phoneme variations of the language and also to ease and prepare the speaker for the spontaneous dialogues) plus embedded sentences. Second, we record the main part, the spontaneous dialogues – regarding the differences between real and acted speech (Wilting et al. 2006), we have decided not to record acted speech. The informal dialogues are about natural topics, mostly about university and other life experiences. The questions are also intending to provoke emotions. The agent starts the conversation with the less personal questions and progresses towards the more personal topics. The transition from one question to the other is as smooth as possible. Although the guided conversation contains a lot of dialogues, it contains as many long stretches of speech from the speaker as possible; instead of yes/no questions, information seeking Wh-questions (why …, when …, what happened …, please tell me …) are used. In order to provoke backchannels and more spontaneous interaction, during the informal conversation, the agent tells his/her own stories (i.e. the agent behaves as an equal partner in the conversation). Third, the formal dialogues of the corpus are produced via simulated job

interviews where the agent is the interviewer and the speaker is the interviewee. When the recording process is over, the agent records speaker related data (age, sex, dialectal region/city, identification number). Time demand per speaker is 30 minutes, out of which we record cca 4 minutes of reading and 26 minutes of dialogues.

3 Speech annotation

Annotation means transcription of the utterance and some segmentation by using timestamps. For speech annotation, we examined the possible tools (see a review in Pápay 2010) and finally decided to use the Praat speech analyser program (Boersma & Weenink 2007); see Figure 4. Counting with 100 speakers, 50 hours of audio and video material has to be annotated.

After examining other corpora and annotation methods, or prosodic analysis methods (Beckman et al. 1992, Burkhardt et al. 2005, Douglas-Cowie et al. 2003, Hirschberg 2007, Vicsi & Sztahó 2009, Varga 1999-2001, 2002), we have created an annotation method for our research purposes. Annotators must produce a verbatim (word-for-word) transcript of everything that is said within the file. The words transcribed within each segment boundary must correspond exactly to the timestamps that have been created by the segmentation, so that the audio file is aligned with the transcript. The main role of the transcription is that all well audible sound events (speech and eventual noises) should be marked. Speaker data is to be fixed in the file naming. The following types of speaker data are to be fixed: speaker ID, sex of the speaker (male or female), age of the speaker and place of origin – based on this, the file naming convention is the following: 001fc25_F_shure, where the three digit code stands for the speaker ID, f or m stands for sex, a two digit code stands for the speaker's age, F or I stands for formal or informal conversation style and shure stands for the microphone type. The audio file and the file holding its transcription has the same name (of course the file extensions are different). As mentioned above, segmentation and transcription are closely related. Timestamps are placed within the transcription by using Praat's segmentation utility. Segmentation (timestamp placement) is necessary for segmenting speech into shorter parts. Sentence, head clause and subordinate clause boundaries are detected automatically by running a phrase boundary detector program (developed by Technical University of Budapest, Laboratory of Speech Acoustics) on the recordings. The next step of the annotator is manual checking of automatically generated segments, exact marking of phrase boundaries and assigning label symbols (matching the appropriate intonational, emotional and discourse phrase type) to each phrase.

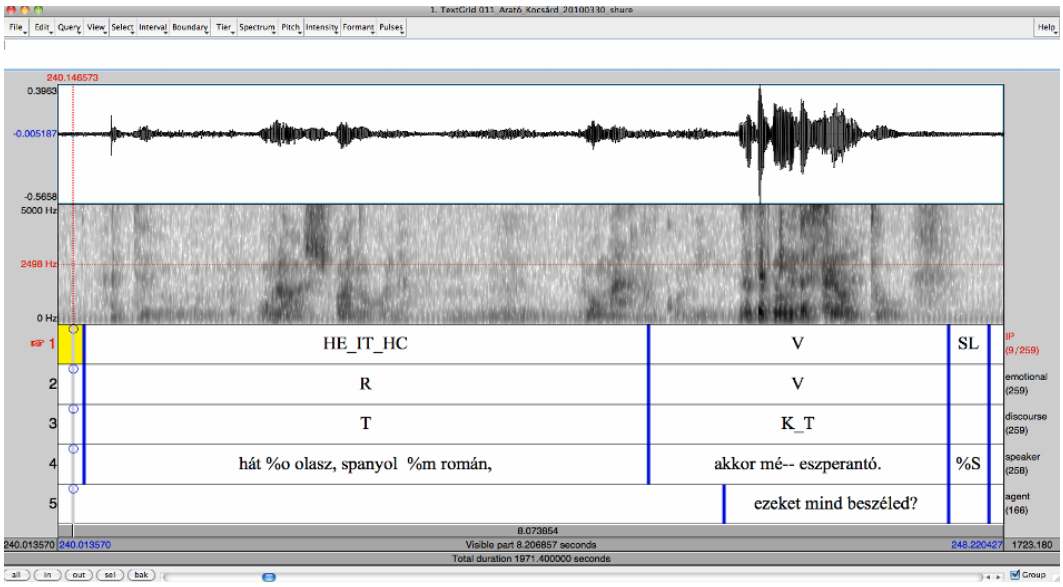

Figure 4: Using Praat for a five-level speech annotation

Speech annotation process is simultaneous at five levels (see Table 1): three functional and two transcription levels of the dialogue. The functional level has three sub-levels: intonational, emotional and discourse phrase types. With the sound files and transcriptions of the intonational phrase level and emotional level, the phrase detector program and the emotion recognizer of BME Laboratory of Speech Acoustics (Németh et al. 2007, Vicsi & Szaszák 2008, 2010, Szaszák 2009, Tóth et al. 2007, Vicsi & Sztahó 2009) or Seppänen et al. (2004) can be retrained and thus improved for recognizing the phrase boundaries and the emotions of spontaneous speech. Labels are abbreviations of intonational phrase types, emotional and cognitive state types and discourse phrase types. These are summarized in Table 1 and described below in detail.

Level 1: **Intonational phrases** (IP labels)	Level 2: **Emotions** (emotional labels)	Level 3: **Diaogue turns** (discourse labels)	Level 4: **Transcription of speaker's speech** (speaker text)	Level 5: **Transcription of agent's speech** (agent text)
HC (head clause)	**N** (neutral)	**T** (turn-take)	text + symbols	text + symbols
SC (subordinate clause)	**S** (sad)	**G** (turn-give)		
EM (embedding)	**H** (happy, laughing)	**K** (turn-keep)		
IN (insertion)	**P** (surprised)	**B** (backchannel)		
BC (backchannel)	**R** (recalling, thinking)			
HE (hesitation) + linking, e.g. HE_HC1	**T** (tensed)			
RE (restart) + linking	**O** (other)			
IT (iteration) + linking				
SL (silence) + linking	**SL** (silence)	**SL** (silence)		
V (overlapping speech)	**V** (overlapping speech)	Overlapping speech, e.g. K_T		

Table 1: Audio annotation levels

3.1 Labels of intonational phrase types

The following categories are established based on Hunyadi's (2006, 2009, 2010) research on embeddings, iterations and insertions. Keszler (1989) did also make some interesting remarks on the acoustic properties of insertions: the inserted clauses have lower fundamental frequency, monotonous intonation and faster speech rate than the neighbouring ones.

HC = head clause – in case of embeddings or insertions, annotators mark the divided clauses with HC1, HC2
SC = subordinate clause – in case of embeddings or insertions annotators use SC1, SC2
EM = embedding
IN = insertion
BC = backchannel

HE = hesitation – annotators use an underscore _ for linking the clause type in which it occurs, e.g. HE_SC
RE = restart – annotators use an underscore _ for linking the clause type in which it occurs, e.g. RE_HC1
IT = iteration – annotators use an underscore _ for linking the clause type in which it occurs, e.g. IT_IN
SL = silence – annotators mark it only if it exceeds 250 ms. If it occurs within a clause, annotators use an underscore _ for linking the clause type in which it occurs, e.g. SL_HC

In case of overlapping speech, annotators mark the given clause with a V. When some words in the clause are chopped off, and there are no restarts in the given clause, it is marked as HC-, SC- etc. One clause can contain more phenomena, in this case each of them is separated by an underscore _ (without spaces).
Comparing to Pápay 2009, there are some changes in the label categories on the IP (intonational phrase) level.

3.2 Labels of expressed attitudes

Annotated expressed attitudes are the following: neutral, happy, surprised, sad, tensed, recalling/thinking and other. See the labels below:

N = neutral
S = sad
H = happy, laughing
P = surprised
R = recalling, thinking
T = tensed
O = other
SL = silence – annotators mark it only if it exceeds 250 ms

In case of overlapping speech, we use the label V.

3.3 Labels of discourse types

Besides, we also implement discourse-level annotation using the following labels: turn-taking (T), turn-giving (G), backchannel (B), turn-keeping (K).

T = turn-take
G = turn-give
K = turn-keep
B = backchannel
SL = silence – annotators mark it only if it exceeds 250 ms

In case of overlapping speech, annotators use the T/G/K/B_T/G/K/B convention's first part for the speaker, second part for the agent.

3.4 Transcription levels

We use two separate levels for the two participants of the discourse; level four is for the speaker (interviewee) and level five is for the agent (interviewer). In case of a discourse, the speech transcription alternates between the two levels. As far as the orthographic annotation is concerned, the following fields are to be considered: spelling, capitalization, punctuation, numbers, acronyms, spoken letters, disfluent speech (including hesitations, partial words, restarts and mispronounced words), noise (including speaker and external noises) and hard-to-understand sections. Regions of disfluent speech are particularly difficult to transcribe. Speakers may stumble over their words, repeat themselves, utter partial words, restart phrases or sentences, and use lots of hesitation sounds. Annotators must take particular care in sections of disfluent speech to transcribe exactly what is spoken, including all of the partial words, repetitions and filled pauses used by the speaker. Table 2 summarizes our solutions (the symbol system is adapted from the TALK 2007 project) for how to mark the above mentioned speech phenomena.

Condition	**Markup**	**Example**	**Explanation**
Numbers	spelled out	nyolcszázöt	Write out full text, not digits.
Punctuation	comma, question, explanation, period	, ? ! .	Do not use other symbols.
Acronyms	@	@MÁV, @DE-BTK	Letters in caps, no space between
Spelling	~	~B ~M ~E	All with spaced caps
Filled pause, pause	%	%o, %m, %s	Filled pauses limited to these 2 items, and signing each lengthened character
Partial words	--	természe--	Transcribe as much of the word as you hear. No spaces preceding/following the word!
Restart	< >	azt hi-- <azt> hiszem	Use it if speaker stops and restarts
Mispronunciation	+	+pszichológus	Uncorrect pronunciation. Note: non-standard, but correct pronunciations are to be accepted!

Speaker noise	{}	{b} {c} {l} {s} {l}	Non-phoneme sounds produced by the speaker. Use only these 5 categories! Mark up only well audible speaker noises.
Instantaneous non-speaker noise	[ ] [b] – for beep sounds	mit [mondasz]?	Short intermittent noise. Mark up only well audible noises.
Semi-intelligible speech	((transcript))	itt van a ((szomszédban))	If you are uncertain about what is said
Unintelligible speech	(( ))	(( ))	If you do not understand what is said
Idiosyncratic words	*	*drrr	Made-up word
Foreign word	[Language: text] [foreign]	[Hunglish: you tube-ról]	For foreign sentences, use only [foreign] and quarantine them by using timestamps.

Table 2: Summary of transcriptional level symbols

3.5 Second passing

Second passing is used as a quality control measure to ensure the accuracy of segmentation, transcription (including markup), and speaker identification. After the initial file has been fully segmented and transcribed, a new annotator listens to the entire recording while viewing the corresponding transcript, and makes adjustments to the timestamps or transcription as needed. Second passing entails a mix of manual and programmatic checks on the transcript files. The particular types of checks conducted during second passing are described below.

Second pass annotators verify that each timestamp matches the corresponding transcript or label exactly. Annotators play each timestamp in turn and make sure that the audio, video transcript and labels for that segment are an exact match and make any necessary corrections. Annotators also check that the timestamp has been placed in a suitable location – between phrases, sentences, or breaths – and that the timestamp does not chop off the start or end of any word.

During the transcript checking phase of second passing, annotators examine the transcript in detail, checking for accuracy, completeness and the consistent use of transcription conventions. Annotators pay particular attention to a handful of areas that are particularly difficult to transcribe, in particular unintelligible speech sections and areas of speaker disfluency. Any proper names whose spelling could not be verified during the initial transcription process are corrected and standardized within the file.

4 Technical implementation

4.1 Preprocessing

First of all, we had to prepare the audio recordings for the annotation method. It means that audio files have to be converted (from stereo to mono) and cut along their different parts because of the following reasons:

- computer memory limitations
- to separate the different parts of the recordings (phonetically rich sentences and words, embedded sentences, formal dialogues, informal dialogues)
- to remove the unwanted parts of the recordings
- to make the annotators' work easier
- to synchronize audio and video recordings

We included *Sox* command line converter program into a Unix script, which converts and stores audio files in a given path. The output contains the following audio properties: 44100 Hz sample rate, 16 bit rate, 1 audio channel.

During the cutting method, we use a Praat script, which automatically stores the timestamps of the cutting points in a structured text file (the recordings contain "beep" signs to make this task easier). This method is very useful, because the video recordings have to be cut along the same points as the audio recordings after synchronization.

The next step is the automatic segmentation: a preprocessing script made by BME Laboratory of Speech Acoustics (Vicsi & Szaszák 2008) is used for segmenting speech recordings automatically. The script was written in Perl and based on HTK (Hidden Markov Model Toolkit, Young et al. 2005). We installed them (HTK source and Active Perl) under Unix, which is our preferred operating system for these preprocessing tasks. The BME script segments speech recordings to sentences and clauses, and uses the following categories:

- declarative sentence (S)
- pause (U)
- declarative clause (T)
- open question (K)
- yes or no question (E)
- imperative and exclamatory sentences (FF)
- optative sentence (O)
- bad sample (e.g. nosie) (R)

The outputs are stored in *.rec files, which must be converted to Praat .TextGrid format, which is used during the manual annotation. The converter program is an improved C Sharp code that adds more tiers (annotation levels) to the .TextGrid files. A Unix shell script was also developed to make the preprocessing method.

This part of the preproccessing was later canceled and became a future plan, because it turned out that the BME segmenter works well only if it is trained with spontaneous speech instead of the currently used read speech.

4.2 Annotation with Praat scripts

The biggest part of the annotation process has to be solved manually with the Praat speech processing program. To make this manual annotation method easier, we use Praat scripts, which are written in Praat's own interpreter language.

In the main script, the path of the annotations can be added; and the script checks if there are any existing annotations (a .TextGrid file, which might be a BME script output or a suspended work) with the same name as the selected speech recording. If there are any, the

script automatically opens the already existing transcription with the .wav file in the Praat editor window. In the other case, the script generates a new transcription file, which contains the tiers added in the window (see Figure 5). The annotator can choose the "using autosave" option, in this case the work is saved in the given annotation path automatically, as soon as the "Continue" button is clicked.

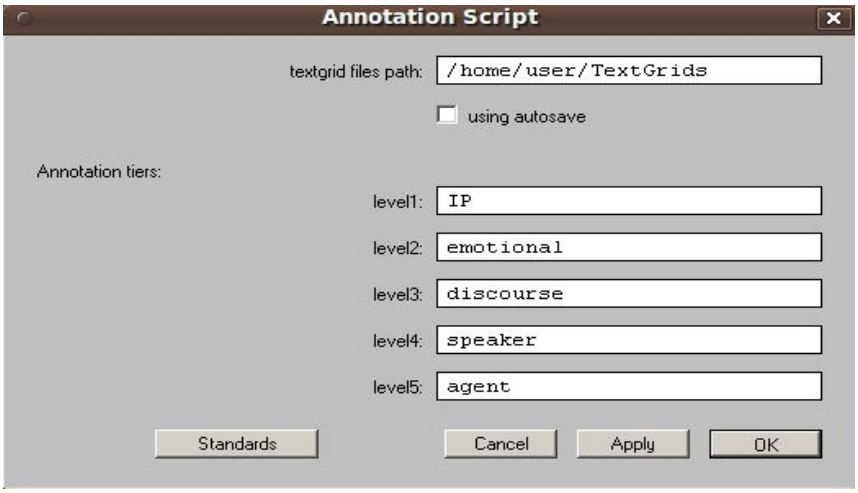

Figure 5: Generating transcription levels with a Praat script

4.3 Checking works

Another Praat script was also developed, which can check the annotations, e.g. the segmentation and the syntax of them. For instance, the script can check whether the annotators use only the allowed symbols. Some correlation between the labels can also be checked but only if the segmentation is already corrected, which means that the 2nd, 3rd and the 4th transciption levels have the same segmentation as the first level has. If any mistake is found, the script puts a sign to that place where the mistake is found in the .TextGrid file. To implement this error signal, two additional tiers are added into the .TextGrid file. The segmentation errors are signed in the first additional tier with the number of the tier where the error was found. The syntactical errors are signed in the second additional tier in the same way (see Figure 6).

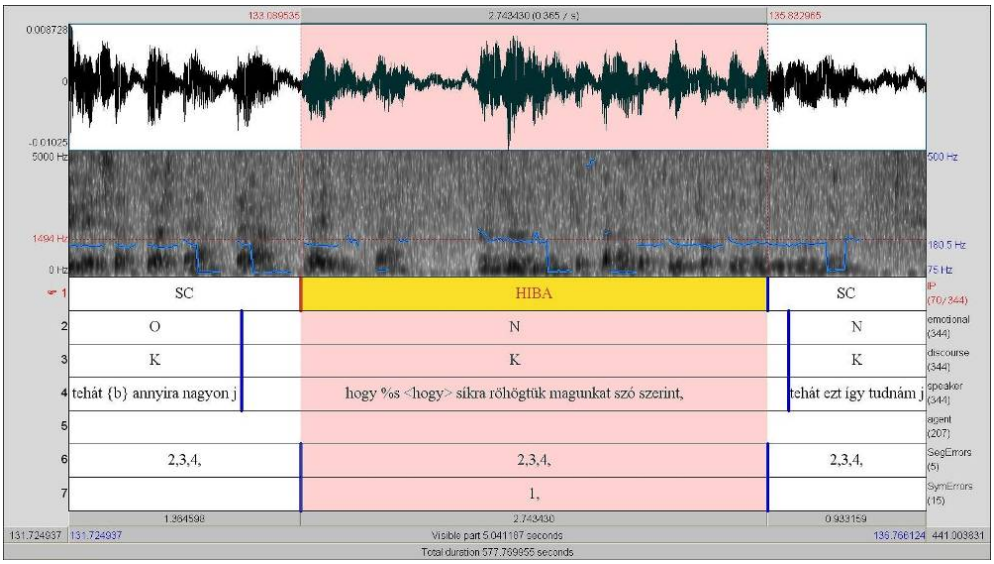

Figure 6: Displaying mistakes in the .TextGrid file

When the scipt terminates, all mistakes and their summaries are listed and saved in a log file as an error message.

Some types of errors can be automatically corrected after the detection, and another script finds and displays errors step by step in the Praat editor window.

5 Annotating nonverbal and multimodal behavior

There are many different approaches for annotating nonverbal human behaviour. The *structural transcription* deals with recognizing the boundaries of gestures, or a gesture sequence, and the segmentation of gestures by changing the direction of movements or dynamism. *Descriptive transcription* considers body parts and its joint points using degrees of freedom (Martell 2002). It is a fairly objective way of describing gestures, and similar to MPEG-4 Body Animation Parameters (Koenen 2002), which is often used for virtual human animation. The *functional transcription* supports the analysis of gesture meanings. It is the complex annotation schema, which codes gesture type, form, and meaning (Steininger et al. 2002).

One of the goals of video annotation is to develop machine learning algorithms to classify facial expression, hand gesture, etc. Another goal is to test or measure the qualities of different algorithms, but our principal aim is to make the recordings adequate for analyzing multimodal human behaviour. Therefore our annotation technique is of a functional type. We also deal with descriptive transcription but instead of manual annotation, we develop image processing methods for annotation form of gesture and pose automatically.

5.1 Annotation tool

HuComTech uses its own annotation tool, which follows the hierarchical annotation model. The levels, groups, events, their attributes, and some other features of annotation can be described with its .xml schema. During the annotation process, the user can only choose one

Kinga Pápay, Szilvia Szeghalmy & István Szekrényes: HuComTech Multimodal Corpus Annotation
Argumentum 7 (2011), 330-347
Debreceni Egyetemi Kiadó

of the predefined labels for the event, or an empty label (see Figure 7). Thus annotators can edit and also delete labels.

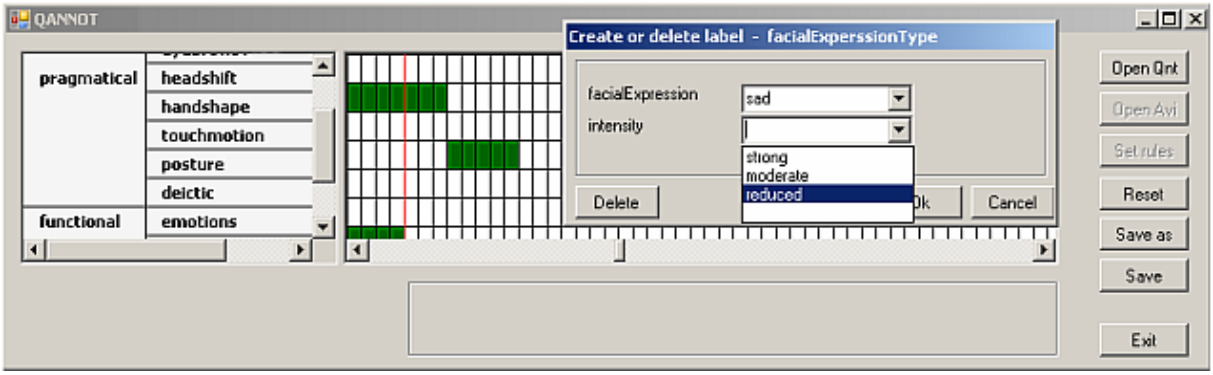

Figure 7: Video annotataion tool in use: a "sad" facial expression event is just created

5.2 Annotation scheme

In this part, the annotation scheme of nonverbal behavior is described, which is similar to MUMIN (Allwood et al. 2006) and is extended with the HuComTech communication model. The model defines three levels of communication: physical, functional and basic levels. The basic level contains basic rules of conversation. The physical level includes different forms of gestures; torso, head, shoulders, hand movement, etc. The highest level is the functional level that includes interpreted mimics and gestures, thus it is the most subjective part of the annotation process.

Using the hierarchical annotation model, the gestures belonging to the same function-class can be determined from the corpus. For example if the interviewee has doubt about something, they may shrug their shoulder or scowl their eyebrows, lean their head or do something else. On the other hand, the function-classes are independent from each other, so a sub-corpus can easily be created, which contains only a selected part of the video corpus; for instance, all the frames where the facial expression group was labelled as happy. This feature of functional classes is useful for developing machine learning methods using this corpus.

The video annotation scheme (see Table 3) also contains the two main discourse markers (start and end) of the communication events.

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 3: Video annotation scheme

5.3 Annotation protocol

The annotation process is quite subjective, therefore it is very important to design strict annotation protocols. Requested quality can be ensured by the common use of an exact labelling system, rules and validation method.

The first problem during the annotation process is to find the boundary of the gestures, which is pretty hard due to the transition of gestures. Some gestures are isolated, so the hand is in a rest postiton at the time of the beginning of the gesture, and arrives back there. But there is not a pause between the gestures many times. The change of the vehicle or trajectory of movement shows only, that the new gesture begins. Therefore it was neccessary to decide how to manage these transition parts. Another problem is classification. A lot of gesture belongs to more than one functional classes, the same function can be expressed in several ways. If the gesture is ambiguous, it is possible that different annotators assign different labels to it. Since it is important that the annotated corpus could be used for machine learning methods, the annotator attaches a label to the frame only if they are is shure which label belongs to it. It is one of the main rules of the annotation process.

It also has to be guaranteed, that watching or listening to an event are not influencing each other's annotation, hence the video and audio parts on the physical level and on the basic level are annotated independently. The events of the functional level, where both channels are considered, are annotated in the last turn.

Finally, during the verification or validation process of nonverbal signal labelling, the annotator checks that the right label category has been choosed and properly timestamped.

6 Conclusion and future prospects

The annotation process of the HuComTech corpus was described in this paper. It is a multimodal annotated formal and informal dialogue corpus. The novelty that the HuComTech corpus combines various modalities: audio, visual or nonverbal only, and complex audiovisual. Our team hopes that we can provide motivation for further research into the development and analysis of other Hungarian multimodal corpora as well. Segmentation of communicative events will involve the simultaneous observance of verbal and nonverbal cues as well. Our future work will center around the evaluation and synthesis of the results of our team's subprojects, namely, HCI communication modelling, digital image processing, and general linguistic subprojects, in order to be able to fully comprehend, and then also model the inherently multimodal nature of communication.

After analysing the speech material, the next step is setting up prosodic rules, statistical modelling and their implementation to the HTK speech recognizer (Young et al. 2005). In

case of statistical modelling, the suprasegmental feature vectors extracted from the preprocessing of the speech files and the segmentation and labeling data can be used for building prosodic models. Besides the purposes of speech and image recognition, after the annotation of the audio and the visual, nonverbal channel, we can examine and implement multimodal features and their overlaps as well.

References

Augmented Multi-party Interaction – AMI (2004) project:
website: http://www.amiproject.org (last downloaded: 2 May 2011)

Allwood, J., Cerrato, L., Dybkjaer, L., Jokinen, K., Navaretta, C. & Paggio, P. (2005): The MUMIN Multimodal Coding Scheme. *NorFA Yearbook,* 129-157.

Beckman, M., Hirschberg, J., Pierrehumbert, J., Pitrelli, J., Price, P., Silverman, K. & Ostendorf, M. (1992): TOBI: A standard scheme for labeling prosody. *Proceedings of the International Conference on Spoken Language.* (last downloaded: 8 May 2010)

Boersma, P. & Weenink, D. (2007): *Praat: Doing Phonetics by Computer 5.0.02*. http://www.praat.org (last downloaded: 9 January 2011)

Burkhardt, F., Paeschke, A. et al. (2005): A database of German Emotional Speech. *Proceedings of Interspeech*, 1517-1520.

Douglas-Cowie, E., Campbell, N., Cowie, R. & Roach, P. (2003): Emotional Speech: Towards a New Generation of Databases. *Speech Communication* 40, 33-60.

EmoTV (2005) corpus description web site:
http://www.limsi.fr/RS2005/chm/tlp/tlp3/index.html (last downloaded: 2 May 2011)

Hirschberg, J. (2007): Pragmatics and Intonation. In: Horn, L.R. & Ward, G. (eds.): *The Handbook of Pragmatics*. Oxford: Blackwell Publishing.

Hunyadi, L. (2006): Grouping, the Cognitive Basis of Recursion in Language. *Argumentum* 2, 67-114.

Hunyadi, L. (2009): Experimental Evidence for Recursion in Prosody. In: Benjamins, J., Diken, T. ten & Vago, R. (eds.): *Approaches to Hungarian* 11, 119-141.

Hunyadi, L. (2010): Cognitive Grouping and Recursion in Prosody. In: van der Hulst, H. (ed.): *Recursion and Human Language*. Berlin & New York: Mouton de Guyter.

Keszler, B. (1989): Die grammatischen und satzphonetischen Eigenschaften der Parenthesen. In: Szende, T. (ed.): *Proceedings of the Speech Research '89 International Conference*, *June 1-3, Budapest, Magyar Fonetikai Füzetek* 21. Budapest: MTA Nyelvtudományi Intézet, 355-358.

Koenen, R. (2002): *MPEG-4 Overview*. http://mpeg.chiariglione.org/standards/mpeg-4/mpeg-4.htm (last downloaded: 3 May 2010)

Markó, A. (2005): *A spontán beszéd néhány szupraszegmentális jellegzetessége*. PhD thesis, Budapest: ELTE BTK.

Martell, C. (2002): FORM: An Extensible, Kinematically-Based Gesture Annotation Scheme. *Proceedings of the International Conference on Spoken Language*, 353-356.

Németh, Zs., Szaszák, Gy. & Vicsi, K. (2007): Prozódiai infromáció használata az automatikus beszédfelismerésben; mondatmodalitás felismerése. In: Alexin, Z. & Csendes, D. (eds.): *V. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: Szegedi Egyetemi Kiadó, 69-80.

Pápay, K. (2009): A spontán beszéd prozódiai frázisszerkezetének modellezése és felhasználása a beszédfelismerésben. In: Tanács, A., Szauter, D. & Vincze, V. (eds.): *VI. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: Szegedi Egyetemi Kiadó, 373-375.

Pápay, K. (2010): Kísérleti módszerek a beszéd technológiai célú kutatásában. In: Gecső, T. (ed.): *Segédkönyvek a nyelvészet tanulmányozásához*. Budapest: Tinta Könyvkiadó, 232-237.

Seppänen, T., Toivanen, J. & Väyrynen, E. (2004): Automatic Discrimination of Emotion from Spoken Finnish. *Language and Speech* 47 (4), 383-412.

Steininger, S., Lindemann, B. & Paetzold, T. (2002): *Labeling of Gestures in SmartKom – The Coding System*. Berlin: Springer, 1611-3349.

SmartKom (2003) project website: http://www.smartkom.org (last downloaded: 2 May 2011).

Szaszák, Gy. (2009): *A szupraszegmentális jellemzők szerepe és felhasználása a gépi beszédfelismerésben*. PhD thesis. Budapest: BME TMIT.

TALK (2007) project website: http://www.talk-project.org (last downloaded: 2 May 2011)

Tóth, Sz. L., Sztahó, D. & Vicsi, K. (2007): Speech Emotion Perception by Human and Machine. In: *Proceedings of COST Action 2102 International Conference. Revised Papers in Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*. Springer LNCS, 213-224.

Váradi, T. (1998): Manual of the Budapest Sociolingistic Interview Data. *Working Papers in Hungarian Sociolinguistics* 4. Budapest: Linguistics Institute, Hungarian Academy of Sciences. http://www.nytud.hu/buszi/wp4/index.html (last downloaded: 2 May 2011)

Vicsi, K. & Szaszák, Gy. (2008): Using Prosody for the Improvement of ASR: Sentence Modality Recognition. In: *Proceedings of Interspeech*. Brisbane, Australia: ISCA Archive,. http://www.isca-speech.org/archive (last downloaded: 8 May 2010)

Vicsi, K. & Szaszák, Gy. (2010): Using Prosody to Improve Automatic Speech Recognition. *Speech Communication* 52, 413-426.

Vicsi K. & Sztahó D. (2009): Ügyfél érzelmi állapotának detektálása telefonos ügyfélszolgálati dialógusban. In: Tanács, A., Szauter, D. & Vincze, V. (szerk.): *VI. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: Szegedi Egyetemi Kiadó, 217-225.

Varga, L. (1999–2001): The Unit of the Hungarian Intonation. In: Szathmári, I. (ed.): *Annales Universitatis Scientiarum Budapestinensis de Rolando Eötvös nominatae. Sectio Linguistica tomus* XXIV. Budapest: ELTE Eötvös Kiadó, 5-13.

Varga, L. (2002): *Intonation and Stress. Evidence from Hungarian*. Houndmills, Basingstoke: Palgrave Macmillan.

Wilting, J., Kramber, E. & Swerts, M. (2006): Real vs. Acted emotional speech. *Proceedings of Interspeech*, 805-808.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D. et al. (2005): *The HTK Book* (for version 3.3). Cambridge: Cambridge University.

Kinga Pápay
University of Debrecen
Department of General and Applied Linguistics
Pf. 24
H-4010 Debrecen
kinga.papay@gmail.com

István Szekrényes
University of Debrecen
Department of General and Applied Linguistics
Pf. 24
H-4010 Debrecen
xepenator@gmail.com

Szilvia Szeghalmy
University of Debrecen
Faculty of Informatics
Pf. 12
H-4010 Debrecen
szeghalmy.szilvia@inf.unideb.hu