Alexandra Staudt & Kinga Pápay:
The Annotation of the HuComTech Audio corpus in Practice – Observations and Questions Arising
Argumentum 7 (2011), 313-329
Debreceni Egyetemi Kiadó

Working paper

Alexandra Staudt & Kinga Pápay

The Annotation of the HuComTech Audio Corpus in Practice – Observations and Questions Arising[1]

Abstract

The planning and creation of the HuComTech Multimodal Corpus is the latest sub-project of the HuComTech (Human-Computer Interaction Technologies) Research Team of Debrecen, Hungary. The sub-project is part of a project conducting research in multimodal human-machine communication. Currently the annotation of the corpus is underway, the questions arising during the annotation process are important from a linguistic point of view as well. Marking and differentiating the units of spontaneous speech and marking of disfluencies typical of spontaneous speech as well as structural and discourse characteristics have required to discuss the following problems: problems of segment boundary placement, differentiating elements with a pause-filling function, differentiating embeddings and insertions, differentiating iteration and restarting, usage of label hesitation and label backchannel in ambiguous cases. The present paper analyzes these questions and suggests solutions by describing examples from the corpus.
Keywords: spontaneous speech corpus, annotation, prosody research

1 Introduction

The planning and creation of the HuComTech Multimodal Corpus is the latest subproject of the HuComTech (Human-Computer Interaction Technologies) Research Team of Debrecen, Hungary. The main goal of the project is studying the overlap between verbal and nonverbal communication occurrences, as well as using the observations for increasing the naturalness and efficiency of human-machine communication. Within this framework, our research goal is the study of the intonation structure of spontaneous speech and its visual manifestation with special attention to the following phenomena: embedding, insertion, iterations (Hunyadi 2009a-b), hesitation, restarting, backchannel, intonational realization of discourse functions of main and subclauses (turn-taking, turn-giving, turn-keeping), and, in addition, the study of the intonational variations in all these categories depending on emotional states and formal/ informal discourse situations. The observations based on these studies will allow for the studied intonational features of spontaneous speech to be included in speech recognition systems, increasing their efficiency. The measurements, statistics and studies necessary were

[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).

conducted using the annotated corpus. This requires a sufficiently annotated multimodal corpus of adequate size and speaker distribution. The multilevel, multimodal annotation provides an opportunity to study overlaps between various audio annotation levels and modalities, as well as verbal and nonverbal communication, according to the research goals of the project. The audiovisual corpus features speakers primarily from northeast Hungary. Currently, the corpus includes samples from 121 young adults – formal and informal spontaneous speech as guided dialogues – with appropriate ratios of gender. The annotation of the corpus is currently under way, forming a crucial part of the data processing: the labels are placed on the audio material (as well as the transcript) on different levels – those of intonational phrases, emotional/cognitive states, and of discourse. The label groups are classified on the basis of prosody. The questions arising during the annotation process are important from a linguistic point of view as well – marking and differentiating the units of spontaneous speech is not as straightforward as it is in case of written texts, and it poses problems. Marking the disfluencies typical of spontaneous speech (hesitation, restarting, repetition) as well as structural characteristics (inserted, embedded, broken clauses) have required the development of a new annotation system. The present paper analyzes the applicability of these new rules to the occurring phenomena by describing the examples from the corpus.

2 The labels used for audio annotation

The audio annotation includes the segmentation of the audio recordings into clauses which are then labeled and transcribed. The Praat software (Boersma & Weenink 2007) is used for the annotation process, identifying five annotation levels: three of these are functional while two of them are transcriptional. The functional level includes three sublevels: those of intonational phrases, emotional phrases and discourse phrases; the transcriptional levels consist of the word-by-word transcripts. The current paper discusses the difficulties that are posed by the segmentation and labeling processes. The functional labels used during the audio annotation are included in Table 1.c

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: The levels and labels of audio annotation

3 Problems posed by audio annotation

In the current paper we describe and provide examples for the problems that so far have come up during the annotation process in the IP and discourse levels, in addition to the difficulties posed by the segmentation process.

3.1 Segmentation

During the segmentation of the corpus we have relied on research previously conducted on speech technology with similar goals to ours (Szaszák 2009, Vicsi-Sztahó 2009); therefore, we have decided to break up the samples into clauses, in the traditional sense. At the same time, however, spontaneous speech creates difficulties for this type of segmentation as the sentence structures in spontaneous speech are not always clear – partly due to the disfluencies. That's why it is worth differentiating between sentences in the classical sense and their corresponding realizations in spontaneous speech (Gósy [2003], for example, called the latter "virtual sentences"). While annotators rely on commas and conjunctions in breaking sentences down into clauses, it is also important to keep intonational units intact. These pose the following problems:

Problem 1: During sentence segmentation, maintaining the intonational units is difficult because certain conjunctions belong to the intonation pattern of the following clause, while others join the previous clause. This situation usually co-occurs with coarticulation, which makes segmentation even more difficult. The question, therefore, is whether the clause boundary falls before or after the conjunction.

Solution to Problem 1: Relying on the annotators' grammatical knowledge and keeping speed and efficiency in mind, we recommend consistently following the original rule that the segment boundary should fall between the comma and the conjunction. Future research on the already annotated material should reveal how often a conjunction belongs to the previous or the following clause in spontaneous speech, which clause it more often forms an intonational phrase with, and whether this phenomenon is speaker-dependent.

Conjunctions are function verbs joining clauses and they express both subordination and coordination in traditional grammar (Keszler 2001). This definition is also poses problems:

Problem 2: Certain conjunctions and adverbs often occur as expletives in spontaneous speech, hindering the recognition of clause boundaries.

Alexandra Staudt & Kinga Pápay:
The Annotation of the HuComTech Audio corpus in Practice – Observations and Questions Arising
Argumentum 7 (2011), 313-329
Debreceni Egyetemi Kiadó

Using expletives may result from the speaker's uncertainty (Horváth 2009), and, at the same time, in such cases participants of the discourse gain more time for processing the received information, and the speaker has more time to put his or her ideas into words. Adverbs are also likely to aggregate within the sentence as well (see example 1).

(1) Example:
Interviewer: *"és akkor nem ment el, hanem még így ott ült, még így tovább így sokáig."*
"and then he didn't go but he kept sitting there on, and on and on for a long time."

Solution to Problem 2: The annotator's attention has to be called to the fact that not every conjunction is a real conjunction, and not all adverbs are real adverbs; it is important that they recognize when these are used with an expletive function. In Table 2, we have collected conjunctions, adverbs and pronouns with an expletive function that have occurred in the examined transcripts.

Hungarian conjunctions	IPA	In English
akkor	[ɒk:oɾ]	then
aztán	[ɒsta:n]	
szóval	[so:vɒl]	so
úgyhogy	[u:ʝhoʝ]	
tehát (és elharapott alakjai)	[tɛha:t]	therefore (and its shortened versions)
meg	[mɛg]	and
Hungarian expletives	**IPA**	**In English**
így – úgy	[i:ʝ – u:ʝ]	this way – that way
Hungarian pronoun	**IPA**	**In English**
ilyen – olyan	[ijɛn – ojɒn]	this, that, such

Table 2: Conjunctions and adverbs with expletive function in the HuComTech spontaneous speech corpus

The expletive word *úgyhogy (so),* included in Table 2, is also mentioned by Markó (2005b). This word most often occurs at the end of sentences in the HuComTech spontaneous speech corpus (see Example 2). The word *tehát (therefore)* may also behave similarly. The next annotation problem is related to this phenomenon:

Problem 3: In case of *úgyhogy (so)* and *tehát (therefore)* positioned at the end of the sentence, the question arises during annotation of whether the segment should be considered an unfinished sentence with a conjunction or a completed sentence with an expletive. In spontaneous speech, unfinished sentences occur frequently and these require a special labeling in annotation (see Chapter 2).

Unfinished sentences are sentences with a beginning but whose train of thought has been left unfinished for some reason by the speaker. One of these reasons may be that the speaker cannot remember the appropriate word, cannot continue the sentence with the original logical structure and starts from the beginning again, does not see the point in continuing since a new idea has come to his or her mind, or simply because the discourse partner has interrupted the speaker. In these cases, the annotator is not aided by the intonational features of the utterance as the fundamental frequency is rarely falling at the end of sentences in spontaneous speech (see Markó 2005b).

Solution to Problem 3: Since this pattern has occurred quite frequently, we have decided these cases should be considered finished sentences with expletives (rather than unfinished clauses).

Example (2):
Interviewee: *"angolul meg ezek mindenhol beszélnek, úgyhogy."* (Töltelékszó szerep.)
"these speak English everywhere, so." (Expletive function)
Interviewer: *" %igen, igen, igen. úgyhogy az biztos, hogy jó."* (Kötőszó szerep.)
"yes, yes, yes, so that must be good." (Conjuction function)

It seems to be speaker-dependent which conjunctions are used as expletives or even whether a person uses any of them as expletives. Some samples of the corpus come from Hungarians living in the neighbouring countries. Being bilingual, these speakers use Hungarian less frequently than speakers living in Hungary. Their speech samples contain fewer hesitations of this sort – a fact that may be the result of a more controlled language use.

3.2 Observations concerning the IP level

By our own definition, the intonational phrase or IP is a segmental unit that can be defined by an intonation pattern i.e. tonal pattern. From a technological point of view, a prosodic unit in Hungarian can be a word or phrase based on stress, or a sentence or clause identified on the basis of intonation (Szaszák 2009). Technologically, word boundary detection is considered a simple case of syntactic analysis. However, study of the stressed segments could belong to the field of semantics. For the description of these units, intonation models on word-, phrase-, clause- and sentence-levels were used. The goal of the theoretical part of the research is to break up spontaneous speech into prosodic phrases (IPs), and the identification of their operational markers and prosodic boundaries. We aim to find regularities in the acoustic-suprasegmental features of spontaneous speech and to compare these to the features of reading aloud. In order to make research more manageable, the level of intonation phrases was introduced, allowing for the identification of head clauses (HC) and subordinate clauses (SC) in speech. These are the various combinations of syntactic primitives which can be used for further operations by the speaker. Operations conducted with the IPs (and which can be identified prosodically) include embedding and insertion; operations within the IPs can be iteration, restarting and hesitation. Difficulties in the annotation of the IP level are explained below. The images used as visual aids were created by a Praat program. The images include an oscillogram on top, a fundamental frequency curve underneath, and the four- and five-level annotations at the bottom.

Alexandra Staudt & Kinga Pápay:
The Annotation of the HuComTech Audio corpus in Practice – Observations and Questions Arising
Argumentum 7 (2011), 313-329
Debreceni Egyetemi Kiadó

3.2.1 Embedding and insertion in spontaneous speech

Problem 4: On the level of intonational phrases, differentiating between embedding (EM) and insertion (IN) often poses a problem for annotators.

Embedding refers to the process when a subordinate clause is inserted into the head clause – or in the case of multiple embeddings, into the subordinate clause (see Figure 1). Syntactically, embedding is linked to the first part of the head clause right before the embedded phrase. This occurs, for example, with the sentence *The cat, that was bit by the dog that was rabid, ran away.* Hunyadi (2006, 2009a-b) illustrates that the deeper the embedding the lower the tone's frequency starts. Embedding is a syntactic operation, its result is integrated into the hierarchical structure of the sentence.

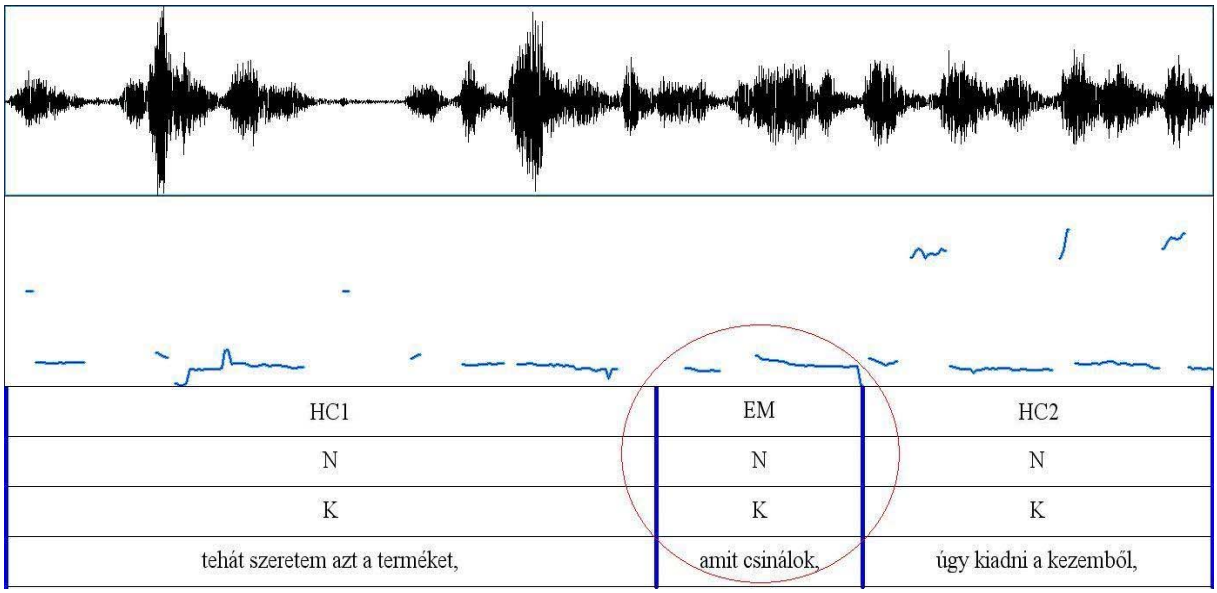

Figure 1: Example for embedding from the HuComTech corpus

The bookmark effect is the prosodic representation of syntactic discontinuity, which means that the tonal contours of the syntactic segments get connected, and thus they can be considered two parts of an IP. The prosodic feature of embedding, i.e. the deeper tone and the bookmark effect prevails on the tonal continuity of the neighboring phrases. Figure 1 depicts the fact that in spontaneous speech, the embedded clause is not necessarily indicated by deeper tones while tonal continuity is still present. (A more detailed statistical analysis will be carried out on the completed annotated corpus.) In spontaneous speech, re-embedding, the return to the tonal features of the IP preceding the embedding, may not happen. One reason for that could be that the speaker has forgotten what was said before the embedding or insertion; that is, the operation is discontinued. However, it is also possible that the speaker changes his or her mind and continues the sentence with a different structure or that the speaker repeats some segment of the discontinued clause to make comprehension easier. The speaker may even continue with referring to the parts before the embedding employing an anaphora (see Example 3).

Alexandra Staudt & Kinga Pápay:
The Annotation of the HuComTech Audio corpus in Practice – Observations and Questions Arising
Argumentum 7 (2011), 313-329
Debreceni Egyetemi Kiadó

Example (3):

Interviewee: *"%o a Szinapszis ~K ~F ~T -nél, ez egy piackutató cég (IN), ennél dolgoztam, mint telefonos %o munkatárs."*
"… at the Szinapszis LTD, this is a market research company (IN), I worked for this as a customer service representative."

On the other hand, insertion is a special type of a head clause: it is a new/independent clause inserted into a head or a subordinate clause (Hunyadi 2006). For example, the sentence *"Meg tudnád mondani, hogy – az én órám megállt – hány óra van?" 'Could you tell me – my watch has stopped – what time it is?'* is a typical case for insertion, where the speaker inserts a comment into his or her idea of the main clause. Insertion is a syntactic operation, its result is not integrated into the hierarchical structure of the sentence. According to Hunyadi (2006), the tonal continuity of the surrounding phrases is also apparent in this case. At the same time, Keszler (1989) believes that the prosodic feature of the inserted clause could also be the flat intonation, the deeper tone, the faster tempo of speech, as well as pausing before and after the parenthetical remark. (Keszler has conducted research in the suprasegmental realizations of insertions as features of sentence structures.) An example for this is from our corpus (Figure 2). We also think that another type of insertion could be the call-out, when we exclaim or say something to someone else while speaking – this could be indicated by higher F0. The examples listed so far support Keszler's observations about texts read aloud: spontaneous speech with insertions is also characterized by a deeper and more monotonous tone. Statistical analysis of the corpus will also be conducted regarding this claim. Hunyadi (2006) states that, prosodically and from a computational aspect, insertion is not different from embedding; the difference is merely syntactic. Besides tonal continuity and deeper tone, another feature of insertions may be a faster tempo and smaller modulation. It remains to be seen whether these are also true for embedding as well.

Solution to Problem 4: The two phenomena can be distinguished by several cues: in the case of insertion, a syntactically independent clause is inserted into the head clause; while in the case of embedding, a subordinate clause is inserted into the head clause or another subordinate clause. Their common features are that they both cut/break the sentence or the clause, and the tonal continuity of the broken phrase (Hunyadi 2006) is present in both cases. Based on our observations, intonational features may differ from those of the surrounding segments more significantly in the case of insertions than in embeddings. Therefore, this distinction supports the recognition of insertions (since in some cases the intonation of the inserted segment is more monotonous than that of the embedded one).

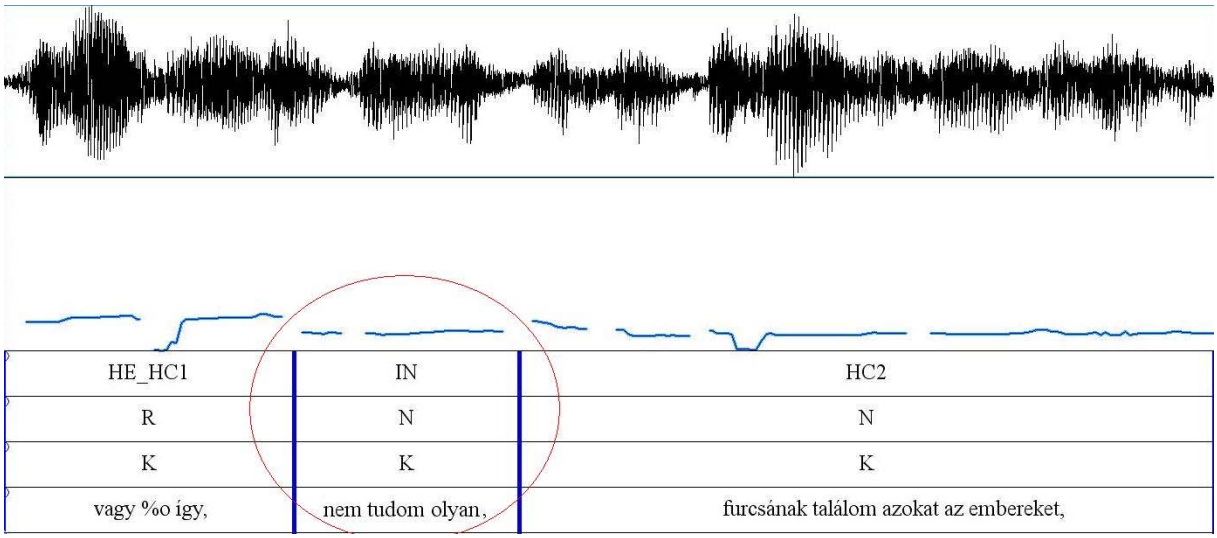

Figure 2: Example for insertion from the HuComTech corpus

3.2.2 Disfluencies: restarting, hesitation

The process of annotation is greatly hindered by the disfluencies so typical of spontaneous speech. In casual speech, we often rephrase sentences that we have started, and repeat words or even whole phrases. This phenomenon may be due to the fact that the speaker changes the linguistic form of the idea to be expressed in midsentence, recognizes mistakes by self-monitoring and alters his or her speech planning strategy, or the speaker may simply become uncertain about how to continue the train of thought. The most frequent phenomena are hesitation and restarting.

Problem 5: During the annotation process, differentiating between hesitation and restarting may pose problems in certain cases.

In case of restarting, the curves of the two segments are more or less identical – this emphasizes the repetition, so the intonational pattern of the previous segment (that could be a clause, a phrase or even a single word) is repeated (see Figure 3). During the annotation process, the following types of restarting have been found so far:

Types of restarting:

1. Restarting incomplete words:
   "úgyse tudnak még rá rea-- <reagálni> sem."
   "they aren't able to rea-- <react> to it yet."
   "nagy kerte-- <nagy kertes> házba lakunk."
   "we live in a big house with a gar-- <garden>."
   "inkább csak né-- vagyis hát <inkább> nézem."
   "I would prefer to just wa-- <rather> watch."
2. Repeating whole words (mostly adverbs, pronouns, articles, conjunctions):
   "ami ilyen <ilyen> nagyon rossz lett volna."
   "that would have been so <so> very bad."

"én <én> nagyon régen hallottam vicceket."
"I <I> haven't heard jokes in a long time."
"mert %o {b} <mert> akkor még nem tudtam,"
"because <because> I didn't know about it at the time,"
*"hogy <hogy> *közbe eszedbe jut,"*
"that <that> you will remember it in the meantime,"
"d%e <de> utána az úgy megszakadt,"
"but <but> it just broke afterwards,"
"hogy %o <hogy> meglenne az %a %s hát ilyen lehetőség,"
"that <that> such opportunity would be there,"

3. Repeating phrases:
"most úgy <most úgy> nem jut eszembe semmi,"
"right now <right now> I can't think of anything,"
"és ő is <és ő> haza <is> ment."
"and he also <and he did> go home."
4. Restarting with a new structure:
"mint hogyha egy teljesen más -- <mint hogyha> mondjuk ő meghalt volna."
"as if a completely diff-- <as if>, say, he had died."

There can be two fundamental reasons for restarting. One of these is that the speaker meant to say something different than what he or she actually did. This case also has two variations: one of them is correcting a slip of the tongue, the other one is a complete recomposition. The other reason for restarting is gaining time by repeating clauses, phrases or words. This latter type occurs mostly with function words. Based on observations by Gyarmathy et. al. (2009), the interrupted utterance is followed by an unfilled pause in nearly half of the cases. However, Horváth (2009) indicates that in many instances the speech process is not interrupted from an articulation aspect, there is no lack of signal, a filled pause signals that the speech production is to be continued.

Solution to Problem 5: These repetitions could fall into the category of hesitation (Horváth 2009). During the annotation of the corpus, the two types were distinguished based on intonational features: hesitation (HE) label was used only for stretching, which is characterized by a longer, monotonous fundamental frequency curve. Restarting, on the other hand, is indicated by the above mentioned repeated intonational pattern. Therefore, in the annotations, hesitation is lengthening: %o [ø:]; %m [m:]; hesitation also refers to any sound lengthening within a word, indicated by a % sign before the lengthened sound. Hesitation is prosodically not independent considering the tonal continuity; thus it is included in the clause segments during segmentation.

Alexandra Staudt & Kinga Pápay:
The Annotation of the HuComTech Audio corpus in Practice – Observations and Questions Arising
Argumentum 7 (2011), 313-329
Debreceni Egyetemi Kiadó

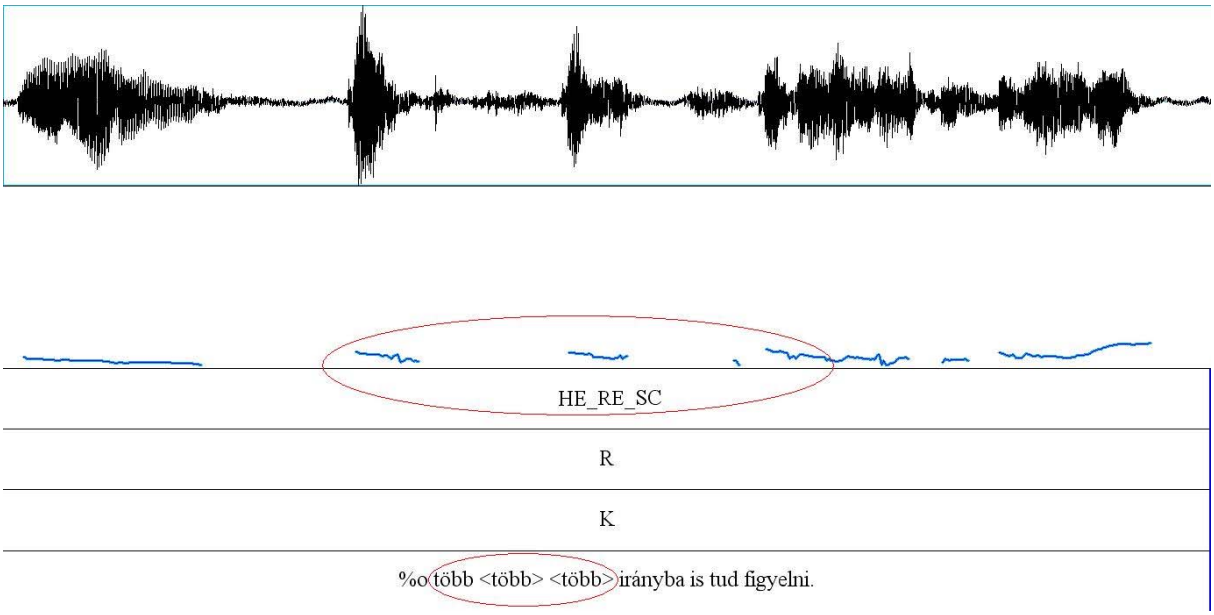

Figure 3: Example for hesitation and restarting from the HuComTech corpus

3.2.3 Iteration: enumeration, reinforcement

Problem 6: Figure 3 depicts an example that represents the problem of differentiating between iteration and restarting.

Iteration includes enumeration (example from the corpus: *"táncoltunk, pörögtünk, tűzzsonglőrködtünk"* [we danced, whirled, juggled with fire], (see also Figure 4) as well as repetition for reinforcement (for example, *"igen, igen"* [yes, yes]). The distinction between these two is necessary because we suppose that the intonation of iteration is repeated with a regular melody in both casual speech and in reading a written text aloud. Through this feature, iterations could be recognized and utilized within speech technology. However, the pattern of regularly repeated intonation is often interrupted by conjunctions, for example, *"angolt, németet, meg olaszt"* (English, German and Italian); therefore, we consider enumerations including a conjunction (*"meg olaszt"* [and Italian]) a separate segment. Identifying enumerations as iterations does not pose a problem for annotators, but identifying repetitions with a reinforcing function does. It is not uncommon for speakers to repeat words because of uncertainty or to gain more time but these cases, as previously indicated, are identified as restarting.

Solution to Problem 6: In this case, what annotators have to consider is whether the repetition is for reinforcement or if it is due to uncertainty – listening to the context is crucial for making a decision in this situation as analyzing the text level alone may lead to incorrect identification. Distinguishing between the two is important because repetition for enforcement is realized with stress, i.e. has different prosody.

Alexandra Staudt & Kinga Pápay:
The Annotation of the HuComTech Audio corpus in Practice – Observations and Questions Arising
Argumentum 7 (2011), 313-329
Debreceni Egyetemi Kiadó

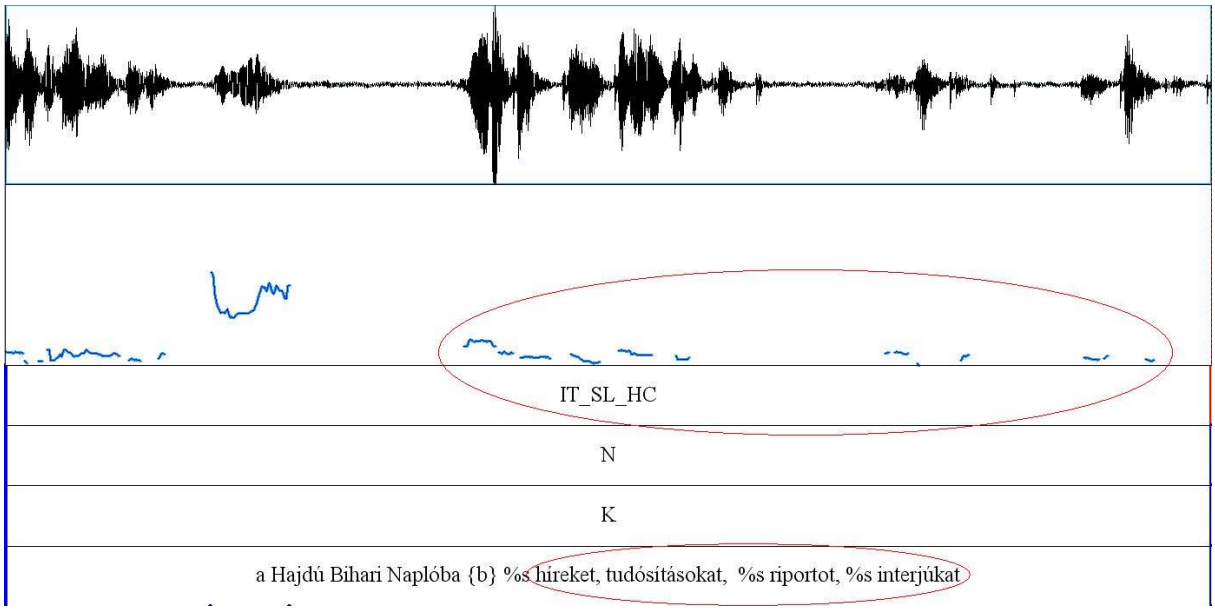

Figure 4: Example for iteration from the HuComTech corpus

3.3 Discourse level labels: turn-taking or backchannel?

We included the labels of discourse turn-taking in a separate level in order to be able to observe regularities in turn-taking as well as how they impact the IP level. After starting the audio annotation we realized that the use of labels such as "turn-take"(T), "turn-give" (G) and "turn-keep"(K) was not sufficient on the discourse level since in spontaneous speech the discourse partner often reacts to what has been said – see Figure 5. With the interviewee's speech, the responses were labeled B on level 3 and BC on level 1. For the annotation process this was a problem as these utterances often overlapped with the speech of another speaker.

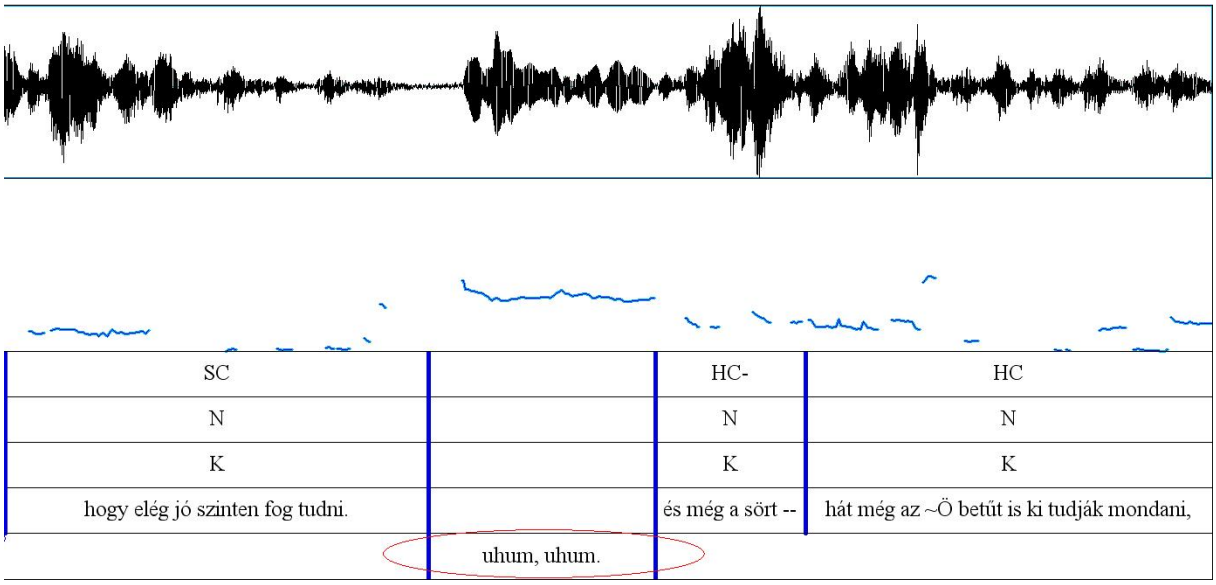

Figure 5: Example for backchannel from the HuComTech corpus

Since the responses are sometimes not words found in dictionaries, the method of their transcription had to be agreed upon, and the annotators must follow it consistently. It is clear from the relevant literature that these types of responses are present in each culture and language, although the frequency of their use may vary. The realization of backchannels is different in every language due to the phoneme set specific to each language. The international literature includes a great number of studies on backchannels in casual speech, especially for English and Japanese (Ward 2000, Pipek 2007), but there is research about German (Stocksmeier 2007), Chinese and Korean (Young & Lee 2004) as well. Backchannels appear not only in face-to-face spontaneous dialogues but are also present in the chat services on the Internet (e.g. MSN, ICQ, etc.) and in telephone conversations. Since they are put into writing in chats, their spelling is mostly created, shaped and spread by this channel. Their role in writing is even greater that in speech as the discourse participants do not see their partners' reactions; therefore, they are necessary for maintaining attention and indicating presence online.

Clarifying the definition is important. Markó (2005a), referring to various earlier studies, such as Vértes O. (1987), uses the term humming for the vocal phenomena we use to reassure our discourse partner that we are following what he or she is saying, and we might even present various emotional reactions simultaneously during a discussion. Backchannels are also typical for spontaneous discourse as the listener's reactions to the speaker's utterances, which can be verbal, such as approval or humming, and can also be nonverbal, like nodding or frowning. The purpose of backchannels is not to provide information but to indicate the listener's attention and interest as well as encouraging the speaker to go on. Backchannels are mostly realized as single words or humming but they might also be sentences of a few words (see Example 4). It must be kept in mind, however, that not all humming is backchanneling, and not all backchanneling is humming, as backchannels may be dictionary words. The hummings discussed by Markó do not always appear as merely concomitant phenomena, as they may constitute complete answers as well.

Potential classification of backchannels:

1. 1. backchannels of the humming sort (e.g. *uh-huh, hmm*)
2. approval expressed by single words or phrases (e.g. *oh yes, of course, sure*)
3. "echo-sentences" (see Figure 6): repeating the speaker's phrases – we believe that this category is different from that of one-word sentences, and it functions like other backchannels.

Example (4):

Interviewee: *"csak feküdni kellett, meg pihenni otthon."*
"I just had to stay in bed and rest at home."
Interviewer: *"uhum."* (B).
"mhm."
Interviewee: *"hát meg egy sokkoló %s volt az egész."*
"well, the whole thing was pretty shocking"
Interviewer: *"igen.*(B) *hát ez sok.* (B) *igen."*(B)
"Yes. (B) It's too much. (B) Yes. (B)"

Alexandra Staudt & Kinga Pápay:
The Annotation of the HuComTech Audio corpus in Practice – Observations and Questions Arising
Argumentum 7 (2011), 313-329
Debreceni Egyetemi Kiadó

Problem 7: It happens occasionally that the listener corrects the speaker's utterances or he or she says the word(s) that is/are slipping the speaker's mind at the moment. The question is whether these are actually backchannels or rather they function more as turn-taking/HC devices.

Solution to Problem 7: Completions including new information are not considered backchannels but completions with a pure approving purpose are (see Example 5).

Example (5):
Interviewer: *"de {b} hát náluk a ház a%z tele van ((így)) ezekkel a* (K)"
"but {b} our house i%t is full of ((thus)) these" (K)
Interviewee: *"trófeával. {l}"*(G)
"trophy. {l}" (G)]
Interviewer: *"trófeákkal. {l}"* (K)
"trophies. {l}" (K)]

Interviewee: *h%át %o végül is* (K)
"w%ell% actually (K)"
Interviewer: *kevésbé.*(G)
"less." (G)
Interviewee: *kevésbé.* (K)
"less." (K)

The listener expresses approval and reinforcement with the backchannels, which may also carry emotional content. The backchannels included in the corpus are mostly emotionally neutral but certain backchannels, such as *"uh!", "hmm!", "hah!"* are specifically used for expressing emotions like surprise and amazement; other emotionally charged backchannels like *"really?"and "hmm?"* express great interest. Backchannels may take several forms even though their meanings are similar. The most frequently used backchannels in the corpus are: *uh-huh, uh, sure, hmm, right (right, right), yes (yes, yes), I see* and laughter {l} – their confirmation and refinement are to be carried out by statistical analysis.

Alexandra Staudt & Kinga Pápay:
The Annotation of the HuComTech Audio corpus in Practice – Observations and Questions Arising
Argumentum 7 (2011), 313-329
Debreceni Egyetemi Kiadó

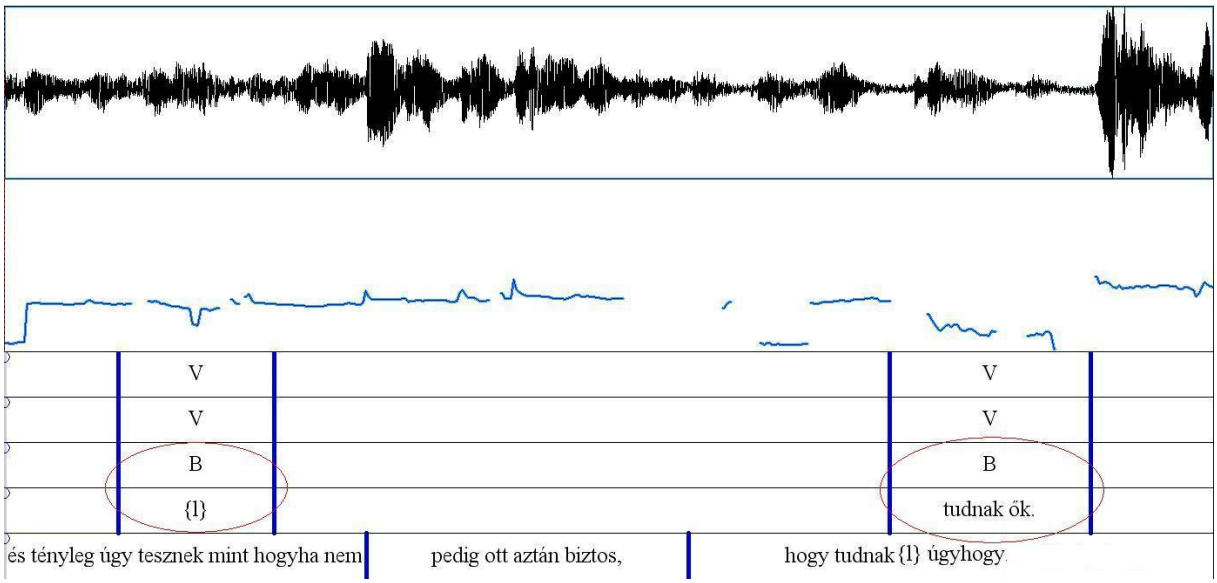

Figure 6: Example for laughter and "echo"-backchannels from the HuComTech corpus

Turn-taking happens when the speaker's signals clearly indicate that he or she is ready to stop or start talking. These signals can be a pause or intonation but this assumption has not been confirmed by research. Therefore, annotators have to rely on the recordings and their own experiences in recognizing the speaker's intentions, which are not always obvious.

Certainly when differentiating between turn-taking and backchannels, one fact to be kept in mind is that turn-taking usually includes new information and beginning of a new discourse, while backchannels maintain the current turn rather than starting a new one. In some cases the listener may use a backchannel but, at the same time, takes over the turn by continuing to speak. These situations are relatively easy to identify for the annotator as the classic backchannel is classified as a separate backchannel segment and the following segment receives the turn-take/turn-give label (see Example 6).

Example (6):

Interviewee:	*hogy jól legyenek me-- melegedve az izmok, mert* – (K)
	"so that the muscles are well wa-- warmed up because (K)"
Interviewer:	*uhum.* (B) *%s tehát azt nem javaslod, hogy csak úgy elkezdjen nyújtani az ember?* (G)
	"mhm. (B) so you don't recommend starting stretching just like that?" (G)
Interviewee:	*az %o nem annyira hatékony.* (T)
	"it isn't so effective." (T)

Problem 8: The backchannels pose another problem, though. The feedbacks/responses are sometimes difficult to set apart from the head or sub clauses of "virtual sentences". Annotators are often confused when one of the speakers answers the other's question by backchannel (see Example 7). It is obvious that the speaker has answered the question, but he or she did so using a frequently used backchannel instead of a verbal approval (e.g. *yes, sure*).

Example (7):
Interviewer: *"ja, koleszos vagy?"*
"oh, do you live in the dorm?"
Interviewee: *"uhum."*
"mhm."

Solution to Problem 8: In these cases, annotators have to watch for turn-giving in a question form before the backchannel.

4 Conclusions

In the present paper the difficulties of annotation were discussed and analyzed from various aspects. The different kinds of overlaps observed among the studied labels have been illustrated, along with the difficulties that these phenomena pose for annotators. It has also been demonstrated that backchannels can be realized by humming and even sentences, and vice versa, answers to questions may also take the form of backchannel humming. Novelties, such as restarting and repetition, which would not occur while reading a text aloud, were also discussed. We may make the following statements about the above mentioned phenomena:

1. Segmentation is done on the clause level; on the text level, the segment border always falls between the comma and the conjunction.
2. Annotators need to recognize elements/components with a pause-filling function.
3. The words *úgyhogy* [so], *tehát* [therefore] positioned at the ends of sentences indicates the end of a sentence – if the speaker's intentions indicate it as well.
4. Annotators need to keep in mind that the difference between embedding and insertion is that in case of an insertion, the speaker inserts a syntactically independent clause into the utterance; an embedded segment is syntactically connected to the clause or sentence in which it is embedded, usually with conjunctions. Their common feature is that both of these operations interrupt the sentence that includes their results, but while an embedding's result is integrated into the hierarchical structure of the sentence, an insertion's result is not.
5. The hesitation label (HE) is only used in case of stretching.
6. When differentiating between iteration and restarting, it has to be considered whether the repetition is for reinforcement (iteration- IT label) or the result of uncertainty (restarting – RE label).
7. Complementation with new information is not included in the backchannel category but the complements used purely for approval are.
8. When one of the speakers answers a question with a backchannel, the answer is considered turn-take/turn-give.

These suggestions and observations are not clear in the current annotation manual and the above mentioned phenomena create uncertainty for annotators. Following the existing and the above discussed annotation rules consistently is necessary for the future analysis and utilization of the corpus.

References

Boersma, P. & Weenink, D. (2007): *Praat: doing phonetics by computer* 5.0.02. http://www.praat.org (10.05.2011).

Gósy M. (2003): Virtuális mondatok a spontán beszédben. In: Gósy M. (szerk.): *Beszédkutatás 2003*. Budapest: MTA Nyelvtudományi Intézet, 19-44.

Gyarmathy D., Gósy M. & Horváth V. (2009): A rejtett és a felszíni önmonitorozás temporális jellemzői. In: Keszler B. & Tátrai Sz. (szerk.): *Diskurzus a grammatikában – grammatika a diskurzusban.* Budapest: Tinta Kiadó, 46-55.

Horváth V. (2009): *Funkció és kivitelezés a megakadásjelenségekben.* PhD értekezés. Budapest: ELTE.

Hunyadi, L. (2006): Grouping, the cognitive basis of recursion in language. *Argumentum* 2, 67-114.

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

Hunyadi, L. (2009b): Cognitive grouping and recursion in prosody. In: van der Hulst, Harry (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.

Keszler B. (szerk.) (2001): *Magyar Grammatika*. Budapest: Nemzeti Tankönyvkiadó.

Markó A. (2005a): "Szavak nélkül". Nonverbális vokális közlések fonetikai elemzése. *Magyar Nyelvőr* 129:(1), 88-104.

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

Pipek, V. (2007): *On Backchannels in English Conversation.* PhD Thesis. Brno: Masaryk University.

Stocksmeier, T., Kopp, S. & Gibbon, D. (2007): Synthesis of prosodic attitudinal variants in German backchannel "ja". *Proc. of Interspeech 2007*. Antwerpen.

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

Vicsi K. & Sztahó D. (2009): Ügyfél érzelmi állapotának detektálása telefonos ügyfélszolgálati dialógusban. *VI. Magyar Számítógépes Nyelvészeti konferencia*. Szeged, 217-225.

Ward, N. & Tsukahara, W. (2000): Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics* 32, 1177-1207.

Young, R.F. & Lee, J. (2004): Identifying units in interaction: Reactive tokens in Korean and English conversations. J*ournal of Sociolinguistics* 8/3, 380-407.

Vértes O. A. (1987): Bevezetés a magyar hangstilisztikába. *Nyelvtudományi Értekezések* 124. Budapest: Akadémiai Kiadó.

Alexandra Staudt
University of Debrecen
Department of General and Applied Linguistics
Pf. 24
H-4010 Debrecen
staudt.alexandra@arts.unideb.hu

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