Human-human, human-machine communication: on the HuComTech multimodal corpus

L. Hunyadi
Department of General and Applied Linguistics, University of Debrecen, Debrecen, Hungary
hunyadi@undieb.hu

T. Váradi
MTA Institute of Linguistics, Research Group on Language Techonology Budapest, Hungary
varadi.tamas@nytud.mta.hu

Gy. Kovács
MTA SzTE Research Group on Artificial Intelligence, Szeged, Hungary, Embedded Internet Systems Lab, Luleå University of Technology, Luleå, Sweden
gykovacs@inf.u-szeged.hu

I. Szekrényes
Institute of Philosophy, University of Debrecen, Hungary
szekrenyes.istvan@arts.unideb.hu

H. Kiss
Department of General and Applied Linguistics, University of Debrecen, Debrecen, Hungary
kiss.hermina@arts.unideb.hu

K. Takács
Department of Phonetics, Eötvös Loránd University, Budapest, Hungary
karolin3813@gmail.com

Abstract

The present paper describes HuComTech, a multimodal corpus featuring over 50 hours of video taped interviews with 112 informants. The interviews were carried out in a lab equipped with multiple cameras and microphones able to record posture, hand gestures, facial expressions, gaze etc. as well as the acoustic and linguistic features of what was said. As a result of large-scale manual and semi-automatic annotation, the HuComTech corpus offers a rich dataset on 47 annotation levels. The paper presents the objectives, the workflow, the annotation work, focusing on two aspects in particular i.e. time alignment made with the Leipzig tool WEBMaus and the automatic detection of intonation contours developed by the HuComTech team. Early exploitation of the corpus included analysis of hidden patterns with the use of sophisticated multivariate analysis of temporal relations within the data points. The HuComTech corpus is one of the flagship language resources available through the HunCLARIN repository.

1 Introduction

In the age of the ubiquitous smart phones and other smart devices, robots and personal assistants, the issue of human-machine communication has acquired a new relevance and urgency. However, before communication with machine systems can become anything approaching the naturalness and robustness that humans expect, we must first understand human-human communication in its complexity. In order to rise to this challenge, we must break with the word-centric tradition of the study of communication and we must capture human-human communication in all the richness of the settings that it normally takes place. The foremost requirement for such an enterprise is richly annotated data, which is truly in

Laszlo Hunyadi, Tamás Váradi, István Szekrényes, György Kovács, Hermina Kiss and Karolina Takács 2019. Human-human, human-machine communication: on the HuComTech multimodal corpus. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 56–65.

short supply given the extremely labour intensive nature of the manifold annotation required. The ambition of the HuComTech project, which goes back to 2009, is to provide a rich language resource that can equally fuel application development as well as digital humanities research.

The HuComTech corpus is the first corpus of Hungarian dialogues that, based on multiple layers of annotation offers the so far most comprehensive information about general and individual properties of verbal and nonverbal communication. It aims at contributing to the discovery of patterns of behaviour characteristic of different settings, and at implementing these patterns in human-machine communication.

The paper will be structured as follows. In section 2 we will describe the data (the informants, the settings of the interviews, the size and main characteristics of the data set etc.) and will discuss the annotation principles and will provide a brief overview of the various levels of annotation. Section 3 discusses two automatic methods used in the annotation: forced alignment at the word level using the WEBMaus tool available through Clarin-DE as well as the automatic identification of intonation contours. Section 4 will preview some tentative exploration of the data, describing an approach that is designed to reveal hidden patterns in this complex data set through a sophisticated statistical analysis of the temporal distance between data points.

2 Description of the data and its annotation

2.1 General description of the corpus

The data for the HuComTech corpus was collected in face-to-face interviews that were conducted in a lab. The informants were university student volunteers. During the interviews informants were asked to read out 15 sentences, and were engaged in both formal and informal conversations, including a simulated job interview. The corpus consists of 112 interviews running to 50 hours of video recording containing about 450 000 tokens. Both the verbal and non-verbal aspects of the communication between field worker and informants were recorded through suitably positioned video cameras and external microphones.

The corpus offers a huge amount of time aligned data for the study of verbal and non-verbal behaviour by giving the chance to identify temporal patterns of behaviour both within and across subjects. The native format is .eaf to be used in ELAN (Wittenburg et al 2006), but a format for Theme (Magnusson, 2000), a statistical tool specifically designed for the discovery of hidden patterns of behaviour is also available for a more advanced approach of data analysis.

Through a database the data of the corpus will be made completely available for linguists, communication specialists, psychologists, language technologists.

A non-final version of the HuComTech corpus is already available online and it can be explored using *Trova* and *Annex (*Beck & Russel 2006) tools developed by the Max Planck Institute for Psycholinguistics within the framework of *The Language Archive* project. From there one can also download media and annotation files for academic research purposes.

2.2 The annotation protocol and the annotation scheme

The annotation followed the independent tagging for each of the more than 30 levels. It means that each level was annotated without any information about tags entered on another level. Each level of each file was annotated by two annotators independently, and a third annotator made possible corrections. The annotators formed groups in which they regularly discussed emerging issues, too. It assured a satisfactory inter-annotator agreement.

The annotation, comprised of about 1.5 million pieces of data ranges from the description of nonverbal, physical characteristics of 112 speakers (gaze, head-, hand-, body movements) to the pragmatic, functional description of these characteristics (such as turn management, cooperation, emotions etc.) The annotation of verbal behaviour includes the phonetics of speech (speech melody, intensity, tempo), morphology and syntax. The more than 450000 running words are time aligned enabling the association of the text with non-verbal features even on the word level.

A special feature of the annotation is that, whenever applicable, it was done both multimodally (using signals both from audio and video channels) and unimodally (using signals from either channel). Of course we subscribe to the view that both the production and the perception of a communicative event is inherently multimodal, yet the rationale for separating the two modalities was that the analysis and the generation of such an event by a machine agent needs to set the parameters of each of the modalities separately. Apart from this technical implementational perspective, we believe that by annotating some communicative/pragmatic functions both multimodally (using information from both video and the audio channel) and unimodally (relying on information from either the video or the audio) may pinpoint the primary source of the given function as either a single modality or a complex of several ones.

Accordingly, the annotation layers are organized into the following six annotation schemes in terms of the modalities involved: audio, morpho-syntactic, video, unimodal pragmatic, multimodal pragmatic and prosodic annotation.

The *audio annotation* is based on the audio signal using intonation phrases (head and subordination clauses) as segmentation units (Pápay et al 2011). The annotation covered verbal and non-verbal acoustic signals and included the following elements: transcription, fluency, intonation phrases, iteration, embeddings, emotions, turn management and discourse structure. The annotation was done manually using the Praat tool (Boersma &Weenink, 2016), validation was semi-automatic involving Praat scripts.

The *morpho-syntactic* annotation was done both manually and automatically, covering different aspects. Automatic annotation included tokenization, part of speech tagging and parsing (both constituent and dependency structure) The HMM-based toolkit *magyarlánc* (Zsibrita et al, 2013) developed at Szeged University was used for the automatic morpho-syntactic annotation. In addition, syntax is also annotated manually both for broader linguistic and for specific non-linguistic (especially psychology and communication) purposes (focusing on broader hierarchical relations and the identification of missing elements).

Video annotation included the following annotation elements: facial expression, gaze, eyebrows, head shift, hand shape, touch motion, posture, deixis, emblem, emotions. Annotation was done manually and, where possible, automatically using Qannot tool (Pápay et al, 2011) specially developed for the purpose.

Unimodal pragmatic annotation used a modified (single-modal) version of conversational analysis as its theoretical model and with the Qannot tool manually annotated the following elements: turn management, attention, agreement, deixis and information structure.

Multimodal pragmatic annotation used a modified (multimodal) version of Speech Act Theory and using both verbal and visual signals covered the following annotation elements: communicative acts, supporting acts, thematic control, information structure. The annotation was done manually with the Qannot tool.

Prosodic annotation (see Section 3 below) was prepared automatically using the Praat tool and covered the following elements: pitch, intensity, pauses and speech rate.

As the above detailed description of the annotation schemes reflects, a large part of the annotation was done manually. This was inevitable given the fact that the identification of

perceived emotions as well as a large number of communicative as well as pragmatic functions require interpretation, which are currently beyond the scope of automatic recognition, therefore they have to be determined and annotated manually.

2.3 Automatic annotation of prosody

In this section we describe a method developed for the automatic annotation of intonation, which, however, can be used not just for the HuComTech corpus, and therefore, we feel, deserves discussion in some detail. Our method does not follow the syllable-size units of Merten's Prosogram tool (Mertens, 2004) but an event can integrate a sequence of syllables in larger trends of modulation, which are classified in terms of dynamic, speaker-dependent thresholds (instead of *glissando)*. The algorithm was implemented as a Praat script. It requires no training material, only a two-level annotation of speaker change is assumed.

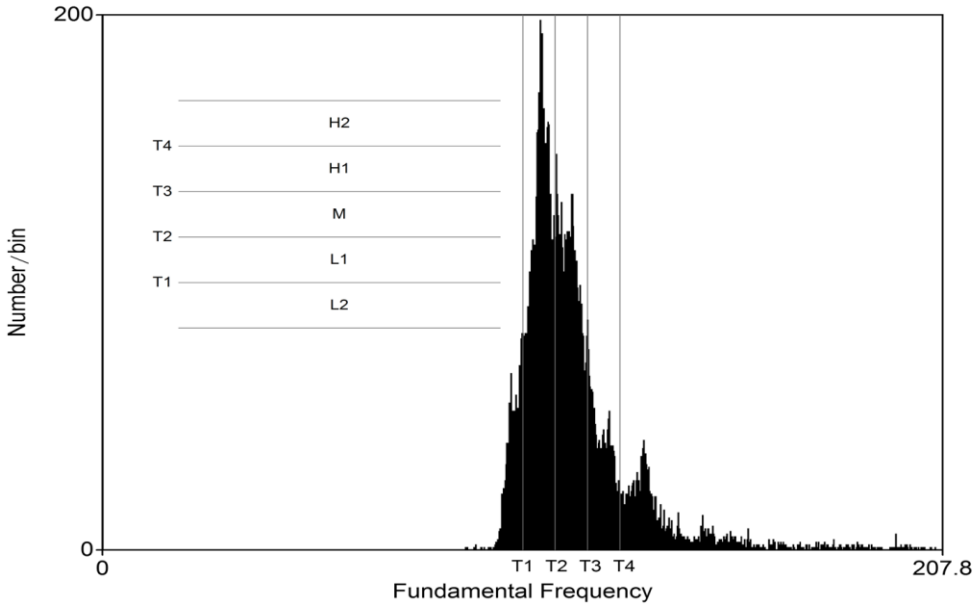

Figure 1: Calculating individual pitch ranges of the speaker based on the F0 distribution: L2 < L1 < M < H1 < H2.

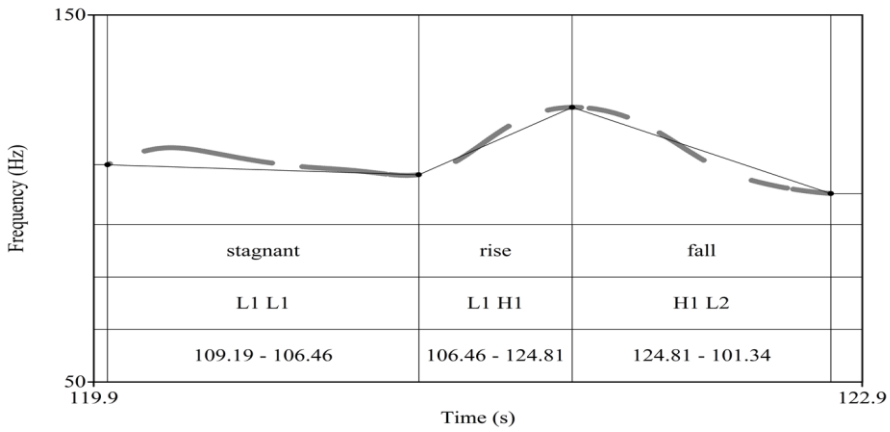

Figure 2: An output sample from the HuComTech Corpus

The output of the algorithm (Szekrényes 2015) contains larger, smoothed and stylized movements of the original data (F0 and intensity values) where the values indicate the contour (descending, falling, rising etc.), the absolute and relative vertical position of every single prosodic event through their starting and ending points. The "relative position" means that we located breaking points of intonation trends in the individual vocal range of the speaker which was divided into five levels based on the distribution of F0 values (see Figure 1).

The resulting labels representing modulations and positions of the prosodic structure can be considered as an automatically generated but perceptually verifiable music sheet of communication based on the raw F0 and intensity data. In Figure 2, one can see an output sample from the HuComTech corpus.

3 Exploring the corpus

We report two preliminary explorations of the HuComTech corpus. Experiments have been conducted with a view to modelling turn management through machine learning using neural networks. Second, through the use of a sophisticated statistical analysis tool we sought to explore hidden patterns within the complex multimodal data sets on the basis of temporal distance between them.

3.1 Modelling turn management: automatic detection of turn taking

The HuComTech corpus provides detailed data on turn management. For each discourse unit it contains manual annotation to indicate topic initiation, topic elaboration, topic change (and the absence of these categories, we will refer to as "no contribution"). Such comprehensive annotation invites experimentation for machine learning to automatically model turn management. Indeed, it is very important for a machine agent to be able to establish if the human interlocutor is keeping to the topic at hand or when they are veering away from it either by opening a completely different topic or slightly altering the course of the conversation. Conversely, it is also important for the machine agent to know when the human interlocutor is not doing any of the above (i.e. not contributing to the topic). Given that – depending on the situation (i.e. the machine agent speaking or not) – this can mean that the human is merely providing backchannel feedback (thus the agent does not have to relinquish the speaking turn) or that the human has finished its turn and the agent can take the speaking turn instead without the risk of interrupting or speaking over the human interlocutor.

Earlier studies on topic structure discovery relied mostly on lexical information (Holz and Teresniak, 2010), prosody (Zellers and Post, 2009), or a combination of the two (Shriberg et al., 2000). The HuComTech corpus, on the other hand, contains a more extensive annotation, facilitating the use of much wider sources of information as cues. This, among others, include such cues as gaze, facial expression, hand gestures, head movements, and so on. Despite the abundance of available information, however, the task is still challenging, and the experiments so far represent only tentative initial steps. One particular difficulty of topic classification is the class imbalance inherent to the task. In a conversation, one naturally spends more time with either not contributing to the topic (by providing only backchannel feedback, or not speaking at all) or elaborating the current topic than with changing the topic (either only slightly altering the course of the conversation, or completely veering away from the current topic).

The effect of this imbalance is twofold, as it can affect both the training and the evaluation of our models. For evaluation purposes the most common metric applied in classification tasks (particularly in multi-class classification) is the accuracy:

$$Accuracy = \frac{\sum_{i \in Classes} \quad No.of correctlyidentifiedinstances \text{in} Classi}{\sum_{i \in Classes} \quad No.of allinstances \text{in} Classi}$$

It is easy to see, however, how class imbalance can introduce a bias into this metric, favouring models that perform well on the majority class. In the HuComTech database, for example, the two majority classes (topic elaboration and no contribution) make up 82% of all instances to be classified, which means that a model correctly classifying these instances (but none of the instances from the two minority classes - topic change and topic initiation) can attain an accuracy score of 82%. An accuracy score of 82% may seem like a reasonable performance, but a model that cannot identify the change in topic at all is clearly ill-equipped for the task of topic unit classification. For this reason we suggested the use of a different metric for evaluating topic unit classification models, namely the Unweighted Average Recall (UAR):

$$UAR = \sum_{i \in Classes} \frac{No.of correctlyidentifiedinstances \text{in} Classi}{No.of allinstances \text{in} Classi \cdot No.of classes}$$

The benefit of using UAR (for further information, see Rosenberg, 2012) is that it assigns the same importance to each class, regardless of the cardinality of the classes. This means that a model performing well on the majority classes but bad on the minority classes would receive the same score as a model performing well on the minority classes, but bad on the majority classes.

Another problem class imbalance can cause is a bias towards the majority classes in the model trained. By training a neural network using several examples from certain classes, and relatively few examples of others, we may inadvertently train the network to disregard the minority classes. One technique that can be used to avoid this is that of downsampling, where we use the same amount of samples from each class during the training process. This, however, means that we disregard a large portion of our labeled samples. Another possibility would be to collect more samples from the minority classes. This, however, is both time-consuming, and costly, rendering this option infeasible in most applications. But it is also possible to "fool" the model by using the samples from the minority classes several times during the process of training. One technique that enables us to do so is that of probabilistic sampling (for further information, see Tóth and Kocsor, 2005), where a parameter is used to control the uniformity of class distribution during the training process. When value of the parameter is set to 0, the number of samples used from each class is equal to the cardinality of that class; when the value of the parameter is set to 1, however, the same number of samples are used of each class.

Kovács et al. (2016) built a topic unit classifier with the use of Deep Rectifier Neural Nets (Glorot et al, 2011), applying the technique of probabilistic sampling. We demonstrated in several experiments that this method attains a convincingly better performance than a support vector machine or a deep rectifier neural net by itself. For further information see (Kovács et al. 2016). In our tentative experiments we have found that the same holds true for other neural networks architectures – such as Long Short-Term Memory Unit (LSTM - Hochreiter and Schmidhuber, 1997) networks, and Gated Recurrent Unit (Cho et al., 2014) networks – as well. Our preliminary results show that the application of probabilistic sampling significantly increases the UAR scores attained in both of these models as well.

Given the rich annotation available for the dialogues in the HuComTech corpus, another promising direction of inquiry is to examine the rate of contribution different types of features had towards the identification of the correct topical unit label. For this we used five of the six categories of annotation described in Section 2.2 (morpho-syntactic annotation,

video annotation, unimodal annotation, multimodal annotation, prosodic annotation). First, we examined the performance attainable with Deep Neural Networks when using features from only one annotation category. We found that by using the features from multimodal annotation only (with the exception of the topic unit labels, which were used as targets), an UAR score can be attained on the task of topic unit classification that is competitive with those scores we attain when using all features. What is more, in most cases we got a better UAR score by using only the multimodal features than that we got by using all available features. In the next stage we employed a classifier combination method on the models trained on individual feature categories. Here, we took the weighted average of the posterior probability estimates provided by the five different models. We found using the proper combination of our five models, we can further improve the classification performance. What is more, we also found that we can attain the same performance using only two categories, that is multimodal and morpho-syntactic annotation. For further information, see (Kovács et al. 2017)

3.2 T-pattern analysis to discover hidden patterns of behaviour

Undoubtedly, the HuComTech corpus contains a bewildering number and complexity of annotation data. The possibility to use this rich database to explore possible interdependencies between data points recorded at numerous levels of annotation is an exciting prospect as well as a serious challenge.

The difficulty lies not simply in the number of data points to consider but rather, it is of a theoretical nature. The capturing of a given communicative function cannot usually be done by describing the temporal alignment of a number of predefined modalities and their exact linear sequences, since for the expression of most of the functions a given list of participating modalities includes optionalities for individual variation, and sequences are not necessarily based on strict adjacency relations. As a result, traditional statistical methods (including time series analysis) are practically not capable of capturing the behavioural patterns leading to functional interpretation.

We apply an approach of discovery on multivariate analysis of temporal relationships between any annotation elements within a given time window. T-pattern analysis (Magnusson, 2000) was developed for the discovery of hidden patterns of behaviour in social interactions using the software tool *Theme*. *Theme* is a unique software environment that is intended to override the usual challenges of behavioural research, namely, patterns are composed of events which do not necessarily follow one another in an immediate sequence, and also, these events may be optional in many cases. Accordingly, when searching for patterns for a given communicative function, this function needs to be identified even if its constituents are not adjacent or, in certain cases, an event stereotypical for the given function is not even present at all. The authors of this paper had the chance to participate in the further development of this software by exposing it to the very large and annotationally complex HuComTech corpus.

The T-Pattern analysis offers a framework to meet these serious challenges by simulating the cognitive process of human pattern recognition. The result is a set of patterns as possible expressions of a given function with their exact statistical significance. Moreover, it also suggests which of the constituting elements (events) of a given pattern can predict or retrodict the given function as a whole.

Without exceeding the limits of this paper let us have a few examples of results from Hunyadi (2017) showing how *Theme* can capture the above challenges and dynamic of multimodal communications, based on the HuComTech corpus:

Example 1: f055 (formal dialogue, female subject), ID975:

example for a pattern composed of events from syntax and gaze direction

(([1 (end of incomplete clause, end of coordination)][2(end of gaze forward, start of blink)])([3 (end of blink, start of blink)][4 (gaze down, end of blink)]))

Annotated as:
(([1 (miss,e,yes co,e,yes)][2(v_gaze,e,forwards v_gaze,b,blink)])([3 (v_gaze,e,blink v_gaze,b,blink)][4 (v_gaze,b,down v_gaze,e,blink)]))

Text:
1: *és úgy gondolom [and I think]*,T=A_speaker_text,B=41188,E=42643
2.3.1.0.0.0.4,6,9.,T=S_clauses,B=41188,E=42643

2: forwards,T=V_gazeClass,B=42655,E=43455
blink,T=V_gazeClass,B=43455,E=43855

3: blink,T=V_gazeClass,B=43455,E=43855
blink,T=V_gazeClass,B=44255,E=44655

4: down,T=V_gazeClass,B=44655,E=45055
blink,T=V_gazeClass,B=45055,E=45455

Example 2: f055 (formal dialogue, female subject), ID 508:

pattern of multimodal behaviour

([1 (speaker, end of new information, spekaer, end of topic elaboration)]([2 (agent, beginning of directive, agent, beginning of new information)][3 (agent, end of directive, agent, end of new information)]))

Annotated as:
([1 (mp_spinf,e,new mp_sptopic,e,t_elab)]([2 (mp_agcommact,b,directive mp_aginf,b,new)][3 (mp_agcommact,e,directive mp_aginf,e,new)]))

Preceeded by:

{b} %m *rendben* [all right].,T=S_text,B=23719,E=25109
mivel pályakezdő vagyok {since I am starting my career],,T=S_text,B=25109,E=26429
nem volt még előző {p} *munkahelyem* [I have not had a previous workplace],,T=S_text,B=26429,E=28569
%s *a tanulmányaim eléggé* %s %o %s *jól sikerültek* [I was fairly successful with my studies].,T=S_text,B=28569,E=33748
tehát úgy érdemjegyileg [as for marks] %o %s *mindenben* [and everything] %s {l} *meg vagyok elégedve vele* [I am satisfied with it].,T=S_text,B=34207,E=41188
és úgy gondolom [and I think],,T=S_text,B=41188,E=42643
hogy [that] e— %o %*ezt tudnám kamatoztatni a munkámban is* [I could benefit from it in my work].,T=S_text,B=42643,E=46574)

1: new,T=MP_speaker_Information,B=23695,E=45775
topic_elaboration,T=MP_speaker_Topic,B=23695,E=45775

2: directive,T=MP_agent_CommunicativeAct,B=46720,E=48320
new,T=MP_agent_Information,B=46720,E=48320

és milyen szakot végzett?[and what did you study?],T=A_agent_text,B=46574,E=48228

3: directive,T=MP_agent_CommunicativeAct,B=46720,E=48320
new,T=MP_agent_Information,B=46720,E=48320

Example 3: f060 (formal dialogue beszélgetés, female subject), ID 2030

a pattern of multimodal pragmatics and posture

((([1 agent, beginning of directive] [2 agent, topic initiation])][3 speaker, communicative act, multimodal: none)](([4 speaker, beginning of leaning right speaker, end of leaning right])[5 (speaker, end of leaning right speaker, beginning of leaning right)]))

Annotated as:

((([1 mp_agcommact,b,directive] [2 mp_agtopic,e,t_init])][3 mp_spcommact,b,none)](([4 v_post,b,right,lean v_post,e,right,lean])[5 (v_post,e,left,lean v_post,b,right,lean)]))

1: directive,T=MP_agent_CommunicativeAct,B=52800,E=55360
(the referenced text: {p} {b} %o {p} {b} **mért jelentkezett a felhívásra?* [why did you respond to the call?], T=A_agent_text, B=51358, E=55239)

2: topic_initiation,T=MP_agent_Topic,B=52800,E=53760
(part of the above the referenced text: **mért* [why])

3: none,T=MP_speaker_CommunicativeAct,B=55046,E=56326
(nonverbal backchannel)

4: lean-right,T=V_postureClass,B=57206,E=57606
(the referenced text: *mert* [because] %o *szeretnék munkába lépni* [I want to get the job],, T=S_text, B=56679, E=58614)

5: lean-left,T=V_postureClass,B=57606,E=58006
lean-right,T=V_postureClass,B=58006,E=58806

4 Conclusion

In this short article, we provided a brief overview of the multimodal HuComTech corpus. It is offered as a richly annotated language resource that can serve a number of purposes ranging from supporting application development in the area of human-machine to empirical based research leading to a better understanding of the complex interplay of numerous factors involved in human-human multimodal communication. The corpus is available through the HunCLARIN repository and is made public with the expectation that it will generate further research into multimodal communication.

References

[Boersma &Weenink, 2016] Boersma, D., Paul & Weenink. 2016. Praat : doing phonetics by computer [computer program]. version 6.0.22. http://www.praat.org/. (retrieved 15 November 2016)

[Beck & Russel, 2006] Berck, P. and Russel, A. 2006. ANNEX – a web-based Framework for Exploiting Annotated Media Resources. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*. Genoa: European Language Resources Association, 2006.

[Cho et al., 2014] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259, 2014

[Glorot et al] Glorot, X., Bordes, A., Bengio, Y. 2011. Deep Sparse Rectifier Neural Networks. In: Gordon, G. J., Dunson, D., B. Dudík, M. (eds): *AISTATS JMLR Proceedings 15*. JMLR.org. 315-323.

[Hochreiter and Schmidhuber, 1997] S. Hochreiter and J. Schmidhuber. 1997. Long Short-Term Memory. Neural Comput. 9, 8 (November 1997), 1735-1780.

[Holz and Teresniak, 2010] Holz and Teresnai, 2010: F. Holz and S. Teresniak, "Towards automatic detection and tracking of topic change", in Proc. CICLing, 2010, pp. 327-339

[Hunyadi 2017] Hunyadi, L. 2017. A multimodális kommunikáció grammatikája felé: szekvenciális események rekurzív hierarchikus struktúrája. In: Bánréti, Z. (ed.) Általános Nyelvészeti Tanulmányok XXIX (2017), pp. 155-182.

[Kovacs et al] Kovács, G., Grósz, T., Váradi, T. 2016. Topical unit classification using deep neural nets and probabilistic sampling. In: *Proc. CogInfoCom*, (pp. 199–204)

[Kovács et al. 2017] Kovács, Gy., Váradi, T. 2017. A különböző modalitások hozzájárulásának vizsgálata a témairányítás eseteinek osztályozásához a HuComTech korpuszon, in: Vincze, Veronika (ed.) XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017) Szeged, Szegedi Tudományegyetem Informatikai Tanszékcsoport, (2017) pp. 193-204. , 12 p.

[Magnusson, 2000] Magnusson, M. S. 2000. Discovering hidden time patterns in behavior: T-patterns and their detection behaviour research methods. *Behavior Research Methods, Instruments, & Computers*, 32:93–110.

[Mertens, 2004] Mertens, P. 2004. The prosogram : Semi-automatic transcription of prosody based on a tonal perception model. In *Proceedings of speech prosody*.

[Pápay et al, 2011] Pápay, K., Szeghalmy, S., and Szekrényes, I. 2011. Hucomtech multimodal corpus annotation. Argumentum 7:330–347.

[Rosenberg, 2012] A. Rosenberg, "Classifying skewed data: Importance weighting to optimize average recall" in Proc. Interspeech, 2012, pp. 2242-2245

[Shriberg et al., 2000] E. Shriberg, A. Stolcke. D. Hakkani-Tür, G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics", Speech Commun. Vol 32, no. 1-2, pp 127-154, 2000

[Szekrényes 2014] Szekrényes, I. 2014. Annotation and interpretation of prosodic data in the hucomtech corpus for multimodal user interfaces. *Journal on Multimodal User Interfaces* 8:(2):143–150.

[Tóth and Kocsor, 2005] L. Tóth and A. Kocsor, "Training HMM/ANN" hybrid speech recognizers by probabilistic sampling

Wittenburg et al 2006] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. 2006. Elan : a professional framework for multimodality research. In *Proceedings of LREC 2006* (pp. 213–269)

[Zellers and Post, 2009] M. Zellers, B. Post, "Fundamental frequency and other prosodic cues to topic structure", in Workshop on the Discourse-Prosody Interface, 2009. Pp. 377-386

[Zsibrita et al] Zsibrita, János; Vincze, Veronika; Farkas, Richárd 2013: magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: *Proceedings of RANLP 2013*, pp. 763-771.