Language technology tools and resources for the analysis of multimodal communication

László Hunyadi
University of Debrecen
H-4032, Egyetem tér 1
Debrecen, Hungary
hunyadi@unideb.hu

Tamás Váradi
Research Institute for Linguistics,
Hungarian Academy of Sciences
Budapest, Hungary
varadi.tamas@nytud.mta.hu

István Szekrényes
University of Debrecen
H-4032, Egyetem tér 1
Debrecen, Hungary
szekrenyes@unideb.hu

Abstract

In this paper we describe how the complexity of human communication can be analysed with the help of language technology. We present the HuComTech corpus, a multimodal corpus containing 50 hours of videotaped interviews containing a rich annotation of about 2 million items annotated on 33 levels. The corpus serves as a general resource for a wide range of research addressing natural conversation between humans in their full complexity. It can benefit particularly digital humanities researchers working in the field of pragmatics, conversational analysis and discourse analysis. We will present a number of tools and automated methods that can help such enquiries. In particular, we will highlight the tool Theme, which is designed to uncover hidden temporal patterns (called T-patterns) in human interaction, and will show how it can applied to the study of multimodal communication.

1 Introduction

Following the origins of digital humanities ("literary and linguistic computing"), the text has always been central to it. However, when one gives a closer look at the words these texts are made of, a whole world opens before our eyes: words are just the expression of what can hardly be expressed by words: *human behavior*. The question then arises: can one truly understand and interpret *thoughts*, *reflections*, *intensions* based on words alone? And also: how much of all this can be traced back by following nonverbal events? Do gestures contribute to disambiguating words, or do they rather mask the unspoken context? And in any case: how objectively can we judge and react to a social interaction of competing verbal and nonverbal events?

The objective of the present paper is to show how language technology can help the investigation of such research questions, extending the horizons of Digital Humanities research. The rest of the paper is structured as follows. In section 2, we describe the HuComText Corpus, an extensively annotated corpus of 50 hours of video-recorded interviews, containing 450 000 running words and altogether 2 million annotation items. Section 3 describes automated methods used to facilitate corpus annotation. The corpus was basically annotated manually by trained annotators. There were three aspects of the annotation of the corpus that was automated by language technology tools. Section 3.1 describes how the prosodic annotation describing *pitch, intensity* and *speech rate* was prepared automatically using a language independent tool, *ProsoTool* (Szekrényes 2015). Section 3.2 introduces a web-service for the morphological annotation of the corpus. EmMorph is part of a newly developed open language technology processing chain for Hungarian available at https://e-magyar.hu/parser. Section 3.3 describes our experiences in using the WebMAUS service (Kisler et al. 2016) to prepare the time-alignment of the corpus at the word level.

Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH), pages 117–124, Osaka, Japan, December 11-17 2016.

2 About the HuComTech corpus

2.1 General description

The multimodal HuComTech corpus is a set of approximately 50 hours of video recorded dialogues comprising of 111 formal conversations (simulated job interviews) and 111 informal ones (guided by a standard scheme). The participants were 111 university students (aged 18-29), and the language of the dialogues in both settings was Hungarian. Each interview consisted of 15 sentences read aloud, a 10 minute guided conversation and 15 minutes of free conversation. The whole corpus contains about 450 000 word tokens. The initial aim of building the corpus was to acquire a wide range of data characteristic of human-human interaction in order to make generalisations for their implementation in more advanced human-machine interaction systems (Hunyadi, 2011).

2.2 Annotation principles

The annotation system of the HuComTech corpus involves 33 different levels of annotation including video (labelling both physical attributes such as gaze, head, eyebrows, hand, posture, and their interpretations for emotions, communicative, discourse and pragmatic functions), audio (labelling for the physical attributes of F0, intensity, speech rate, and their interpretations for emotions and communicative and discourse functions), and functional labelling of combined video+audio, following a partly manual, partly automatic scheme of labelling (Abuczki and Esfandiari-Baiat, 2013).

Annotations include video annotations for gaze, head movement, hand movement, posture, facial expressions, audio annotations for transcription, fluency of speech, turn management, emotions, as well as prosody (done automatically) for pitch movement, intensity and pause. The manual annotations are uniquely extended to spoken syntax and, to our knowledge also as first of its kind, to unimodal (video only) pragmatics complementing multimodal pragmatic annotation. All these different layers of annotation are meant to be studied simultaneously, allowing for the study of the eventual temporal and structural alignments of all available multimodal markers.

In addition to its automatic parsing, for syntactic labelling a special manual scheme is designed to code incompleteness of structure, highly specific to spoken language but difficult for machine recognition. Morphological parsing is complemented with the automatic time alignment of each running word in the text. This extensive annotation of 50 hours of dialogues results in about 2 million distinct pieces of data.

As a special feature of the corpus, annotation was done, when applicable, both multimodally (using the video and audio signals at the same time) and unimodally (relying only on the video cues without the sound channel). The rationale behind it was that whereas it is generally accepted that both the production and the perception/interpretation of a communicative event is essentially multimodal due to the participation of a number of (verbal and non-verbal) channels (modalities), both the analysis and generation of such an event by the machine agent needs to follow a complex of individual modalities, i.e. by the setting of the parameters of each of the modalities separately. Due to the highly structured and detailed nature of our material, in what follows we are going to restrict our presentation to data from prosody and morphology only. Even with these restrictions, the presentation will be an example of the multimodal nature of communication: the several levels of prosody and morphology annotation also yield a rich set of data which combine into complex multimodal patterns of elements (events) which contribute to the expression of various communicative functions in the usual multimodal way of optionality.

3 Automated methods of multimodal corpus annotation

3.1 Automatic annotation of prosody

Among manually annotated non-verbal modalities of the interactions, the annotation of prosody was implemented by a computer algorithm in the HuComTech corpus. Therefore not only the resulting prosodic labels, but the methodology itself can be used as a speech processing resource for the prosodic analysis of any spoken language corpora. The algorithm of *Prosotool* was implemented as a Praat script (Boersma and Weenik 2016) for transcribing the temporal modulation of three important prosodic phenomena: *pitch*, *intensity* and *speech rate*. Unlike other existing tools such as ToBI, *Prosotool* is language independent and does not require any training material. Only a pre-created, acoustic rep-

resentation of speaker change is needed, marking utterance units in different annotation tiers per different speakers (see Figure 1.) in Praat TextGrid format.

During automatic pre-processing of recordings, the utterances of different speakers are separated from each other (excluding overlapping segments of speech) in order to make possible the isolated, preliminary analysis of the individual prosodic behavior of every participant, distinguishing four ordinal levels for every prosodic feature based on the individual distribution of the measured, physical

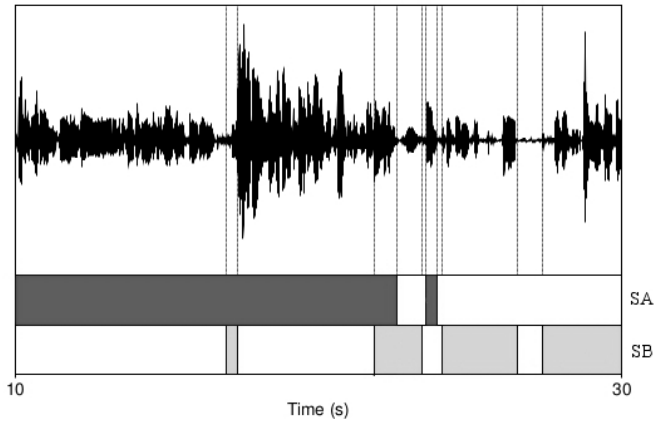

Figure 1. Acoustic representation of speaker change

values (F0, intensity and the actual syllable rates). In Figure 2, these consecutive levels (L1 < T1 < L1 < T2 < M < T3 < H1 T4 < H2) are defined using certain thresholds of F0 distribution.

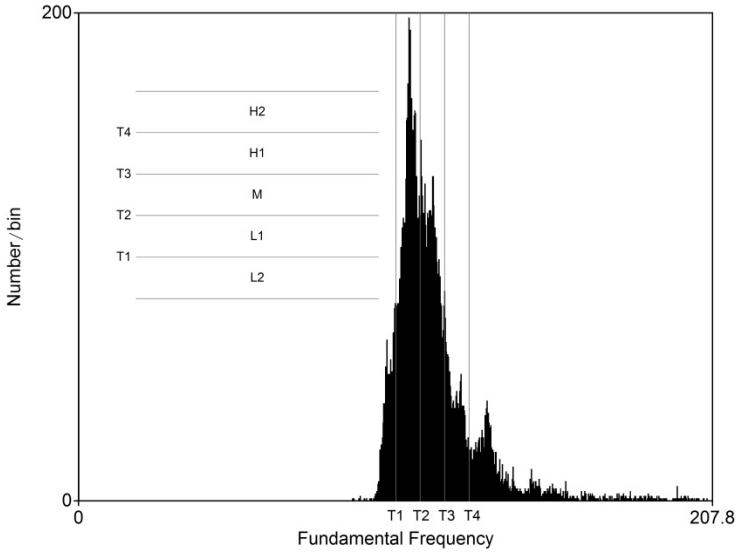

Figure 2. Individual vocal levels of the speaker

In the final output, the resulting annotation labels (in Praat TextGrid format) are not aligned with segmental units of speech (syllables, sentences etc.) but instead they follow the prosodic segmentation of interactions, the stylized segments of F0 and the intensity or syllable rate contour (see Figure 3.). This kind of prosodic segmentation aims at indicating those modulations of prosodic features which are (1) possibly independent from segmental units of speech (for instance, a tonal contour can integrate more syllables or words etc.) and (2) perceptually relevant, significant movements rather than momentary excursions of the measured physical values.

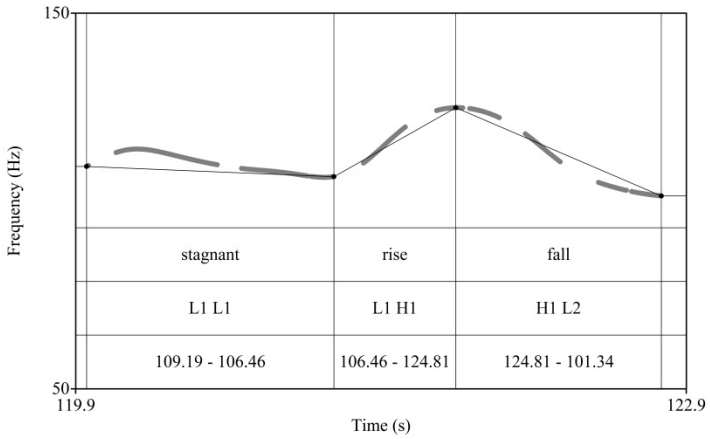

Figure 3. The result of prosodic annotation

As displayed in Figure 3, the resulting annotation consists of a three-level analysis of prosodic features. On the first level, the shape of every prosodic segment is classified using five categories of *rise*, *fall*, *descending*, *ascending* and *stagnant* accents. The classification is based on two main parameters, the duration and the amplitude of the movements. On the third level, the modulations are described as point-to-point vectors of the measured physical values (in Hertz, decibel or Syllable/second), while on the second level, the same vectors are associated with the four-level categories of the relative, individual scale as shown above.

It is intended that the *ProsoTool*'s algorithm will be freely available for research purposes within the framework of *E-magyar* digital language processing system[1]. The concept of stylization was inspired by Merten's *Prosogram* (Alessandro & Mertens 2004) and the psychoacoustic model of tonal perception (Hart 1976). The parameters of the pitch accents classification can also be found in the *Tilt* intonation model (Taylor 2000).

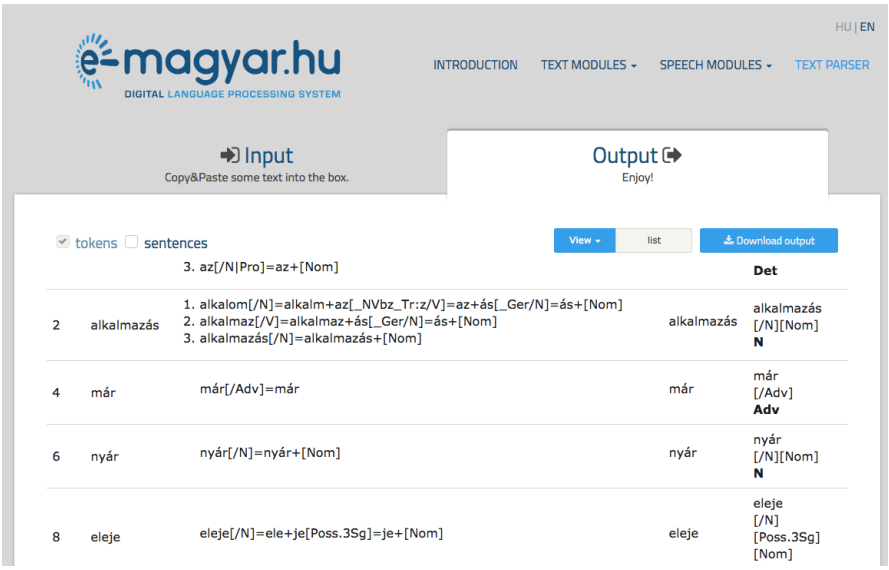

Figure 4. The emMorph web-service output

3.2 Morphological annotation

Morphological annotation of the corpus will be prepared with the emMorph morphological analyser, that is integrated in the recently developed e-magyar.hu language technology infrastructure. The e-magyar toolchain (https://e-magyar.hu) is a comprehensive Hungarian digital processing set prepared

[1] http://e-magyar.hu/

as a collaborative effort of the Hungarian NLP community. It integrates and enhances most of the tools developed in various labs from tokenizer to dependency parser. The flagship product of this new open infrastructure is the morphological analyser, which builds on the state-of-the art morphological analyser HUMOR (Prószéky and Tihanyi 1993) but is using a new annotation set (developed in consensus with theoretical linguists) and is implemented in finite state technology (using the HSFT engine). The infrastructure is open not just in the sense that most of its modules are available in open source to language technology specialists but also it makes its services available to non-developers, eminently targeting digital humanities researchers or even the general public. This is achieved through the web-service that allows the users to copy and paste an input text in a textbox on the website, select the kind of processing required and retrieve the results of the analysis. Figure 4 shows the morphological analysis in vertical form output of text copied into the input text box.

3.3 Automatic word alignment

The forced alignment of the HuComTech transcript with the speech signal is prepared using the WebMAUS tool developed and operated as a web service by the Institute of Phonetics and Speech processing of the Ludwig Maximilan University, München (http://clarin.phonetik.uni-muenchen-.de/BASWebServices/#/services.) The web service accepts a speech file and its transcript in a number of languages, including, fortunately, Hungarian. The size of the speech input file is limited to 200 Mb but the system also operates in batch mode i.e. it can accept a number of speech and text file pairs. There is extensive help and even a couple of YouTube videos to help the non-specialist digital humanities researches.

Even a carefully compiled and checked transcript such as the HuComTech corpus requires preprocessing before WebMAUS can be applied without a hitch. Some of the preprocessing steps are amenable to routine automation like filtering the transcript from any characters that do not correspond to sounds such as any codes, metainformation contained in brackets. Some of the phenomena that may cause misalignment are more subtle and have to do with the mismatch between the raw acoustic signal and its perception by the annotators. Apart from the length of stop consonents, items involved are often voiced hesitation phenomena orthographically put down as a single "ö" in Hungarian whereas its length may far exceed the maximum length of a segment the WebMAUS system is trained to accept. Another typical phenomenon (illustrated in Figure 5) concerns lenition or complete disappearance of sounds in words that are transcribed in full orthographic form.

Figure 5 records the case where the final "t" sound in the word "mint" is dropped between the word final and word initial nasals.

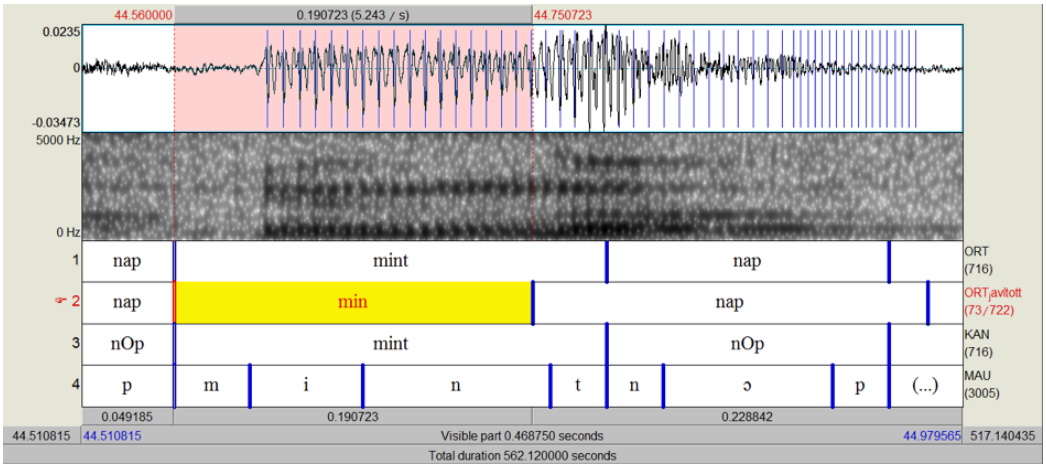

Figure 5. Adjustment of lenition phenomena in WebMAUS output

4 T-Pattern analysis

Multimodal communication involves a myriad of interlocking signals that are quite difficult to tease out and establish their interplay on each other. The HuComTech corpus with its annotation system involving as many as twelve tiers offers such a rich dataset of widely differing modalities (gaze, gesture,

posture, prosody, emotions etc.) that it pauses a challenge for analysis and interpretation. In this section we present a radically new approach, the T-pattern analysis (Magnusson 1996) that promises to uncover hidden patterns in behaviour including human communication. The approach uses sophisticated multivariate analysis that establishes a hierarchical system of recurrent sequence of phenomena forming patterns (T-Patterns) based on the time length between their regular occurrences within a critical interval. Figure 6 shows a schematic way how a sequence of characters with no apparent structure to it reveals an increasingly complex pattern, as T-pattern analysis removes extraneous data and reveals the hidden patterns established on the basis of the temporal relationship between elements of patterns in a given time window. For more information on T-Pattern analysis and Theme, the software tool specially developed to detect T-patterns see Magnusson 2000 and Magnusson et al. 2016.

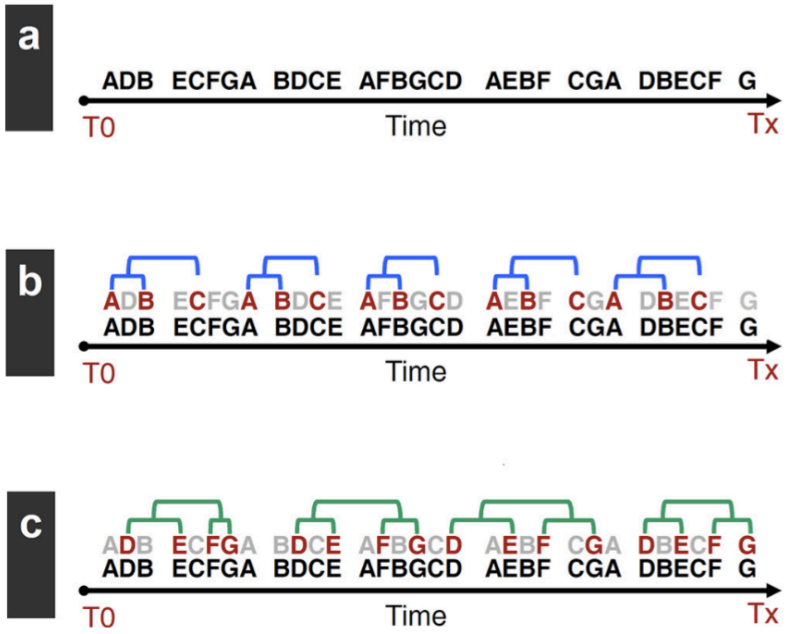

Figure 6. Emerging T-patterns through successive levels of analysis

T-pattern analysis has been applied to a wide range of phenomena and a software tool called Theme is available from PatternVision (http:// http://patternvision.com/) to carry out the research. It is suggested that it provides an exciting new perspective from which the hidden temporal structure of such complex phenomena as multimodal communication can be captured. The T-pattern analysis of the HuComTech corpus poses a challenge not only due to data size, but also the complexity of the nature of multimodal communication: 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.

Hunyadi et al. 2016 contains a tentative first analysis and as a follow-up we present Figure 7 showing T-patterns in multimodal topic management in the 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.

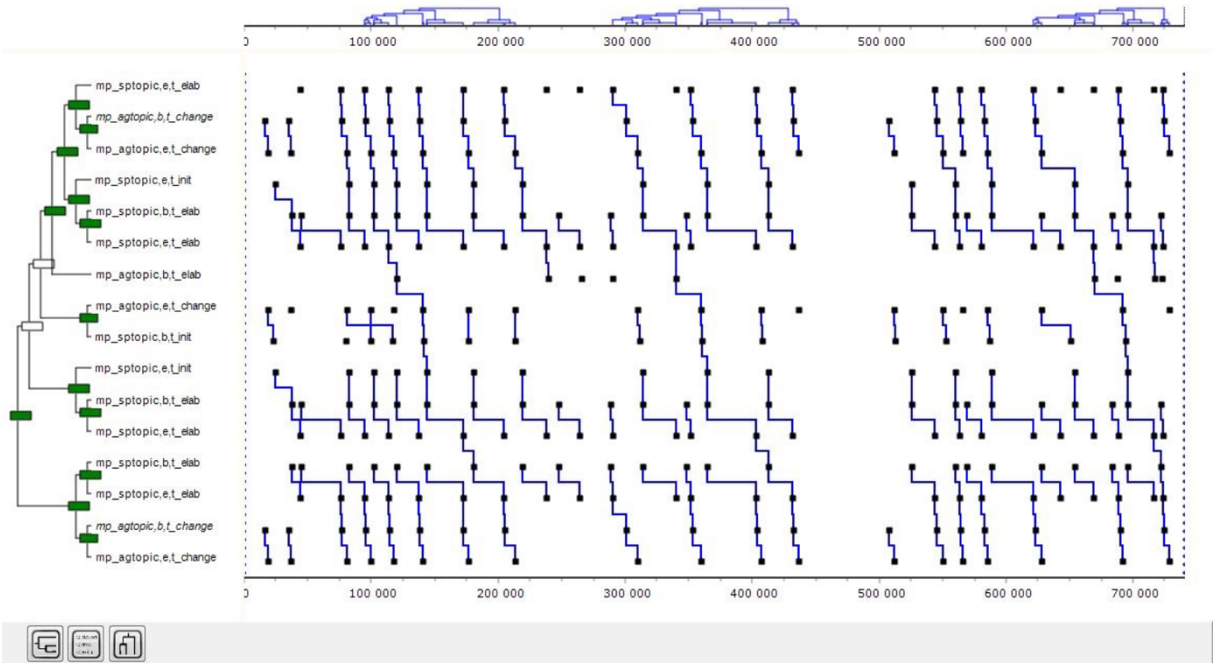

Figure 7 T-patterns in multimodal topic management in the HuComTech corpus

5 Conclusions

In this short paper, we showed how language technology can facilitate the analysis of complex research questions that arise from the study of multimodal human communication. To this end, we introduced a multimodal corpus containing 50 hours of dialogues annotated in rich detail in both unimodal and multimodal manner. We described three tools that were deployed to automate the annotation work, such as recording the prosodic phenomena of pitch, intensity and speech rate, the forced alignment of transcripts and speech signal and the morphological analysis of the text levels of annotation. We showed how T-pattern analysis can present an intriguing perspective to uncover hidden patterns in the temporal structure of multimodal communication.

Acknowledgement

The research reported in the paper was conducted with the support of the Hungarian Scientific Research Fund (OTKA) grants # K116938 and K116 402.

References

Ágnes Abuczki and Ghazaleh Esfandiari-Baiat, 2013 An overview of multimodal corpora, annotation tools and schemes. *Argumentum*, 9:86–98.

Christolihe d'Alessandro and Piet Mertens. 2004. Prosogram: semiautomatic transcription of prosody based on a tonal perception model. In *Proceedings of the 2nd International Conference of Speech Prosody*, pp. 23-26.

Luigi Anolli, Starkey Duncan Jr., Magnus S. Magnusson and Guiseppe Riva (eds.) 2000. *The Hidden Structure of InteractionFrom Neurons to Culture Patterns* http://www.emergingcommunication.com/volume7.html

Paul Boersma and David Weenik (2016). *Praat: doing phonetics by computer* [Computer program] version 6.0.13 http://www.praat.org

J. t'Hart(1976) Psychoacoustic backgrounds of pitch contour stylization. In *IPO- Annual Progress Report 11*, Eindhoven, The Netherlands, pp. 11–19.

László Hunyadi. 2011. Multimodal human-computer interaction technologies. Theoretical modeling and application inspeech processing. *Argumentum* 7:313–329 http://argumentum.unideb.hu/magyar/archivum.html-#7_(2011)

László Hunyadi, Tamás Váradi and István Szekrényes. (2016) The Multimodal HuComTech Corpus: Principles of Annotation and Discovery of Hidden Patterns of Behaviour in *Proceedings of LREC Workshop Multimodal Corpora: Computer Vision and Language Processing* available from https://db.tt/QWI9lLmA

Thomas Kisler, Uwe D. Reichel, Florian Schiel, Christoph Draxler, Bernard Jackl, and Nina Pörner. 2016. BAS Speech Science Web Services - an Update of Current Developments, *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, paper id 668

Thomas Kisler, Florian Schiel and Han Sloetjes. 2012. Signal processing via web services: the use case WebMAUS *Proceedings of Digital Humanities Conference 2012* available from https://www.researchgate.net/publication/248390251_Signal_processing_via_web_services_the_use_case_WebMAUS

Magnus S. Magnusson, M. 1996. Hidden real-time patterns in intra- and interindividual behavior: Description and detection. *European Journal of Psychological Assessment*, 12, 112-123.

Magnus S. Magnusson. 2000. Discovering hidden time patterns in behavior: T-patterns and their detection. *Behavior Research Methods, Instruments, & Computers* 2000, 32 (I), 93-110.

Magnus S. Magnusson, Judee Burgoon and Maurizio Casarrubea (eds.) 2016. *Discovering Hidden Temporal Patterns in Behavior and Interaction. T-Pattern Detection and Analysis with THEME™.* Springer, New York.

Gábor Prószéky and László Tihanyi. 1993. Humor: High-Speed Unification Morphology and Its Applications for Agglutinative Languages. *La tribune des industries de la langue*, No. 10. 28–29., OFIL, Paris, France

István Szekrényes. 2015. ProsoTool, a method for automatic annotation of fundamental frequency In Baranyi Peter (ed.) *2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom*). New York: IEEE, 2015. pp. 291-296.

Paul Taylor. 2000. Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America*, 107(3):1697-1714.