On some linguistic properties of spoken Hungarian based on the HuComTech corpus

László Hunyadi
Dept of General and Applied Linguistics, University of Debrecen, Hungary

The HuComTech corpus is the first multimodal corpus of verbal and nonverbal behaviour based on Hungarian. Its building started in 2009 and has by now been virtually completed. It includes about 50 hours of interactions of formal and informal dialogues between two agents and 110 subjects. The formal dialogues (average duration: 8 minutes) have the form of a job interview, the informal dialogues (average duration: 18 minutes) are actually guided interactions on a variety of everyday topics. The original aim of the corpus was to study and learn those verbal and nonverbal aspects of interactions which can be formalised enough to be implemented in human-machine interfaces. It is based on a generative model of multimodal interactions with the essential goal to capture perception/analysis and production/synthesis in a unified system that can both be instrumental in building such interfaces and modelling human cognition.

Annotations are done at multiple levels most of them with a time resolution of 300 ms. As for the medium to be annotated, there are three kinds: video only, audio only, video+audio. Annotation levels include those requiring the description of some physical, measurable properties as well as those requiring the interpretation of observable events. As for video, annotations include physical properties such as gaze, head movement, hand shape, posture, touch motion, interpretations include perceived emotions. As for audio, there are annotations of various aspects of prosody (absolute values of and stylised F0 and intensity), speech rate and silence as well as the interpretations of perceived emotions. Also on the audio level, force alignment of each running word (their beginning and end) is now being implemented using the Webmaus service and manually adjustment. The combined video+audio medium is mainly used for a wide range of pragmatic and turn management annotations.

Special effort has been made to provide sufficient material for traditional linguistic purposes as well. Accordingly, the corpus includes full morphological annotation (using automatic parsing), an automatic shallow syntactic parsing and, manually, a special annotation for syntactic incompleteness, a well known property of spoken language.

At present the size of the corpus is approaching 2 million single annotations. In order to retrieve information from this vast body of data a database has been implemented. The format is .eaf and it can be searched using the freely available ELAN tool. The database is designed to be accessible remotely from two web sites (The Language Archive, Nijmegen, and RIL, Budapest); until all work has been completed, only a subset of the data I can be reached publicly.

The challenge of studying multimodal human behaviour is that virtually none of the stereotypical primitives of an event of any kind is obligatory and – in sharp contrast to syntax or morphology – the primitives constituting an event may or may not follow adjacency so that a chain of them may also include "noise", i.e. irrelevant to the event data. In order to cope with this seemingly discouraging situation we are applying a special research environment with a specific methodology designed to understand this complex nature of social interaction. The talk will present some research data and results based on pattern recognition using the Theme software. We will highlight verbal and nonverbal behavioural patterns of spoken dialogues drawing examples from syntax organisation as well as prosodic patterning as a function of certain thematic and pragmatic properties, all reflecting the variability of spoken interactions.