Ágnes Abuczki & Esfandiari Baiat Ghazaleh:
An overview of multimodal corpora, annotation tools and schemes
Argumentum 9 (2013), 86-98
Debreceni Egyetemi Kiadó

Ágnes Abuczki & Esfandiari Baiat Ghazaleh

An overview of multimodal corpora, annotation tools and schemes*

Abstract

The aim of the present study is to provide a general overview of multimodal corpora along with their annotation tools and schemes. In the first section an introduction is given to justify the necessity of multimodal approach not only in linguistics and communication studies but also in dialog modeling. Following the justification and the definition of multimodal corpora, three relatively well-known corpora are presented and compared. The considerable differences in their platforms and schemes highlight the necessity of standardization in order to enhance the usability and interoperability of their resulting datasets. Finally, the limitations of multimodal corpus studies are briefly listed.
Keywords: multimodality, corpus linguistics, multimodal corpora, corpus annotation

1 Introduction

Corpus linguistics (henceforth: CL) studies natural language use, that is, performance, in contrast with theoretical linguistics where the focus is on the study of competence. It can be defined as a methodological research approach to studying language variation and use (Biber & Reppen 2012). Since CL uses a large and principled collection of natural texts as a basis for quantitative and qualitative as well as traditional and computational analyses, its research findings have great generalizability and validity. The goals of CL include providing explicit descriptions of language-in-use, finding sequential patterns in language, verifying models of language use and developing algorhythms in various fields of language technology.

Methodological innovations in CL allow linguists to raise fundamentally different kinds of research questions, sometimes providing different perspectives on language variation and language use (Biber & Reppen 2012). For instance, as computing capacities began to increase in the 1970s, CL research interest shifted from the study of written to spoken interaction. It was uncovered with the help of CL that spoken language greatly differs from written text since not all sentences are well-formed, sentence boundaries are often unclear, and that different principles and maxims guide spoken interaction. All these findings led to the necessity of novel units of segmentation in spoken corpora, such as turns or utterances. However, as Knight et al. (2006) argue, spoken corpora are limited as they only have the

* The work/publication is supported by the TÁMOP-4.2.2/B-10/1-2010-0024 project. The project is co-financed by the European Union and the European Social Fund.

provision for presenting data in a single format; that is text, in the form of transcripts of interactions. Spoken corpora with transcripts alone are not sufficient for uncovering the nonverbal-visual aspects of interaction since 'the reflexivity of gesture, movement and setting is difficult to express in a transcript' (Saferstein 2004: 213). As a result, CL research interest has recently shifted to the study of multimodal interaction and its capturing in corpora.

Multimodal corpus (henceforth: MM corpus) research faces two major problems: (1) the time consuming nature of annotation process, therefore, relatively small sizes of annotated MM corpora are available; (2) the lack of existing annotation standards (tools, formats and schemes), especially for coding nonverbal behavior. There are several annotation schemes available that code the nonverbal behavior of speakers; however, most of them do not integrate talk and gesticulation in a coherent fashion. It would be crucial at least to recognize the transitions between intentional and non-intentional movements (e.g. the difference between a real gesture and an accidental movement).

2 The necessity of a multimodal approach in communication studies

The aim of this section is to highlight the multimodal nature of human expression and perception that must be considered in designing corpora. We attempt to briefly answer why we need a multimodal approach in studies of language use. In general, most multimodal studies address the questions how modalities work together as well as to what extent verbal and nonverbal expressions are synchronized.

2.1 The multimodal nature of human interaction

McNeill (1992: 30–32) proposes that gestures must be regarded as part of language: speech and gesture are two modes of expression and they can be regarded as two aspects of the process of an utterance. Similarly, Kendon (2004) argues that a theory of utterance should not begin with a division between speech and gesture. In Kendon's definition, a gesture is a form of human expression, an activity that is significant for the understanding of a speaker's expression, and they often express something complementary to what is being expressed verbally (2004: 107).

2.2 Multimodal perception

Massaro (1987) points out that it is not only our expressions but also our perception that is of multimodal nature since we process not only what we hear but also what we see (facial expressions, lip movements, eyebrow movements, hand gestures of speakers, etc). Inherently, there is both auditory and visual relation among speakers in which the modalities normally complement, sometimes contrast each other. Simultaneous auditory and visual perception is called multimodal perception, and the processing of stimuli coming from various modalities is largely an unconscious process (Massaro 1987). Since auditory and visual perception is both simultaneous and unconscious, the nonverbal components of interaction must also be taken into account when interpreting a communicative event.

Ágnes Abuczki & Esfandiari Baiat Ghazaleh:
An overview of multimodal corpora, annotation tools and schemes
Argumentum 9 (2013), 86-98
Debreceni Egyetemi Kiadó

3 The definition and requirements of MM corpora

A 'multi-modal corpus' is defined as 'an annotated collection of coordinated content on communication channels including speech, gaze, hand gesture and body language, and is generally based on recorded human behavior' (Foster & Oberlander 2007: 307–308). The integration of textual, audio and video records of communicative events in MM corpora provides a platform for the exploration of a range of lexical, prosodic and gestural features of conversation, and for investigations of the ways in which these features interact in real, everyday speech (Knight 2009: 15).

Within various types of MM corpora, we can distinguish two basic types:

- video recordings supplemented with only transcriptions;
- video and audio recordings annotated at multiple levels (based on both audio and video separately).

All three corpora presented in this study belong to the second category which is considered more valuable in communication studies.

Biber & Reppen (2012) list the following requirements of corpora:

- representativity
- validity
- generalizability
- standardized format

We would also complement this list in connection with MM corpora with the requirement that their annotation schemes should be domain and tool-independent, and their labels (within a single level at least) should be mutually exclusive. Moreover, besides its audio and video contents, a usable MM corpus must also have metadata description, annotation guidelines and user's guide in order to provide rigorous guidelines to its coders as well as to ensure its usability for researchers.

4 Annotation tools and query options related to MM corpora

4.1 Annotation and querying tools

Generally, different annotation tools are designed and used to annotate the audio and video contents of a corpus that can later be merged in query systems or databases. For instance, video contents of the HuComTech corpus were annotated in Qannot (Pápay et al. 2011: 330–347), while audio contents were annotated in Praat, a fine grained audio analysis tool (Boersma & Weenink 2007) which enables a much more precise and detailed acoustic analysis than compact multimodal annotation software such as Anvil (Multimodal Annotation and Visualization Tool[1]) or ELAN (Brugman & Russel 2004: 2065–2068). However, Anvil and ELAN offer a lot of benefits to their users since they enable the simultaneous streaming and annotation of both audio and (even multiple) video files in separate windows, and users can specifically design their own annotation scheme and attach multiple tags to one segment in both software programs. Moreover, Anvil allows multiple annotators to work on the same file, and therefore it is able to measure inter-annotator agreement. Concerning the video annotation tool of the HuComTech corpus, a new software program, Qannot was designed instead of Anvil because Anvil sometimes seemed to fail to handle large files, and there was a

[1] ANVIL is freely available at: http://www.anvil-software.org/

risk that annotations might be behind time in these large files. As the annotations were complete, the various annotation files of the HuComTech corpus were merged in an SQL database. Annotations are still stored in SQL and can also be queried in a very user-friendly way using the ELAN software (Brugman & Russel 2004). Custom query options of ELAN include: *N-gram within annotations; Structured search of multiple files; Find overlapping labels within a file;* and *Find left overlaps within a file*, etc. The availability of multimodal annotation tiers enables the systematic and joint search of the temporal alignment and/or synchronous co-occurrences of turns, clauses or specific lexical items with the use of manual gestures, head movement types, gaze directions, eyebrow movement types and posture changes in spontaneous interaction corpora.

4.2 Usability of datasets in novel corpus-driven research areas

With the help of MM corpora searches, the investigation of the temporal alignment (synchronized co-occurrence, overlap or consecutivity) of gesture and talk has become possible. Similarly to corpus-driven approaches that study *lexical bundles* (multi-word sequences) (Biber 2010: 170–172), some of the MM corpus researches are inspired by the notion of *semiotic bundles* (Arzarello et al. 2005) where modeling language production includes the manipulation of resources as well as gesture and talk. Some functional annotation schemes (Allwood et al. 2007) try to code the meaning relations between gestures and co-occurring speech in a systematic way, and label communicative events according to the alignment of speech and gesture. Gestures often co-occur with speech; however, their discursive functions are not always identical. The basic functions of the gestures and speech either ‘overlap’ or are ‘disjunct’, and sometimes synchronous verbalisations and gestures may be more ‘specific’ than the other sign at a given timestamp in the annotation (Evans et al., 2001: 316). Frequency evidence (of any sequential linguistic pattern and co-occurring nonverbal phenomena) found in corpora supports the application of statistical methods in language analysis and modeling. The huge amounts of synchronized data enable the practical and fruitful use of such advanced statistical methods as factor analysis or multidimensional analysis in order to uncover the prototypical features that simultaneously occur in certain communicative acts. Therefore, these methods contribute to the solution of a challenging task in dialog modeling and dialog management, the automatic identification of dialog structure and communicative act types.

5 Examples of MM corpora

This section aims at providing a general overview of MM corpora by describing a few examples of such corpora. The corpora chosen for this purpose are AMI, SmartKom and HuComTech. These three different corpora were chosen in a way to represent the variety of approaches and aims involved in structuring MM corpora. Therefore, they can be contrasted in terms of their different types of discourse following different scenarios, such as meetings, task-based interaction, simulated job interviews and informal conversations. In the following section, each one of them will be described briefly, providing their particular aim, context of use, structure and annotation scheme.

5.1 AMI Corpus

5.1.1 Aim and Context of Use

The AMI or Augmented Multi-party Interaction Corpus is a large MM corpus, involving 100 hours of meetings. Its aim was to develop and integrate meeting browsing technologies in order to support human interaction in meetings. The corpus focuses on language use in a single setting, which is a meeting room, so it is contextually specific, and it only features extracts from one specific discourse context (i.e. meeting discourse), thus its usefulness is limited in studying more informal, interpersonal aspects of language use (Carletta et al. 2005).

5.1.2 Corpus Design

While some of the meetings in this 100-hour long corpus were naturally occurring (35 hours), the majority (65 hours) was elicited using a scenario in which groups of three to four participants played different roles as employees working on a design project in a design team. The data was collected in three smart meeting rooms. In each room 4 cameras, 24 microphones and special tools to capture handwriting and slides were used (McCowan et al. 2005). The language of communication in all meetings was English, while most of the participants were non-native English speakers. Due to this fact, a higher degree of variability in speech patterns can be observed in this corpus compared to other corpora.

5.1.3 Annotation Scheme

The data has been annotated at a number of levels covering various verbal and nonverbal features. Table 1 summarizes the annotation scheme used in this corpus (Carletta et al. 2005).

Levels of annotation	Annotated elements
Speech transcription	orthographic transcription of speech, also annotating speaker change boundaries and word timings
Named entities	reference to people, artifacts, times and numbers
Dialogue acts	act typology used for group decision-making
Topic segmentation	major topic and sub-topic segments in meetings
Group activity	activities that groups are engaged in
Abstractive summaries	decisions that were made during the meeting, problems or difficulties that occurred during the meeting, next steps
Extractive summaries	extract a subset of the dialogue acts of the meeting, such that they form a kind of summary and then link those extracted dialogue acts with sentences
Emotion	different dimensions which reflect the range of emotions that occur in the meetings
Head and hand gestures	movements of both the head and the hands of the participants
Location of the individual	location of the individual in the room or the posture if seated
Focus of attention	what the participants are looking at (which people or artifacts)

Table 1: Annotation scheme used in AMI corpus

The AMI Meeting Corpus is publicly available at http://corpus.amiproject.org containing media files (audio files, video files, captured slides, whiteboard and paper notes) and also all annotation dimensions described in Table 1. However, the annotated dimensions as well as

the implicit metadata for the corpus are difficult to exploit by NLP tools due to their particular coding schemes.

5.2 SmartKom Corpus

5.2.1 Aim and Context of Use

The SmartKom corpus was built as part of the SmartKom project in Germany with the goal to develop an intelligent computer-user interface allowing for more natural interaction for users. SmartKom is one of the first corpora that combines the analysis of acoustic, visual and tactile modalities. It is a task-oriented corpus since that data were gathered and annotated having specific aims and has therefore a limited re-usability for other purposes (Schiel et al. 2002).

5.2.2 Corpus Design

The data were gathered using so called Wizard-of-Oz experiments. In this experiment, participants were asked to work on a specific task while cooperating with the system. The subjects thought that they were really interacting with an existing system, but in reality the system was simulated by two humans from another room. 96 different users were recorded across 172 sessions of 4.5 minutes each. In each Wizard-of-Oz session, spontaneous speech, facial expressions and gestures of the subjects were recorded and later annotated. The language of communication was German in all recorded sessions (Steininger et al. 2002).

5.2.3 Annotation Scheme

The data has been annotated on several levels covering various features. Table 2 summarizes the annotation scheme used in this corpus (Steininger et al. 2002). This corpus is available for academic use only through the META-SHARE website[2]. META-SHARE is an international organization which builds a multi-layer infrastructure and aims at providing an open, distributed, secure, and interoperable infrastructure for the language technology domain. Release SKAUDIO 1.0 contains all audio channel recordings of the SmartKom corpus covering all three scenarios (Public, Home and Mobile) used in the technical setup.

Levels of annotation	**Annotated elements**
Speech transliteration	orthographic transliteration on word level of spontaneous dialogue between user and machine
Head gestures	three morphological categories, head rotation, head incline forward/backward, head incline sideward
Hand gestures	functional and intentional (not morphological), based on the intention of the user's assumed goal
Emotional facial expressions	joy/gratification, anger/irritation, helplessness, pondering/reflecting, surprise, neutral, unidentifiable episode
Prosody	pauses between phrases, words and syllables, irregular length of syllables, emphasized words, strongly emphasized words, clearly articulated words, hyper articulated words, words overlapped by laughing

Table 2: Annotation scheme used in SmartKom corpus

[2] META-SHARE website: http://www.meta-net.eu/meta-share

The annotation of the nonverbal-visual components of interaction in both AMI and SmartKom is somewhat incomplete and inapplicable for an in-depth analysis of interpersonal communication since they both predominantly aim at capturing movements and fail to label the visual features with their meanings or functions in the particular discourse context. For instance, AMI annotates movements of the head and the hands of the participants and SmartKom annotates head gestures based on three morphological categories, head rotation, head incline forward/backward, head incline sideward. At the same time, we can find alternative annotation schemes among MM corpora which try to integrate talk and gesticulation in a coherent, truly multimodal scheme, such as MUMIN (A Nordic Network for MUltiModal INterfaces) developed by Alwood et al (2007) or HuComTech (described in Section 5.3 below, and in Hunyadi et al. 2012a in detail).

5.3 HuComTech Corpus

5.3.1 Aim and Context of Use

The MM HuComTech corpus was built in the framework of the Human-Computer Interaction Technologies project[3]. Hungarian was the language used in all recorded conversations. The aim of building the corpus was to investigate the nature and temporal alignment of verbal and nonverbal features of spontaneous speech as well as to compare the characteristics of formal and informal communication as the corpus involves both formal and informal conversations (between dialogue partners). It is useful to include two types of conversation, formal and informal for purposes of comparative analysis since formal conversations follow rules and strong social norms and involve the use of keywords, symbolic gestures, high conscious control, while the structure and scenario of informal conversations are not so strict (overlapping turns, inconsistencies, discrepancies between modalities, iconic gestures, other casualties often occur). This distinction is important for the sake of defining spontaneity within interaction, and drawing our technological limits (Pápay et al. 2011).

5.3.2 Corpus Design

The material contains 50 hours of both formal and informal dialogues from 121 speakers. The dialogues were recorded in a soundproof studio. The participants were both audio and video taped during their conversations. The informal dialogues centered on everyday topics, mostly about university and other life experiences while formal dialogues followed the typical scenario of simulated job interviews. Both the formal and informal dialogues were guided by pre-designed questions that intended to provoke various emotions such as happiness, sadness, anger and surprise (Pápay et al. 2011).

5.3.3 Annotation Scheme

The data was annotated on different levels coding various features. The annotation was carried out based on either one modality (audio only or video only) or two modalities (audio and video). This corpus also includes syntactic, prosodic and pragmatic annotation. The syntactic annotation was restricted to the identification and classification of clauses and sentences (Hunyadi et al. 2012a). In the prosodic annotation, the F0 movements were annotated

[3] HuComTech website: https://hucomtech.unideb.hu/hucomtech/

(Hunyadi et al. 2012b). Table 3 and 4 briefly summarize the annotation schemes used in this corpus.

Levels of annotation	**Annotated elements**
Speech transcription	orthographic transcription of speech for both speakers
Discourse labels	turn take, turn give, turn keep and backchannels
Emotions	happy, tense, sad, recall, surprise, neutral, other
Intonational phrases	head clause, subordinate clause, embedding, insertion, back channel, hesitation, restarts, iterations and silence

Table 3: HuComTech annotation scheme based on audio-only

Levels of annotation	**Annotated elements**
Facial expressions	happy, tense, sad, recall, surprise, neutral, other
Gaze	gaze direction of the speaker using various directional labels
Eyebrow	movement of the speaker's eyebrow using various directional labels
Head shifts	movement of the speaker's head using various directional labels
Hand shape	shape of the speaker's hand
Touch motion	the speaker touching one or some of his/her body parts
Posture	body shifts of the speaker using various directional labels
Deictic	the speaker points at him/herself or something else present in the room
Emotion	happy, tense, sad, recall, surprise, neutral, other
Emblems	attention, agree, doubt, disagree, refusal, block, doubt-shrug, finger-ring, hands-up, more-or-less, number, one-hand-other-hand, surprise-hands and other

Table 4: HuComTech annotation scheme based on video and audio

The pragmatic annotation was carried out on two separate levels, multimodal (based on both audio and video) and unimodal (based on video only), the latter being a novel approach in pragmatic corpus annotation.

Multimodal pragmatic annotation codes communicative functions and speaker intentions, not necessarily mirrored in surface structure. For instance, an interrogative sentence may express a directive function. The major aim of the multimodal pragmatic annotation was to find the underlying structure of communicative behavior as well as the visual, acoustic and verbal correlates of different communicative acts (Abuczki et al. 2011: 179–201).

As for the unimodal annotation, the aim was to grasp communicative events based solely on visual input. Table 5 and 6 outline the pragmatic annotation schemes used in this corpus.

Levels of annotation	**Annotated elements**
Communicative act types	constative, directive, commissive, acknowledging and indirect
Supporting acts	backchannel, politeness marker and repair
Thematic control	topic initiation, topic elaboration and topic change
Information	units of new information

Table 5: HuComTech multimodal pragmatic annotation scheme

Levels of annotation	**Annotated elements**
Turn management	start speaking successfully, breaking in, intend to start speaking and end speaking
Attention	call attention, pay attention
Agreement	agreement and disagreement and its degree: default case of agreement, full agreement, partial agreement, uncertainty, default case of disagreement, blocking and uninterested
Deixis	deictic gestures not annotated in the video annotation
Information structure	received novelty was annotated

Table 6: HuComTech unimodal pragmatic annotation scheme

This corpus is not publicly available yet. It is available for academic use only through the META-SHARE website.

6 Standardization

In the previous section a brief overview of three different MM corpora was provided. These three MM corpora were different in their approaches and also annotation schemes. In each one of them, different nonverbal behaviors were selected and annotated using different labels defined in specific ways serving their own purpose of study. Therefore, in designing a MM corpus there are no conventionalized prescriptions that determine which behaviors to mark-up, how to describe these behaviors, which labels to use in the annotation scheme and how to integrate everything in the corpus database to cover all multimodal elements of discourse. As a result, generalizing standards for codification of visual and spoken data should be considered as a priority in multimodal research (Knight 2009). Recently, many researchers and research teams have started to lay the foundations for designing a standardized scheme for annotating various features of spoken utterances, gaze movement, facial expressions, gestures, body posture and combination of any of these features. They have the aim to integrate these aspects to develop re-usable and international standards for investigating language and gesture-in-use in user-friendly environments. The outcome of such international interdisciplinary initiations and cooperations are for instance the META-SHARE, the HUMAINE[4] (Human-Machine Interaction Network on Emotion) and the SEMAINE[5] (The Sensitive Agent) projects. The HUMAINE project developed the XML-coded EARL (Emotion Annotation and Representation Language) scheme[6] to annotate the dimensions and intensity of emotions. However, it can only be used with the Anvil software. Its restricted usability highlights the necessity of tool- and domain-independent annotation schemes.

The SAIBA project developed the tool- and domain-independent Behavior Markup Language (BML) (Vilhjalmsson et al. 2007). BML is a widely used method to unify the key interfaces in multimodal human behavior generation processes. ISO standard 24617-2 for dialogue acts developed in recent years is an example of a widely accepted international standard (Bunt et al. 2012). It is an application-independent dialogue act annotation scheme that is both empirically and theoretically well founded. It covers typed, spoken, and multimodal dialogue, and it can be effectively used by both human annotators and automatic

4 HUMAINE: http://emotion-research.net/projects/humaine/aboutHUMAINE
5 SEMAINE: http://www.semaine-project.eu/
6 EARL-scheme: http://emotion-research.net/projects/humaine/earl

annotation methods. In designing this ISO standard for dialogue act annotation, most concepts were applied from the DIT++ taxonomy of dialogue acts.[7] Table 7 summarizes the annotation scheme used in this ISO standard.

General-purpose functions	**Information-seeking functions**: *propositional questions, check questions, set questions and choice questions*
	Information-providing functions: *inform, agreement, disagreement, answer, confirm and disconfirm*
	Commissive functions: *promise and offer*
	Directive functions: *instruct/command and request*
Dimension-specific functions	**Auto-feedback functions**: *positive and negative*
	Allo-feedback functions: *positive and negative*
	Time management functions: *stalling and pausing*
	Turn management functions: *turn accept, turn assign, turn grab, turn keep, turn release and turn take*
	Discourse structuring functions: *interaction structuring and opening*
	Own and partner communication management functions: *completion, correct misspeaking, signal speaking error, retraction and self correction*
	Social obligation management functions: *initial greeting, return greeting, initial self introduction, return self introduction, apology, accept apology, thanking, accept thanking, initial goodbye and return goodbye*

Table 7: ISO standard for dialogue act annotation scheme

7 Limitations

This section aims at outlining some of the major limitations in multimodal corpus research. First, each corpus with its design and annotation scheme serves specific purposes. Therefore, it does not cover all elements and types of communicative events. Most gesture annotation schemes only describe the size, trajectory and direction of movements that gestures are comprised of. Some of them are restricted to the analysis of only one large component of human nonverbal expressions according to kinesic properties, such as the detailed analysis of face by the Facial Action Coding System (FACS scheme) (Ekman et al. 2002). Another movement-based annotation scheme comes from McNeill (1992) who labels a variety of manual gestures as well besides the facial expressions and head movements, but it still fails to integrate talk and gesticulation in a truly multimodal framework. Second, the relatively small size of all corpora (tens or hundreds of hours of annotated interaction) makes the observation of all sorts of linguistic phenomena in a natural proportion impossible (Knight 2009: 99–100). Third, participants are typically seated, which highly restricts their movements and posture changes. Fourth, the number and positioning of cameras restrict the size of analyzable environment related to each speaker. Frequently, each of the speakers is recorded by different cameras, which makes it impossible to investigate such phenomenon as joint attention or eye contact. Synchronization of the cameras and their joint streaming in annotation software such as ELAN can be an acceptable solution in these cases. Last but not least, we must mention that annotation (for instance, pragmatic annotation) involves interpretation to some extent, so it is never completely free of intuitions. However, this problem can be handled if the

[7] DIT++ taxonomy is available at http://dit.uvt.nl/

annotation guidelines are consistently followed and if annotators regularly attend meetings where ambiguous cases can be discussed.

Although we are aware of the existing shortcomings of MM corpora, we believe there is no better alternative resource available for analyzing spontaneous language-in-use than a MM corpus, since current corpus linguistics provides us various methodologies to process and analyze interaction in its entire multimodal nature.

References

Abuczki Á., Bódog A. & Németh T. E. (2011): A pragmatikai annotáció elméleti alapjai az ember–gép kommunikáció modellálásában. In: Németh T. E. (szerk.): *Ember–gép kapcsolat: A multimodális ember–gép kommunikáció modellezésének alapjai.* Budapest: Tinta Kiadó, 179–201.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C. & Paggio, P. (2007): The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Proceedings of the International Language Resources and Evaluation Conference (LREC)* 41(3-4), 273–287.

Arzarello, F., Ferrara, F., Paola, D. & Robutti, O. (2005): The genesis of signs by gestures. The case of Gustavo. In: Chick, H.L. & Vincent, J.L. (eds.): *Proceedings of the 29th Conference of the International Group for the PME*, 1. Melbourne, AU: University of Melbourne, 73–83.

Biber, D. (2010): Corpus-based and corpus-driven analyses of language variation and use. In: Heine, B. & Narrog, H. (eds.): *The Oxford Handbook of Linguistic Analysis*.

Biber, D. & Reppen, R. (2012): *Corpus Linguistics*. SAGE Publications Ltd.

Boersma, P. & Weenink, D. (2007): *Praat: doing phonetics by computer 5.0.02.* University of Amsterdam: Institute of Phonetic Sciences. http://www.praat.org

Brugman, H. & Russel, A. (2004): Annotating multi-media / multi-modal resources with elan. In: Lino, M., Xavier, M., Ferreire, F., Costa, R. & Silva, R. (eds.): *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. Lisbon: Portugal, 2065–2068. (The ELAN 4.5.1 software can be downloaded at http://tla.mpi.nl/tools/tla-tools/elan/download/)

Bunt, H., Alexandersson, J., Choe, J.W., Fang, A.Ch., Hasida, K., Petukhova, V., Popescu-Belis, A. & Traum D. (2012): ISO 24617-2: A semantically-based standard for dialogue annotation. *Proceedings of the International Language Resources and Evaluation Conference (LREC)*, 430–437.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D. & Wellner. P. (2005): The AMI Meeting Corpus: A Pre-Announcement. *Proceedings of the Second international conference on Machine Learning for Multimodal Interaction*, 28–39.

Ekman, P. & Friesen, W.V. (1969): The repertoire of non-verbal behavior: Categories, origins, usage and coding. *Semiotica* 1(1), 49–98.

Ekman, P. & Friesen, W.V. (1978): *Facial Action Coding System (FACS): A technique for the measurement of facial action*. Palo Alto, CA: Consulting Psychologists Press.

Ekman, P., Friesen, W.V. & Hager, J.C. (2002): *Facial Action Coding System: The Manual on CD ROM*. Salt Lake City: A Human Face. (Description of FACS and information on the CD ROM is available at: http://face-and-emotion.com/dataface/facs/description.jsp)

Ekman, P. & Rosenberg, E.L. (1998): *What the Face Reveals*. Oxford: Oxford University Press.

Enfield, N.J. (2009): *The Anatomy of Meaning. Speech, gesture, and composite utterances.* Cambridge: Cambridge University Press.

Evans, J.L., Alibali, M.W. & McNeill, N.M. (2001): Divergence of verbal expression and embodied knowledge: Evidence from speech and gesture in children with specific language impairment. *Language and Cognitive Processes* 16(2-3), 309–331.

Foster, M.E. & Oberlander, J. (2007): Corpus-based generation of head and eyebrow motion for an embodied conversational agent. *Proceedings of the International Language Resources and Evaluation Conference (LREC)* 41 (3/4), 305–323.

Hunyadi L., Földesi A., Szekrényes I., Staudt A, Kiss H., Abuczki A. & Bódog A. (2012a): Az ember–gép kommunikáció elméleti-technológiai modellje és nyelvtechnológiai vonatkozásai. In: Kenesei I., Prószéky G. & Várady T. (szerk.): *Általános Nyelvészeti Tanulmányok XXIV. Nyelvtechnológiai kutatások.* Budapest: Akadémiai Kiadó, 265–309.

Hunyadi, L., Szekrényes, I., Borbély, A. & Kiss H. (2012b): Annotation of spoken syntax in relation to prosody and multimodal pragmatics. *Proceedings of IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom),* 537–541.

Kendon, A. (2004): *Gesture. Visible Action as Utterance.* Cambridge: CUP.

Knight, D., Bayoumi, S., Mills, S., Crabtree, A., Adolphs, S., Pridmore, T. & Carter, R.A. (2006): Beyond the Text: Construction and Analysis of Multi-Modal Linguistic Corpora. *Proceedings of the 2nd International Conference on e-Social Science*, Manchester, 28–30 June 2006 [online]. Available at: http://www.ncess.ac.uk/events/conference/2006/papers/ abstracts/ KnightBeyondTheText.shtml

Knight, D, (2009): *A Multimodal Corpus Approach to the Analysis of Backchannelling Behaviour* (doctoral dissertation). The University of Nottingham.

Massaro, D.W. (1987): *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Hillsdale, New Jersey: Lawrence Erlbaum.

McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., Post, W., Reidsma, D. & Wellner. P. (2005): The AMI Meeting Corpus. *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research.*

McNeill, D. (1992): *Hand and Mind: What Gestures Reveal about Thought.* Chicago: University Of Chicago Press.

Pápay, K., Szeghalmy, S. & Szekrényes, I. (2011): HuComTech Multimodal Corpus Annotation. *Argumentum* 7, 330–347.

Saferstein, B. (2004): Digital technology- methodological adoption: Text and video as a resource for analytical reflectivity. *Journal of Applied Linguistics* 1.2, 197–223.

Schiel, F., Steininger, S. & Türk, U. (2002): The SmartKom Multimodal Corpus at BAS. *Proceedings of the International Language Resources and Evaluation Conference (LREC).*

Steininger, S., Rabold, S., Dioubina, O. & Schiel, F. (2002): Development of the User-State Conventions for the Multimodal Corpus in SmartKom. SmartKom. *Proceedings of the International Language Resources and Evaluation Conference (LREC).*

Vilhjalmsson, H., Cantelmo, N., Cassell, J., Chafai, N.E., Kipp, M. & Kopp, S. (2007): The Behavior Markup Language: Recent Developments and Challenges. In: *Proceedings of Intelligent Virtual Agents* (IVA 2007). LNAI, 4722. Berlin, Heidelberg: Springer, 99–111.

Ágnes Abuczki
University of Debrecen
Department of General and Applied Linguistics
Pf. 24
H-4010 Debrecen
abuczki.agnes@gmail.com

Ghazaleh Baiat Esfandiari
University of Debrecen
Department of General and Applied Linguistics
Pf. 24
H-4010 Debrecen
esfandiari.gh@gmail.com