The POETICON corpus:
Capturing language use and sensorimotor experience in everyday interaction

K. Pastra[1], C. Wallraven[2], M. Schultze[2], A. Vatakis[1], K. Kaulard[2]

[1]Institute for Language and Speech Processing, Athena Research Center, Athens, Greece

[2]Max Planck Institute of Biological Cybernetics, Tuebingen, Germany

{kpastra,avataki}@ilsp.gr, {christian.wallraven,michael.schultze,kathrin.kaulard}@tuebingen.mpg.de

Abstract

Natural language use, acquisition, and understanding takes place usually in multisensory and multimedia communication environments. Therefore, for one to model language in its interaction and integration with sensorimotor experiences, one needs a representative corpus of such interplay. In this paper, we will present the first corpus of language use and sensorimotor experience recordings in everyday human:human interaction, in which spontaneous language communication has been recorded along with corresponding multiview video recordings, recordings of 3D full body kinematics, and 3D tracking of objects in focus. It is a twelve-hour corpus which comprises of six everyday human:human interaction scenes, each one performed 3 times by 4 different English-speaking couples (interaction between a male and a female actor), each couple acting each scene in two settings: a fully naturalistic setting in which 5-camera multi-view video recordings take place, and a high-tech setting, with full body motion capture for both individuals, a 2-camera multiview video recording, and 3D tracking of focus objects. The corpus has been developed within an EU-funded cognitive systems research project, POETICON (http://www.poeticon.eu), and represents a new type of language resources for cognitive systems. Namely, a corpus that *reveals the dynamic role of language in its interplay with sensorimotor experiences* and which allows *one to computationally model this interplay*.

1. Introduction

The need for intelligent multimedia conversational systems has been indicated since the early days of Artificial Intelligence (AI); for example, Winograd's SHRDLU system was one of the first, primitive prototypes that could answer simple user queries regarding a commonly shared visual scene (Winograd, 1972). The path towards the development of embodied or non-embodied conversational systems has been a long one since; embodied conversational agents, conversational robots of the new millennium, or simply intelligent interfaces that allow for multimodal input and output, are just some recent manifestations of this quest for developing intelligent, cognitive systems, able to communicate with humans in as much a human-like way, as possible (Pastra and Wilks, 2004).

Natural language is undoubtedly, one of the basic communication means, one that is perceived, used and acquired in a multimedia and multisensory environment; therefore, the development of intelligent conversational systems demands that one studies and models language in its interaction with other media and with sensorimotor experiences. There are a number of language resources that capture different aspects of language, as well as its interaction in multimedia settings. However, currently there exists no resource that captures language use along with other sensorimotor experiences.

2. The related resources landscape

Depending on the applications developed for, unimodal, text or speech corpora cover a variety of genres, domains, and languages. Naturally occurring everyday language production in the form of human:human conversations has always been a quest in developing spoken corpora (Wilks et al., 2008); recently, recordings of such language-based interaction in an everyday life setup comprise a unique in its kind, annotated corpus (Sherstinova, 2009).

Multimedia corpora comprising of both language and other media such as video or images are also increasingly built for different system development needs. In their simplest form, such corpora are collections of labeled images, such as the PASCAL object recognition challenge collection. Collections of annotated audiovisual documents such as the COSMOROE corpus of TV travel series (Pastra, 2008b) go a step further capturing a mixture of read and spontaneous speech, and video rich in object/scene depiction, and human activity (i.e. body movements and gestures, facial expressions). These are corpora of rich multimedia information, which capture a mixture of multimodal, individual human behaviour and multimodal human:human interaction in informal, everyday settings.

Conversation corpora of human:human interaction in semi-formal and formal settings, such as meetings, have also been captured in recordings of rich language interaction and videos of corresponding gestures, head movements, and gaze direction (Carletta, 2007). In a semi-formal studio setting, spoken dialogue between highly familiar with each other individuals, forms part of

another corpus which also comprises of the corresponding video recordings of gestures, gaze direction, and facial expressions (van Son et al., 2009). Though multimedia and multimodal, such corpora remain restricted to capturing only two sensory channels, i.e. the auditory (speech perception) and the visual (text, visual action perception, and visual object/scene perception). Actually, beyond language perception, the only sensorimotor experience captured is limited to visual perception of gestures, body-movements, and objects/scenes.

On the other hand, corpora of simulated human:computer interaction which usually rely on Wizard of Oz techniques for inducing human:machine conversations capture guided language-based interaction, but they go beyond video capture of sensorimotor experience to capturing tactile input modalities (e.g., pointing gestures touching the system screen;Scheil et al., 2002, Wilks et al., 2008). Even this very simple, binary form of touch sensing (touch vs. no touch) is something extra in terms of capturing sensorimotor information during interaction, something that could also be captured in recently emerged corpora of human:human interaction and simultaneous speech-based interaction with a machine (handheld device; Schiel et al., 2008).

However, it is not only intelligent human:computer interfaces that have led researchers to building multimedia and multisensorial data collections. The rapidly emerging "network science'", i.e. the study of human behavior not in isolation but in relation to other humans and the environment, dictates that appropriate, rich multisensorial data are systematically collected, measuring all possible aspects of human behaviour and interaction. For example, for the needs of applications related to monitoring the everyday activities of the elderly or of different types of patients, not only video recordings, but also motion acceleration measurements, biosensors measuring human physiological data (e.g., respiration, heart rate etc.) as well as other measurements from environmental sensors (e.g., contact sensors on doors to notify one when door is opening/closing, pressure sensors on chairs that identify sitting events) are being systematically collected (Zouba et al., 2009). The ultimate multi-sensory monitoring data collections come from House_n Placelab, a "living lab" space, properly equipped with cameras and sensors for measuring different aspects of human interaction (Intille et al., 2006, Logan et al., 2007).

A database that captures the full range of human actions as well as both human-object and human-human interactions is the Motion Capture database of the Carnegie Mellon University (CMU; http://mocap.cs.cmu.edu.); the database comprises of 3D full body kinematic data and corresponding videos on a large variety of human actions. However, these are individual recordings of each action separately, in a lab environment and do not capture human activity in an interaction setting. On the contrary, the CMU kitchen capture database comprises of 18 participants performing 5 simple cooking recipes each, in a fully equipped kitchen-lab (Frade et al., 2008). Full body motion capture and acceleration data, limited physiological data, and multiview video recordings are included in the database, as well as RFID-based object identification and tracking. However, the recordings are of individuals carrying out an everyday activity on their own (i.e. no interaction), and actually no language/speech has been included. Last, the Technical University of Muenchen (TUM) Kitchen Data Set comprises also of multiview video recordings of individuals setting a table. RFID and magnetic sensors have been used for object identification and tracking, and full body motion capture and labels of events have also been included (Tenorth et al., 2009). This database does not contain any language-based communication either.

Evidently, language-based human:human interaction in everyday activities along with corresponding sensorimotor measurements are not currently available. In the following section, we present the first such corpus, the POETICON corpus, which we consider a first step towards the goal of capturing as many aspects of human:human natural interaction as possible.

3. The POETICON Corpus

The main objective of the POETICON project is the development and automatic extension of a conceptual knowledge base, the PRAXICON, in which each concept is represented both with symbolic (i.e. language) and sensorimotor representations and it is defined though its network of semantic relations with other concepts (Pastra, 2008). Related software, cognitive and neurophysiological experiments as well as demonstration of the use of the resource in intelligent systems and humanoids are being developed in the project. The POETICON corpus is used for the development/testing of the software, as source of stimuli for the experiments and as pool of information to be extracted for populating the PRAXICON.

The corpus comprises of six everyday human:human interaction scenes, each one performed 3 times by 4 different English-speaking couples (interaction between a male and a female actor), each couple acting each scene in two settings: a fully naturalistic setting in which 5-camera multi-view video recordings take place, and a high-tech setting, with full body motion capture for both individuals, a 2-camera multiview video recording, and 3D tracking of focus objects. All recordings include full language-based interaction (dialogue) which though pre-scripted for providing a guide to the actors, it is natural and spontaneous due to the actors left free to improvise based on the general script lines. Each scene lasts approximately 2-7 minutes depending on the scene and the actors, while the duration of the whole corpus is approximately 12 hours. The scenes are related to activities one may perform in a dinning room/kitchen, such as ***changing the***

pot of a plant, cleaning the room, setting the table, preparing a Greek salad, preparing Sangria, and *making a parcel.* The scenes are scripted following the form of play scripts, in which dialogue as well as guidance on emotions, and direction related explanations are given to the actors (see Figure 1). This was done in order to make sure that the enactment of the scenes will be rich in actions, interaction with objects, emotions, and attitudes on the part of the actors, as well as occurrence of sudden, unexpected events that capture anomalies in one's movements (e.g., someone breaking a plate, slipping and falling on the floor, someone behaving in a childish way by jumping up and down, etc). Furthermore, the scripted dialogue includes not only situated communication with many references to objects in the environment and the task at hand but also non-situated discussion, mainly small talk.

Person B takes the place mats and places them on the table. Then he goes to the kitchen-table/cupboard and opens another cabinet door to get the plates.
(open door)
He picks up two plates and takes them to the dinner table and places them there.
One plate is set on the edge of the table and falls down. Person B turns around and says *"oh!"*
(fear)
Person A looks and asks "Did you break something?"
(anger) (look_back)
Person B: "No…Thank God it does not break easily! Phew" and puts the plate back on the table.
(evasive, then smiling)
When Person B is done with all the plates, he walks to the kitchen table where the mugs are standing.

Figure 1: Sample of the script used in the POETICON corpus. The scripts have been annotated with action- and emotion-related information in order to enrich the interaction and data acquisition.

3.1 Naturalistic Setting

The naturalistic setting of the corpus was recorded in a custom-designed setting in order to look as natural as possible (with furniture, carpets, curtains, and everyday objects; see Figure 2). The scenes were recorded with 5 high-definition camcorders (Canon HF100, resolution of 1960x1400 pixels, 2 of those camcorders have a wide-angle lens, DHG 0,75x Wide Angle Converter 52 mm) from different positions (see Figure 2).

Figure 2: Sample frames of the *naturalistic setting* of the POETICON corpus. Each scene was captured by 5 cameras in different positions: A) overview from one room corner, B) overview from the opposite room corner, C) view of the kitchen table from above, D) view of the cupboard from above, and E) view of the center of the scene where the interaction takes place.

3.2 High-Tech Setting

The high-tech setting of the corpus was recorded using the same furniture set-up as the one used in the naturalistic setting. The scenes were recorded with 2 high-definition camcorders (Canon HF100, resolution of 1960x1400 pixels, wide-angle lens, DHG 0,75x Wide Angle Converter 52 mm) from different positions (see Figure 3). The movement of the 2 actors was captured with 2 Moven motion capture suits (Xsens technologies). The position of the 2 actors was also tracked with the Vicon motion capture system, using 2 helmets with tracking markers. Finally, some of the objects that were central in the interaction tasks were tracked with the Vicon motion capture system (see Figure 4).

3.3 Corpus Animations

The high-tech setting of the POETICON corpus allowed for the generation of the corpus animations. Specifically, the animations were created from the motion capture data using 3ds Max. The animations include the actors, the furniture present in the scene (e.g., kitchen table, cupboard, chairs etc.), and the Vicon-tracked objects (e.g., broom, dustpan etc.; see Figure 5). The motion capture data of the actors from the Moven suits was imported into 3ds Max and positional and rotational drifts were corrected manually using the Vicon data and the high-tech movies as a reference. The furniture and the objects are represented as boxes and can be removed from the scene; they can also be replaced by the images of the real objects/furniture. The objects movement was represented using the Vicon data and the two actors as reference.

3.4 Further processing and annotation

The corpus has been post-processed for synchronization of all cameras, integration of full-body kinematic data and 3D object tracking data, and creation of animation videos of the integrated kinematic data. The videos of all the scenes recorded were cut using iMovie and then exported in QuickTime-movie format (.mov). Camera synchronization was achieved by using the audio track in the naturalistic setting and both the audio track and the start of the actions for the kinematic recordings.

The POETICON corpus has been transcribed using Transcriber (Barras, Geoffrois, Wu, and Liberman 2000) and it has been semantically annotated in Anvil (Kipp, 2004) using the COSMOROE semantic relations (Pastra 2008b).

The POETICON corpus will be available to the research community under a Creative Commons Attribution - Non Commercial - Share Alike license upon completion of the project.

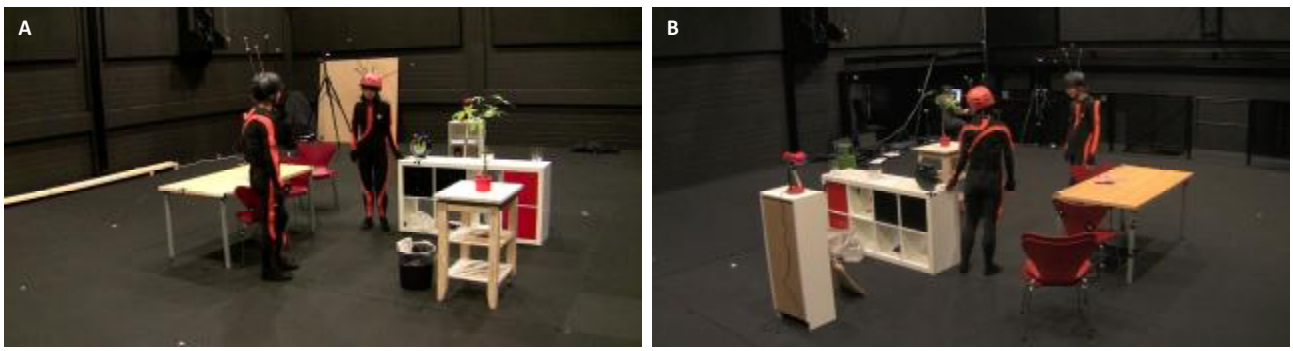

Figure 4: Sample frames of the high-tech setting of the POETICON corpus. Each scene was captured by 2 cameras in different positions: A) overview from one room corner, B) overview from the opposite room corner.

Scene	Cleaning the kitchen	Preparing a greek salad	Lay the kitchen table for a dinner for two	Changing the pot of a plant	Preparing some drinks: Sangria	Sending a parcel
Objects tracked	Table, kitchen-table, cupboard, chairs, broom, dustpan, clock	Table, kitchen-table, cupboard, chairs, broom, dustpan, clock, salad bowl	Table, kitchen-table, cupboard, chairs, broom, dustpan, clock, salad bowl, candlestick	Table, kitchen-table, cupboard, chairs, watering pot	Table, kitchen-table, cupboard, chairs, pitcher	Table, kitchen-table, cupboard, chairs, toy guitar

Figure 3: The POETICON corpus includes 3D tracking data of objects. This figure shows the objects tracked per scene.

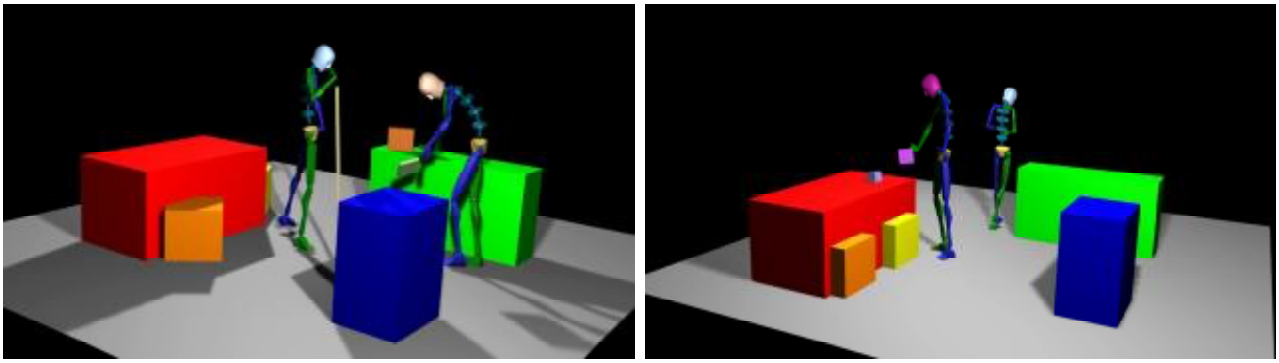

Figure 5: Sample frames of the POETICON corpus animations. These animations were created based on the high-tech recordings for all six scenes.

4. Conclusion

The POETICON corpus demonstrates that, though technologically challenging, it is feasible for one to capture natural language interaction along with sensorimotor experiences. This corpus represents an extension of the state-of-the-art on several levels:

A) A corpus with natural, yet script-controlled interactions at all interaction levels (human:human, human:object) in well-defined scenarios;

B) A corpus that includes data from several different individuals with multiple recordings of the same actors in the same scene, thus allowing for intra-scene, intra-individual variance but also for within actors comparisons of movements etc.

C) A corpus with both "natural" and matched "high-tech" recordings of multiple scenes which makes it suitable both for cognitive experiments and computational modeling.

D) A corpus with a large amount of multisensory information (multiple camera angles, visual and audio data, language, 3D kinematic data, and 3D tracking data of objects) from different sensors and different modalities that can support analysis across a large number of dimensions from 2D analysis of video streams up to complex models of 3D articulation.

We consider the POETICON corpus a first step towards the development of corpora of everyday human:human interaction in which language-based communication is recorded ---and thus further studied and modeled--- in its interplay with environmental and biological sensors, as well as motion capture and object tracking; it is a first step towards corpora that will capture all aspects of human interaction with other humans and with the environment.

5. Acknowledgements

Work reported in this paper is being funded by the European Commission Framework Program Seven in the framework of the POETICON project (FP7-ICT-215843).

6. References

Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (2000). Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication 33*(1-2).

Carletta, J. (2007). Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation, 41*(2), 181--190.

Frade F., J. Hodgins, A. Bargteil, X. Martin Artal, J. Macey, A. Castells, and J.Beltran. (2008). Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) Database. Technical Report CMU-RI-TR-08-22, Robotics Institute, Carnegie Mellon University.

Intille S., K. Larson, E. Munguia Tapia, J. Beaudin, P. Kaushik, J. Nawyn, and R. Rockinson (2006). Using a live-in laboratory for ubiquitous computing research. In K. P. Fishkin, B. Schiele, P. Nixon, and A. Quigley (Eds.) *Lecture Notes in Computer Science*, vol. 3968, Berlin Heidelberg: Springer-Verlag, pp. 349--365.

Kipp, M. (2004). *Gesture Generation by Imitation - From Human Behavior to Computer Character Animation*. Boca Raton, Florida: Dissertation.com.

Logan M., J. Healey, M. Philipose, E. Munguia Tapia, and S. Intille. (2007). A long-term evaluation of sensing modalities for activity recognition. *Lecture Notes in Computer Science*, vol. 4717, Springer-Verlag, pp. 483--500.

Pastra, K. and Wilks, Y. (2004). Vision-Language Integration in AI: a reality check. In Proceedings of the Sixteenth European Conference in Artificial Intelligence, pp. 937--941.

Pastra K. (2008). PRAXICON: the development of a grounding resource. In Proceedings of the Fourth International Workshop on *Human-Computer Conversation*, Bellagio, Italy.

Pastra K. (2008b). COSMOROE: A Cross-Media Relations Framework for Modelling Multimedia

Dialectics. *Multimedia Systems*, vol. 14 (5), Springer Verlag, pp. 299--323.

Schiel, F., Steininger, S., and Türk, U. (2002). The SmartKom Multimodal Corpus at BAS. In Proceedings of the Third Language Resources and Evaluation Conference (LREC 2002), pp. 200--206.

Schiel, F. and Mögele, H. (2008). Talking and Looking: the SmartWeb Multimodal Interaction Corpus. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC08), Marrakech, Morocco.

Sherstinova T. (2009). The Structure of the ORD Speech Corpus of Russian Everyday Communication. In V. Matouˇsek and P. Mautner (Eds.), *Lecture Notes in Artificial Intelligence*, vol. 5729, Springer-Verlag, pp. 258 -- 265.

Tenorth M., Bandouch , J., and M. Beetz, (2009). The TUM Kitchen Data Set of Everyday Manipulation Activities for Motion Tracking and Action Recognition. In the IEEE International Workshop on *Tracking Humans for the Evaluation of their Motion in Image Sequences* (THEMIS).

van Son R., Wesseling, W., Sanders, E., and van den Heuvel, H. (2009). Promoting *free* Dialog Video Corpora: The IFADV Corpus Example. In M. Kipp et al. (Eds.), *Multimodal Corpora. Lecture Notes in Artificial Intelligence*, vol. 5509, Springer-Verlag, pp. 18--37.

Wilks, Y., Benyon, D., Brewster, C., Ircing, P. and Mival, O. (2008). Dialogue, Speech and Images: The Companions Project Data Set. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC08), Marrakech, Morocco.

Winograd, T. (1972). *Understanding Natural Language*. Academic Press.

Zouba N., Bremond, F., and Thonnat, M. (2009). Multisensor Fusion for Monitoring Elderly Activities at Home. In Proceedings of the Sixth IEEE International Conference on *Advanced Video and Signal Based Surveillance*, pp. 98--103.