1st Symposium on Multimodal Communication

University of Malta, Valletta
17-18 October 2013

BOOKLET OF ABSTRACTS

Organising committee
Patrizia Paggio
Jens Allwood
Elisabeth Ahlsèn
Kristiina Jokinen
Costanza Navarretta

Local organisers
Lucienne Bugeja
Albert Gatt
Luke Galea
Patrizia Paggio
Alexandra Vella

Scientific Committee
Albert Gatt, University of Malta
Alexandra Vella, University of Malta
Anton Nijholt, University of Twente
Catherine Pelachaud, CNRS Telecom ParisTech
Dirk Heylen , University of Twente
Isabella Poggi, Università degli Studi Roma Tre
Jean-Claude Martin, CNRS-LIMSI, Paris
Joakim Gustafson, KTH, Stockholm
Jonas Beskow, KTH, Stockholm
Kirsten Bergmann, University of Bielefeld
Maria Graziano, Lund University
Marie Alexander, University of Malta
Mariet Theune, University of Twente
Mary Ellen Foster, Heriot-Watt University
Massimo Zancanaro, FBK, Trento
Michael Kipp, Hochschule Augsburg
Nick Campbell, Trinity College Dublin
Onno Crasborn, Radboud University Nijmegen
Roman Bednarik, University of Eastern Finland
Stefan Kopp, University of Bielefeld
Thomas Hanke, University of Hamburg

Programme

17 October

8:30 -9:00	*Registration*
9:00- 9:15	*Welcome*
9:15-10:15	*Keynote speech 1* Onno Crasborn: "The representativeness of sign language corpora"
10:15-10:45	*Coffee break*
10:45-12:45	*Paper session* – Lucie Metz and Virginie Zampa: "Plat'In: A French sign language teaching/learning platform" – Maria Galea: "The Sutton SignWriting glyph-set and its use for the writing of nine different sign languages" – Jonas Beskow, Simon Alexanderson, Kalin Stefanov, Morgan Fredriksson, Britt Claesson and Sandra Derbring: "The Tivoli System: A sign-driven game for children with communicative disorders" – Emer Gilmartin, Shannon Hennig, Ryad Chellali and Nick Campbell: "Exploring sounded and silent laughter in multiparty social interaction - audio, video and biometric signals"
12:45-14:00	*Lunch*
14:00-15:30	*Paper session* – Jens Allwood, Stefano Lanzini and Elisabeth Ahlsèn: "On the contributions of different modalities to the interpretation of affective-epistemic states" – Farina Freigang and Stefan Kopp: "Exploring the speech-gesture semantic continuum" – Mathieu Chollet, Magalie Ochs and Catherine Pelachaud: "Investigating non-verbal behaviours conveying interpersonal stances
15:30-16:00	*Coffee break*
16:00-17:30	*Paper session* – Magdalena Lis and Costanza Navarretta: "Classifying the form of iconic hand gestures from the linguistic categorization of co-occurring verbs" – Bjørn Wessel-Tolvig: "Up, down, in & out: Following the Path in speech and gesture in Danish and Italian – Magdalena Lis and Fey Parrill: "Referent type and its verbal and gestural representation: test on English multimodal corpus and WordNet 3.1

18 October

9:30- 10:30	*Keynote speech 2* Kristiina Jokinen: "Studying multimodal communication with eye-tracking"
10:30-11:00	*Coffee break*
11:00-12:30	*Paper session* – Héctor P. Martínez and Georgios N. Yannakakis: "Deep learning for multimodal feature extraction" – Patrizia Paggio and Alexandra Vella: "Overlaps in Maltese: a comparison between task-oriented and conversational data" – Jean-Claude Martin: "Quantification of self and Mechanical Turk: Two future pillars for multimodal corpora research?
12:30-14:00	*Lunch*
14:00-15:00	*Paper session* – Isabella Poggi and Francesca D'Errico: "Parody of politicians: a multimodal tool for political satire" – Ekaterina Morozova, Anna Khokhlova and Svetlana Mishlanova: "Verbal and gestural representation of the space-time relation in the oral narrative"
15:00-16:00	*Coffee break and posters* – Ryosaku Makino and Nobuhiro Furuyama: "Relationship between home position-formation and storytelling" – Svetlana Polyakova and Svetlana Mishlanova: "Multimodal Metaphors of Health: an Intercultural Study"
16:00-17:00	*Paper session* – Laura Vincze and Isabella Poggi: "Precision gestures in oral examinations and political debates" – Mary Suvorova and Svetlana Mishlanova: "Multimodal representation of the concept of happiness in Russian students' narrative"
17:00-17:15	*Closing session*

Author list

Mishlanova, Svetlana	Multimodal Metaphors of Health: an Intercultural Study
Morozova, Ekaterina	Verbal and gestural representation of the space-time relation in the oral narrative
Navarretta, Costanza	Classifying the form of iconic hand gestures from the linguistic categorization of co-occurring verbs
Ochs , Magalie	Investigating non-verbal behaviors conveying interpersonal stances
Paggio, Patrizia	Overlaps in Maltese: a comparison between task-oriented and conversational data
Parrill, Fey	Referent type and its verbal and gestural representation: test on English multimodal corpus and WordNet 3.1.
Pelachaud, Catherine	Investigating non-verbal behaviors conveying interpersonal stances
Poggi, Isabella	Parody of politicians: a multimodal tool for political satire
Poggi, Isabella	Precision gestures in oral examinations and political debates
Polyakova, Svetlana	Multimodal Metaphors of Health: an Intercultural Study
Stefanov, Kalin	The Tivoli System - A Sign-driven Game for Children with Communicative Disorders
Suvorova, Mary	Multimodal Representation of the Concept of Happiness in Russian Students' Narrative
Vella, Alexandra	Overlaps in Maltese: a comparison between task-oriented and conversational data
Vincze, Laura	Precision gestures in oral examinations and political debates
Wessel-Tolvig, Bjørn	Up, down, in & out: Following the Path in speech and gesture in Danish and Italian
Yannakakis, Georgios N.	Deep Learning for Multimodal Feature Extraction

Plat'In: a French Sign Language teaching/learning platform --
Preliminary studies and specifications

Abstract

This paper deals with the creation of a French Sign Language teaching/learning platform entitled Plat'In. This platform aims to fill a gap in e-learning tools for French Sign Language (LSF). In this presentation we focus on the methodological choices made to develop Plat'In taking into account the importance of multimodal aspects of LSF.

First of all, in Europe, there are a number of existing tools for Sign Language. One such project is "*DELE*" an e-learning platform intended for Italian Sign Language (CAPUANO *et al*; 2011). Another project was born in 2004 for Greek Sign Language (EFTHIMIOU *et al*, 2004). In France, there has only been one attempt a platform for teaching LSF or teaching in LSF, entitled "*E-LSF*" and created by Dalle in 2009 (DALLE; 2010). The only one functional learning platform that we observed is dedicated to Quebec Sign Language[1] owned by "la fondation des sourds du Québec".

However in France the law passed in 2005 ("*Loi pour l'égalité des droits et des chances, la participation et la citoyenneté des personnes handicapées (loi 2005-102 du 11 Février 2005)*"), has accorded LSF with special status. Since then, LSF has become a recognized language among other natural languages, with the same needs and interests with respect to teaching and learning. Nevertheless no formal method has been conceived yet to support and organize teaching of and in LSF in order to accompany this official acknowledgement. Including the sequence of decision which have been taken after it, such as the creation of an LSF secondary school teaching diploma, among others.

The study of existing LSF distance-learning tools reveals a lot of lacks in their conception. Moreover, we observed that the tools conceived for distance-learning applications aren't adapted to deal with distinctive features of sign languages and specific needs of deaf people. Indeed, we observed the presence of audio information or a lot of written French information in those projects, which excluded most deaf people lacking expertise of written French language[2]. Also, deaf people are a really mixed public and many of them would probably find difficult to work with contents and instructions if it's in written French language (GILLOT; 1998).

A preliminary study to determine the needs of potential users was done with around fourteen researchers and teachers specializing in LSF and deafness. Following this study we decided to distribute a second form another hundred LSF learners to know their specific needs. We wanted to know how the platform would be perceived if it existed.

Results show us its extreme usefulness and the importance to take into account the multimodal aspects of sign language learning particularly because it's a vision and gesture-based communication. We also wanted to collect the fears and expectations of the potential users. First,

[1] http://www.courslsq.net/ewac/lsq/info.php

[2] Exemples among others: "la LSF en 15 étapes" by Monica Companys, "Apprentissage interactif de la LSF" by Patrice Carillo or "la vache et le chevalier" by CRDP.

learners explained their expectations for quality personalized learner assessment, the types of content and the essential aspect of interaction with a teacher and with the group (class).

Our choice to split up Plat'In users' into several categories was made to answer as much as possible to people's needs. First, we defined teacher and learner categories. Next, we decided to focus the distinction on proficiency in writing French language for those two types of users. Lastly we categorized learners by age distinguishing between adult and children and by LSF skills level. Also we established twelve types of learner and four teacher's categories.

Regarding learners, we chose to define an educational pathway fixed for all users. Nevertheless, according learners will remain a certain amount of freedom: they could choose to follow it, at their own rhythm. As well, learners will evolve freely within the platform.

LSF is a vision and gesture-based form of communication; therefore it was essential for us to preserve these multimodal aspects of sign language communication, as it may be the best way to enable users to interact in LSF.

We also considered two types of interactions: (a) Human-machine interaction and (b) interaction between users of the platform.

- **(a)** Human-machine interactions are interactions built between users and the platform. In our case, this interaction is mediated by an animated avatar sign. This avatar has several advantages: learners who don't have enough expertise of written French could navigate easily within the platform; more advanced learners could acquire all the spatial information and grammatical specificities of sign language; etc.
- **(b)** The main objective of interaction between platform users is to create a total LSF language immersion. We found this situation to be optimal when the learner is immersed in into an environment where LSF is the main language used. For those purposes, we suggest to integrate some features of web 2.0 tools like video forums and video chats.

These types of interaction will support learners and teachers. Learners will be corrected by their referent teacher and will keep track of their own progress. They will also be able to practice and improve LSF skills. Teachers, for their part, will not only discuss and exchange in LSF but they will have the possibility to exchange some teaching data thanks to a dedicated module named "*salle des profs*" [staff room].

Forty people mixed between beginners and those who have followed one semester of LSF lessons have already tested some of our choices. This experiment confirmed that the platform was very well received by the testers and has highlighted several further improvements to be made. We noted particularly their attempt of new content, addition of pedagogical feed-back and scores.

Perspectives

To conclude, we are conscious that this is a work-in-progress and will evolve in the future months. However, our recent research tells us that our project rejoins social and public interest. Much research remains to be achieved in particular about choices that have to be made for learners who don't master written French language.

Bibliography

CAPUANO D., TAGARELLI De MONTE M., GROVES K., ROCCAFORTE M., TOMASUOLO E., (2011) "A Deaf-centred E-Learning Environment (DELE): challenges and considerations", *Journal of Assistive Technologies*, Vol. 5 Iss: 4, pp.257 - 263

DALLE P. (2010) « Technologies de l'information et de la communication au service de la LSF ». *Contact n°6 : Grandir et apprendre en langue des signes : oui, mais comment ?* : Actes, journée d'étude du 29 janvier 2011 Groupe d'Etudes et Recherches sur la Surdité. L'Harmattan : 89–109

EFTHIMIOU, E., SAPOUNTZAKI, G., KARPOUZIS, K. & FOTINEA, SE. (2004), "Developing an e-Learning platform for the Greek Sign Language", *Lecture Notes in Computer Science (LNCS)*, in K. Miesenberger, J. Klaus, W. Zagler, D. Burger, Springer (eds), Vol. 3118, 2004, 1107-1113.

GILLOT, D. (1998). Le droit des sourds : 115 propositions. Rapport parlementaire au Premier Ministre

DELE. Plateforme d'apprentissage de la Langue des Signes Italienne. 2010

E-LSF. Plateforme d'enseignement de la LSF URL: http://enseignement-lsf.com/ last view on may, 2nd, 2013

Plateforme d'apprentissage de la Langue des Signes Québécoise (LSQ) http://www.courslsq.net/ewac/lsq/info.php last view on may, 2nd, 2013

The Sutton SignWriting glyph-set and its use for the writing of nine different sign languages

Abstract

The Sutton SignWriting system (ISWA 2010) consists of a BaseSymbol glyph-set of 652 glyphs. However no given sign language is likely to use all these glyphs (Sutton, 2008-2011). This paper provides a snapshot of the use of the ISWA 2010 glyph-set for the writing of nine different sign languages, where it is seen how each sign language uses a smaller glyph-set for the representation of their languages in written form.

SignPuddle 2.0 is an online software program, created by Steve Slevinski, where it is possible to create and store SignWriting text directly on the web. There are public and private SignPuddles and each Puddle has an editor or a group of editors who can moderate and edit what enters the Puddle. There are also 'personal puddles', where using a USB stick a writer can use SignPuddle but does not need internet connection. In the Dictionary SignPuddles, single-sign entries inputted, and in the Literature SignPuddle long-text entries (utterances/sentences) are inputted. This paper evaluates the 81 SignPuddles that are available from the homepage http://www.signbank.org/signpuddle/ .

From these 81 SignPuddles, ten have been identified as 'actively' using the Literature Puddles[1]. These are Puddles representing ASL (American Sign Language), LSM (Maltese Sign Language), Nicaraguan Sign Language, German Sign Language, Tunisian Sign

[1] The standard for a 'active' Literature SignPuddle is based on the Literature Malta Archive Puddle (LMAP) which is the main SignPuddle under investigation. The LMAP contains a substantial amount of children's stories and translated Bible passages and Catholic mass rites and prayers. Therefore the criteria of an 'active' Literature SignPuddle is based on its size that is either similar or larger in size to the LMAP.

Language, Czech Sign Language, Norwegian Sign Language, Brazilian Sign Language (Libras) and Spanish Sign Language. ASL uses two different Literature Puddles: The ASL Bible Puddle and the US Literature Puddle. The remaining 71 'inactive' SignPuddles are not analyzed further.

This paper will discuss some issues related to the evolution of SignWriting: from its use as a writing system for all sign languages to its application when writing a specific sign language; hence the development of SW into different orthographies. The comparative analysis of the ten different Puddle glyph-sets presented in this paper reveals some of these issues.

Discussion is then turned to whether a glyph-set of a given sign language can be further reduced to fully represent a 'grapheme-set' that is more fully representative of the phonological inventory of that given sign language. The focus of this section is on the Malta Literature Puddle and glyph-set. For the other eight sign languages, only a few examples of written signs that use low frequency glyphs are presented in order to illustrate the similar current of development of SignWriting when applied to specific languages.

The findings from the investigation of the active SignPuddles, although only a beginning, point towards the merits of SignPuddle (and SignWriting) for sign linguistic investigation.

References

Sutton, V. (2008-2011). *SignWriting alphabet: Read and write any sign language in the world; ISWA 2010 International SignWriting Alphabet 2010* [Electronic version] . California, San Diego: Centre for Sutton Movement Writing. Retrieved June 18, 2013 from http://www.signwriting.org/archive/docs7/sw0636_SignWriting_Alphabet_Manual_2010.pdf

The Tivoli System – A Sign-driven Game for Children with Communicative Disorders

Jonas Beskow, Simon Alexanderson, Kalin Stefanov
KTH Speech, Music and Hearing
Lindstedtsvägen 24, 10044 Stockholm, Sweden

Britt Claesson, Sandra Derbring
DART
Kruthusgatan 17, 411 04 Göteborg, Sweden

Morgan Fredriksson
Liquid Media
Lindstedtsvägen 24, 10044 Stockholm, Sweden

We describe Tivoli, a multimodal game and training application for Swedish key word signing, targeted at children with communicative disorders. The game has a built-in sign recognizer that allows players to interact with the game through signing. In addition it features a cartoon-like signing avatar with naturalistic movements driven by motion capture.

Key word signing

Sign Language and different forms of sign based communication is important to large groups in society. In addition to members of the deaf community, that often have sign language as their first language, there is a large group of people who use verbal communication but rely on signing as a complement, often known as key word signing. Key word signing is a form of AAC (augmented and alternative communication) and is often used amongst and with individuals with some form of communication disability such as developmental disorder, language disorder, cerebral palsy or autism, as a means of reinforcing the communicative situation.

Key word signing schemes follow the word order of the spoken language, and are as such not to be confused with sign language. However, the individual signs are typically borrowed from sign language. For example, the Swedish form of key word signing, TSS (tecken som stöd – signs as support) borrow its signs from SSL (Swedish Sign Language). As such, these communication support schemes do away with the grammatical constructs in sign language and keep only parts of the vocabulary. One important difference between SSL and TSS is that the latter is poorly formalized and described, and the extent and manner in which it is taught differ widely between different parts of the country. While many deaf children have sign language as their first language and are able to pick it up in a natural way from the environment, potential TSS users often do not have the same opportunity to be introduced to signs and signing. The Swedish Tivoli project aims at creating a fun yet instructive learning environment where children can pick up signs in a game-like setting.

The Tivoli Game

The diverse target audience for Tivoli is children with significant differences in attention span, memory, communicative abilities and motor skills. In order to create a game that was appealing and enjoyable with this in mind, we focused on finding the suitable game mechanics that would result in what the children considered fun. A number of different game prototypes were developed and tested with the children until we found a concept that the test subjects enjoyed so much they did not want to quit playing the game.

The setting of the game is an amusement park after dark. The children are greeted by a cicerone in a high hat who leads them through a sequence of games loosely tied together

in a story where the kids help the cicerone solve situations through playing games. In one of the small games, the animals of the amusement park zoo goes on a roller coaster ride (see figure 1), each time they go around an animal falls of (not hurting itself, naturally) and the player then tells the system who is missing. In another game someone has thrown a lot of things in the wishing well. These items are only partially visible over the surface. The player has to say tell the game what it is in order to fish it out of the water.
The interaction in the game is done by the player signing to the system. The signing is done in order to answer a question or to select something out of a set, and the signs are identified by a dedicated sign recognizer, described below.

Figure 1: Screenshot of the Tivoli game.

Sign Recognition
Sign recognition is a challenging task. The Tivoli recognizer is required to identify the signs produced by the child and distinguish them from other signs, and indicate whether or not it is the right one. Large inter-subject variability is expected which complicates the task. On the other hand, only isolated signs are to be recognized, and the active vocabulary at a given time is small (3-10 words), which helps. Furthermore the sign recognition module should be able to cope with difference in environment, lighting, subject clothing, and background. To this end we have chosen to capture the user using a depth sensor (Microsoft Kinect) which among other benefits provides robust tracking of the user's limbs and face.
For the current first version of the recognizer, we recorded 10 adult novice signers using the Kinect sensor. Each user performed 51 signs, repeated 5 times. HMM-based recognizers were then trained to classify the signs based on the recorded motion, either in speaker dependent or speaker independent mode. In the speaker dependent case, accuracy for the 51-word vocabulary was close to 90% (5-fold cross validation, over the five instances of each sign). For the speaker independent case, accuracy was 63% (10-fold cross validation, over the 10 signers).

Discussion
The Tivoli system is currently in a working prototype phase. The recognizer performance seems to be accurate enough for the current task, given that the active vocabulary in every instant in the game will be much smaller than the 51-word set used in the experiments described above. The next step is to carry out evaluations of the complete system with users in the target group. These evaluations will show whether the system fills the pedagogical function of strengthening the sign vocabulary and general sign usage of the test persons.

Exploring Sounded and Silent Laughter in Multiparty Interaction - Audio, Video and Biometric Signals

Emer Gilmartin[1], Shannon Hennig[2], Ryad Chellali[2], Nick Campbell[1]

[1] Speech Communications Lab, Trinity College Dublin
[2] PAVIS Group, Istituto Italiano di Tecnologia
gilmare@tcd.ie

1 Background

Human conversational interaction is a multi-faceted, multi-modal, and multi-functional activity, in which participants filter information from a bundle of signals and cues, many of them temporally interdependent. Interaction is largely studied through corpus collection and analysis, with recordings capturing as much as possible of the signal bundle through multiple channels, including audio, video, motion-capture, shared whiteboard data, and bio signals. Much work concentrates on corpora of interactions with a clear short-term goal; these 'task-based' interactions include information gap activities (map-tasks [1], spot the difference [2], ranking items on a list [3]) and real or staged business meetings[4], [5]. Such interaction relies heavily on the exchange of verbal information. However, the immediate task in real-life conversation is often not so clear and the purpose of some interaction may be better described as a long term goal of building and maintaining social bonds; non-verbal information may carry more importance in these situations. Communication is situated, and its characteristics can vary with the type of the current interaction or 'speech-exchange system' [6]. Thus, the study of different types of interaction, and indeed stretches of social and task-based communication within the same interaction sessions, should help discriminate which phenomena in one type of speech exchange system are generalizable and which are situation or genre dependent. In addition to extending our understanding of human communication, such knowledge will be useful in human machine interaction design.

As part of our larger exploration of how aspects of interaction vary in social (chat) and task-based scenarios, we are investigating the occurrence and role of laughter in multiparty interaction. Laughter is a universally observed element of human interaction, predominantly shared rather than solo, older than language, and aiding social bonding [7]. It punctuates rather than interrupts speech [8], and manifests in a range of forms – from loud shared bouts to short, often quiet chuckles. It is described as a stereotyped exhalation of air from the mouth in conjunction with rhythmic head and body movement [9], and is thus multimodal, with visual and acoustic elements available to the interlocutor's senses.

In our investigations we use corpora of non-scripted (spontaneous) multiparty interaction: the task-oriented AMI meetings corpus [5], and the conversational TableTalk [10], d64 [11], and DANS (described below) corpora. Previous work explored relationships between laughter and topic change in AMI and TableTalk [12], [13]. We are currently investigating the interplay of laughter and bio signals in conversational interaction. We are particularly interested in any variation in 'task' vs. 'chat' dialogue, in terms of laughter and of measured electro-dermal activity (EDA). EDA has been linked to levels of emotional arousal [14] and to cognitive load [15], with work in psychology observing lower cognitively loaded implicit cognition in social chat [16], while laughter has been observed to be more frequent in social than in task-based dialogue. It may thus be possible to distinguish important or content-rich sections of dialogue using these metrics.

2 Measurement of Laughter

Several of the corpora we use have been annotated previously for laughter. The use of existing annotations is attractive to researchers, but during our early work on laughter it became apparent that some of the laughter annotation in the corpora we used was not adequate for detailed study, as we encountered many of the problems outlined by other researchers[18], [19]. These problems included mixtures of point and interval annotation for laughter, laughter annotated on the transcription tier at insufficient granularity – e.g. segmented only to the utterance level rather than to word level, and no method for dealing with laughter when it co-occurs with speech in the same speaker. It was clear that the best strategy for laughter work was to create separate laughter tiers on annotations. We therefore created a new annotation scheme using Elan [20] with separate laugh tracks for each speaker, annotating laughter according to the MUMIN scheme [21]. While re-annotating the TableTalk corpus we noted that many laughs were not acoustically sounded, (a requirement in the MUMIN definition), or too quiet to be picked up by microphone. To capture and explore this 'silent' laughter for our current work, we have expanded our annotation scheme to include two extra passes over the data - adding silent video and audio only annotation tiers. Below we describe the data and laughter annotation scheme used in this work, and outline our research questions.

3 Data and Annotation

To explore silent and sounded laughter and biosignals in conversation we use the DANS corpus. The corpus comprises three sessions of informal English conversation between two women and three men, four native English speakers (American, British, Irish (x2)) and one near-native English speaker (French). The sessions were recorded over three days in a living-room like setting with sofa and armchair. Participants were free to speak about any topic they liked and to move around as they wished. In addition to video and audio recordings the corpus includes measurements of participants' electro-dermal activity (EDA) or galvanic skin response. All participants wore biological sensors (Q-sensors) on both wrists to measure EDA. Q-sensors [17] are unobtrusive wristbands which measure electrodermal activity (galvanic skin response) in the wrist – the level of conductance based on the level of perspiration on the skin. Annotation of laughter in DANS was performed in three passes – video only, audio only, and video with sound. For the video only ('silent') passes two annotators were provided with silent video of the data and asked them to mark intervals where they saw laughter, while the audio only version was created by two annotators marking sound recordings of the data with intervals where they heard laughter. These two annotations along with the third annotation based on audio and video ensure full coverage of laughter in the corpus and are now being used to analyze silent and sounded laughter. The corpus was also segmented into turns, and annotated for pauses, gaps, and backchannels.

4 Analysis and Future Work

We are currently analyzing the annotated DANS corpus in terms of silent and sounded laughter distribution, bio signals, topic, and turn taking, addressing the following questions:

1. Does EDA vary with silent or sounded laughter, laugher role (speaker/backchanneller), shared/solo laughter?
2. Is silent laughter functionally different to sounded laughter – is it more likely to be solo, backchannel laughter?
3. Can laughter distribution and/or EDA be used as a marker of task-based vs. chat interaction?

We will also extend previous analysis of shared and solo laughter in relation to topic change to the DANS data. The resulting insights will add to our model of chat vs. task-based dialogue.

During our work on laughter in conversational corpora we have noted the need to re-annotate, and then expand our annotation scheme in view of observations during manual annotation. While data annotation is time-consuming and labour-intensive work, it is invaluable for a fuller understanding of the dynamics of human interaction. Indeed, close examination of data has revealed subtleties that may have been missed had we simply used pre-existing annotations.

Acknowledgements

This work is supported by the Fastnet Project – Focus on Action in Social Talk: Network Enabling Technology funded by Science Foundation Ireland (SFI) 09/IN.1/I2631, and by the Università degli Studi di Genova and the PAVIS department at the Istituto Italiano di Tecnologia

References

[1] A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, and others, 'The HCRC map task corpus', *Lang. Speech*, vol. 34, no. 4, pp. 351–366, 1991.

[2] R. Baker and V. Hazan, 'LUCID: a corpus of spontaneous and read clear speech in British English', in *Proceedings of the DiSS-LPSS Joint Workshop 2010*, 2010.

[3] A. Vinciarelli, H. Salamin, A. Polychroniou, G. Mohammadi, and A. Origlia, 'From nonverbal cues to perception: personality and social attractiveness', in *Cognitive Behavioural Systems*, Springer, 2012, pp. 60–72.

[4] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, and A. Stolcke, 'The ICSI meeting corpus', in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, 2003, vol. 1, pp. I–364.

[5] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, and V. Karaiskos, 'The AMI meeting corpus', in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, 2005, vol. 88.

[6] H. Sacks, E. A. Schegloff, and G. Jefferson, 'A simplest systematics for the organization of turn-taking for conversation', *Language*, pp. 696–735, 1974.

[7] P. J. Glenn, *Laughter in interaction*. Cambridge University Press Cambridge, 2003.

[8] R. R. Provine, 'Laughter punctuates speech: Linguistic, social and gender contexts of laughter', *Ethology*, vol. 95, no. 4, pp. 291–298, 1993.

[9] M. Mehu and R. I. Dunbar, 'Relationship between smiling and laughter in humans (Homo sapiens): Testing the power asymmetry hypothesis', *Folia Primatol. (Basel)*, vol. 79, no. 5, pp. 269–280, 2008.

[10] K. Jokinen, 'Gaze and gesture activity in communication', in *Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments*, Springer, 2009, pp. 537–546.

[11] C. Oertel, F. Cummins, J. Edlund, P. Wagner, and N. Campbell, 'D64: A corpus of richly recorded conversational interaction', *J. Multimodal User Interfaces*, pp. 1–10, 2010.

[12] F. Bonin, N. Campbell, and C. Vogel, 'Temporal distribution of laughter in conversation'.

[13] E. Gilmartin, F. Bonin, C. Vogel, and N. Campbell, 'Laughter and Topic Transition in Multiparty Conversation'.

[14] M. E. Dawson, A. M. Schell, and D. L. Filion, 'The Electrodermal System', *Handb. Psychophysiol.*, p. 159, 2007.

[15] Y. Shi, N. Ruiz, R. Taib, E. Choi, and F. Chen, 'Galvanic skin response (GSR) as an index of cognitive load', in *CHI'07 extended abstracts on Human factors in computing systems*, 2007, pp. 2651–2656.

[16] D. Kahneman, *Thinking, Fast and Slow*. Farrar Straus & Giroux, 2011.

[17] M. Z. Poh, N. C. Swenson, and R. W. Picard, 'A wearable sensor for unobtrusive, long-term assessment of electrodermal activity', *Biomed. Eng. IEEE Trans.*, vol. 57, no. 5, pp. 1243–1252, 2010.

[18] K. Laskowski and S. Burger, 'Analysis of the occurrence of laughter in meetings.', in *INTERSPEECH*, 2007, pp. 1258–1261.

[19] K. P. Truong and J. Trouvain, 'Laughter annotations in conversational speech corpora-possibilities and limitations for phonetic analysis', *Proc. 4th Int. Worskhop Corpora Res. Emot. Sentim. Soc. Signals*, pp. 20–24, 2012.

[20] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, 'Elan: a professional framework for multimodality research', in *Proceedings of LREC*, 2006, vol. 2006.

[21] J. Allwood, L. Cerrato, L. Dybkær, and P. Paggio, 'The MUMIN multimodal coding scheme', in *Proc. Workshop on Multimodal Corpora and Annotation*, 2004.

Contributions of different modalities to the attribution of affective-epistemic states

Jens Allwood, Stefano Lanzini and Elisabeth Ahlsén
SCCIIL Interdisciplinary Center
University of Gothenburg

Abstract

Introduction/Background

It is often claimed that multimodal presentation of information provides more redundancy and is therefore easier to interpret correctly than unimodal presentation. There are also studies showing that gestures enhance the comprehension/memory of a spoken message (Beattie and Shovelton, 2011). The contribution of different perceptual modalities to the interpretation of multimodal AES has not been extensively researched. It has, however, been addressed, to some extent, in studies aiming at the automatic recognition and generation of emotional cues in Embodied Communicative Agents (ECA) (see Abrilian et al. 2005). A database of multimodal videorecorded emotional interaction was established by Douglas-Cowie et al. (2000). Complex emotions were analyzed by Buisine et al. (2006) from videorecorded interactions. The analysis was used for creating a model and simulations of combined (superposed or masked) emotions, using an ECA in a perception experiment and the contribution of different modalities was one of the analyzed parameters,

By affective-epistemic states (AES) we mean all those states that involve cognition, perception and feeling (Allwood, Chindamo and Ahlsén, 2012). A perception experiment was performed with the purpose of examining how affective-epistemic states (i.e. emotions, like happy or sad and epistemic states, like surprised) in a dialog were perceived when the data was presented unimodally as either Video only, or Audio only and multimodally in Video+Audio format (cf. also Lanzini, 2013).

Method

Participants
There were 12 participants, 6 men and 6 women, all native speakers of Swedish, at least 20 years old with no background of studies in the field of communication.

Material
Four recordings of First Encounters from the Swedish NOMCO database were displayed to each subject. The NOMCO project collected multimodal spoken language corpora for Swedish, Danish and Finnish, in order to make it possible to carry out collaborative research. The corpora were transcribed and annotated and are available for research. The "First encounter" interactions in the corpora were recorded in a studio setting, where the participants were standing in front of a light background, so that automatic registration of body movements are possible. Gestures were annotated according to an adapted version of the MUMIN annotation scheme for multimodal communication (Allwood et al. 2007) (cf. www.cst.dk/mumin), using the Praat (Boersma and Wenink 2013) and ANVIL (Kipp 2001) tools. Functional annotation is mainly related to communicative feedback (Allwood, Nivre & Ahlsén 1992) and other interaction phenomena (cf. Paggio et al. 2010).
The parts of the recordings that were shown were about 2 minutes long (avg. 2 min, 7 sec).

Procedure

The experiment was administered to each subject individually. Each recording was shown to each participant in a different mode: video, audio or video+audio, i.e.
Participant 1: V+A (rec 1), A (rec 2), V (rec 3)
Participant 2: A (rec 3), V+A (rec 2), V (rec 1) etc.

The subjects were asked to identify which kinds of affective-epistemic states were displayed by the participants in the recording and to provide motivations for their answers.

At the beginning of the experiment, the goal of the study was explained to the subjects and they were given instructions. The meaning of affective-epistemic states was briefly explained and the subjects were told to identify all states including knowledge and/or feelings (Schroder et al., 2011).

The recording was stopped after every 3-4 contributions (avg 15.5 times, avg. 8.3 sec). Every time it was stopped, the subject had to interpret the affective-epistemic state being expressed (if any), how it was expressed (i.e. in what mode) and give a motivation. The subjects were free to choose their own words.The session of the experiment lasted around 75 minutes per subject.

Analysis

The responses were transcribed and coded manually. Around 200 different descriptions of the affective-epistemic states and the behavior used to express them were obtained from the participants in the study. Using intuitive semantic analysis, these were then grouped into the following 7 types of AES: nervousness, happiness, interest, confidence, disinterest, thoughtfulness and understanding.

Results

The results show differences concerning which affective-epistemic states were identified in the different conditions, depending on the mode of presentation. In this study, we focus on the five most common affective-epistemic states, which are: happiness, interest, nervousness, confidence and disinterest.

When video images and sounds are perceived together, one modality can affect the perception of the other modality. This applies to affective-epistemic states like confidence, interest, happiness, disinterest, understanding and a few times to nervousness. When a specific behavior was presented unimodally, the subjects were often more likely to interpret it as a signal of a particular AES, than if it was presented multimodally.

The results are however somewhat different for different AES. An interpretation of happiness was more often given to laughing and smiling in the unimodal video mode than in the video+audio mode, while nervousness was more easily attributed in the unimodal audio mode. This means that a nervously laughing person can, for example, be interpreted as happy in the video mode, but as nervous, when the audio mode is added. For some AES, the multimodal presentation mode resulted in fewer attributions, since the information from the different modalities often, contrary to expectations of redundancy, seem to act as restrictions on each other. In the case of interest, however, the total number of attributions were about the same in the video and audio+video modes, so in this case, multimodality perhaps added to redundancy.

References:

Abrilian, S., L. Devillers, S. Buisine and J.-C. Martin (2005).
EmoTV1: Annotation of real-life emotions for the specification of multimodal affective interfaces. *11th Int. Conf. Human-Computer Interaction (HCII'2005), Las Vegas, Nevada,*

USA, Electronic proceedings, LEA.
Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C. and Paggio, P. (2007) The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing. In J. C. Martin et al. (eds) Multimodal Corpora for Modelling Human Multimodal Behaviour. Special issue of the *International Journal of Language Resources and Evaluation*. Springer.
Allwood, J., Chindamo, M. & Ahlsén, E. (2012). Some suggestions for the study of stance in communication. Paper presented at the *ASE/IEEE International Conference on Social Computing,* Amsterdam, 2012.
Allwood, J., Nivre, J., and Ahlsén, E. (1992). On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9, 1–26.
Beattie, G. & Shovelton, H. (2011). An exploration of the other side of semantic communication: How the spontaneous movements of the human hand add crucial meaning to narrative. Semiotica., 184, 33-51.
Paul Boersma, P. & Weenink, D. (2013):Praat: doing phonetics by computer [Computer program]. Version 5.3.51, retrieved 2 June 2013 from http://www.praat.org/
Buisine, S. Abrilian, S, Niewiadomski, R, Martin, J-C., DeVillers, L. & Pelachaud, C. (2006). Perception of blended emotions: From video corpus to expressive agent. In J. Gratch et al. (eds.) *IVA 2006, LNAI* 4233, pp. 93-106l Heidelberg: Springer-Verlag.
Douglas-Cowie, E., Cowie, R. & Schröder, M. (2000). A new emotion database: considerations, sources and scope. *ITRW on Speech and Emotion*, Newcastle, Northern Ireland, UK, September 5-7, 2000. ISCA Archive. http://www.isca-speech.org/archive.
Kipp, M. (2001). Anvil – A Generic Annotation Tool for Multimodal Dialogue. In *Proceedings of Eurospeech 2001*, pp. 1367 – 1370.
Lanzini, S. (2013). How do different modes contribute to the interpretation of affective epistemic states? University Gothenburg, Division of Communication and Cognition, Department of Applied IT.
Paggio, P., Allwood, J., Ahlsén, Jokinen. K and Navarretta, C. (2010). The NOMCO Multimodal Nordic Resource - Goals and Characteristics. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., & Tapias, D. (Eds.). *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC´10)* Valletta, Malta. May 19-21. European Language Resources Association (ELRA). ISBN 2-9517408-6-7. http://www.Irec-conf.org/proceedings/Irec2010/index.html (PAGGIO10.98).
Schroder, M., Bevacqua, E., Cowie, R., & Eyben, F. et al. (2011) Building autonomous sensitive artificial listeners. *IEEE Transactions. Affective Computing*. Vol. 3:2: 165-183.

Exploring the speech-gesture semantic continuum

Farina Freigang and Stefan Kopp
farina.freigang@uni-bielefeld.de, skopp@techfak.uni-bielefeld.de
Faculty of Technology, Center of Excellence "Cognitive Interaction Technology" (CITEC)
Collaborative Research Center "Alignment in Communication" (SFB 673)
Bielefeld University, P.O. Box 100 131, D-33501 Bielefeld, Germany

In natural conversation, speech and gesture are usually one unit that is either produced or received by a communication partner. However, the relationship between the meaning of speech and the meaning of gesture can differ. Several terms have been used to specify these different relationships, ranging from "redundant" over "supplemental" to "mismatching" information. No consensus about the exact definition of these terms or the appearing variety in how speech meaning and gesture meaning relate to each other has been reached. We argue that this confusion is due to the fact that these terms address different dimensions of the speech-gesture semantic relationship, and therefore can hardly be related directly with each other. In the following, we discuss the terminology and related studies with regard to production and comprehension.

On the side of language *production*, McNeill (1992) already discussed semantic synchrony in general without going further into detail. Alibali and Goldin-Meadow (1993) were the first to report"mismatches" produced by children learning the concept of mathematical equivalence. This term is not completely agreed on by Willems, Özyürek, and Hagoort (2007) who found that the term "mismatch" should be used with an "incongruent" speech-gesture pair and not when gesture conveys "additional" but not contradicting information as speech. They referred to the mismatch phenomenon as speech-gesture "incongruence". Furthermore, Kelly, Özyürek, and Maris (2010) accepted both terms "mismatch" and "incongruence". In this context, other terms have been mentioned, e.g., speech-gesture "concordance", "concurrent" speech-gesture pairs, "redundant" gestures, and "semantic coordination" of speech-gesture pairs. A detailed definition of these terms and a comparison between them is as of yet still missing.

On the side of language *perception*, McGurk and MacDonald (1976) showed that speech perception is not a purely auditory process but that mouth gestures can influence the recipient's interpretation of what has been said by the message giver. Sometimes this interpretation results in a third meaning, different from the speech or mouth gesture own their own. Similar to the McGurk-MacDonald effect, one can assume that observed speech-gesture mismatches or incongruences may lead to a third interpretation by a subject. Habets, Kita, Shao, Özyurek, and Hagoort (2011) looked at semantic congruent and incongruent combinations ("matches" and "mismatches") or "semantic integration" of speech and gesture during comprehension in an EEG study and found that "mismatching gesture-speech combinations lead to a greater negativity on the N400 component in comparison with matching combinations" (p. 1852). This suggests a cognitive basis for what counts as mismatching in terms of whether speech and gesture can be integrated.

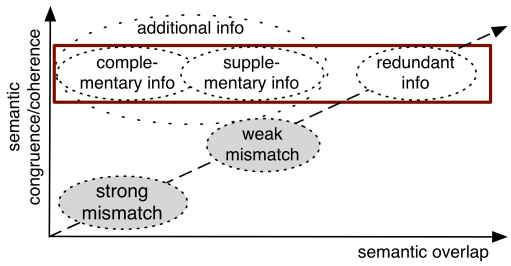

Figure 1: Two-dimensional space of semantic coordination.

From our point of view, the appearances and the understanding of the speech-gesture semantic relationship has a lot more depth to it than sketched so far. In figure 1, we propose a two dimensional space that separates the semantic overlap from the semantic congruence/coherence of speech and gesture. A gesture can convey complementary (*different but necessary*), supplementary (*additional*), or redundant (*corresponding, matching*) information in relation to speech (and vice versa). While this level of semantic overlap has been studied throughout (box in figure 1), it implicitly assumes a high level of coherence between speech and gesture meaning (in the sense of being integrable into a coherent unified interpretation). This congruence, we argue, makes for a second dimension. If a gesture is produced or received with neither semantic overlap nor congruence with speech meaning, we define this as a strong **semantic mismatch**, or hereafter just mismatch. A weaker mismatch is produced or received, if moderate overlap and intermediary congruence between speech and gesture meaning is given.

With this in mind, we define a continuum of mismatches between speech and gesture (dashed arrow in figure 1). In figure 2, three examples along the

continuum are illlustrated in more detail, depending on whether the concepts expressed in speech and gesture are totally different, whether they are derived from the same concept field or whether the concepts are the same. An incongruent speech-gesture pair is a strong mismatch, whereas there are weaker forms up to redundant information in either speech or gesture. Examples of produced or received messages are "high", "wide" and "round", and corresponding gestures which can be used interchangable (cf. figure 2).

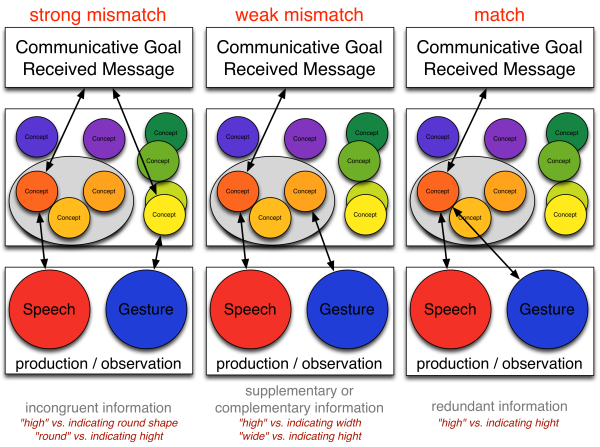

Figure 2: Concepts linked to speech and gesture.

On the basis of this theoretical discussion, how can we explore this phenomenon empirically? The research question on mismatch *production* is, to what extent are speech-gesture mismatches actually produced in natural communication and, since they are hardly found, when and how do they occur in artificial settings? The research question tackling the *comprehension* part of speech-gesture mismatches is, how do subjects cope with conflicting information of the two modalities speech and gesture, do they sometimes interpret a third meaning and are there different levels of impact in each modality?

In preliminary work, we investigated mismatches in natural language production and comprehension. In our first pilot study, the focus laid on mismatches in *production* and we learned that it is difficult to elicit speech-gesture mismatches from adult subjects. The subjects were shown pictures with optical illusions and had to describe the error in the picture either from memory or while looking at it (hands-free), which has been taped on video. We did neither control for the time a mismatch occurs nor for the type, we just created a condition of high cognitive load on the subject. The result was that an adult subject rather interrupts herself during the description process than accepting a semantic mismatch.

In our second pilot study we concentrated on *comprehension* mismatches, similar to Habets et al. (2011). We were able to confirm the tendency of a third meaning emerging from a speech-gesture mismatch (cf. McGurk and MacDonald, 1976) in some cases. In the experiment, we combined conventional gestures like clear pointing gestures up to vague open-arm gestures with spoken words like 'there' and 'everything'/'no clue' and jumbled them. The 64 video snippets (no filler gestures and words) where rather unnatural as they consisted of a single word aligned to one gesture. We decided for the gesture and the word not to appear in some context, since this may influence the subjects interpretation of the related meaning and we were hoping for yet a different meaning than the word and gesture meanings on their own. The results of a subsequently completed questionnaire showed a notable visual impact on the subjects, which may be due to the poor audio quality of the video or to the fact that the visual modality, being quiet dominant, acts as a modifier to the spoken words.

In order to investigate these research questions further, we are about to conduct further experiments. In a first experiment, subjects may have to determine the meaning of certain gestures and certain words independently. We expect clear, vague, and ambiguous meanings in both speech and gesture. Furthermore, we expect some gestures to act as a modifier to the spoken words. Subsequently, we will cluster similar meanings. These clusters will again be checked for their combined meaning by another set of subjects. We considered using predefined gesture lexicons like the "Berliner Lexikon der Alltagsgesten" (BLAG) (Posner, Noll, Krüger, & Serenari, 1999), however, the gesture performance is only shown imprecisely on pictures. Interestingly, the gesture meanings are about the same we have investigated so far. In a second experiment, we may use these predefined gesture and word meanings to conduct the perception experiment again with more attention to detail.

References

Alibali, M. W., & Goldin-Meadow, S. (1993). Modeling Learning Using Evidence from Speech and Gesture. In *Proceedings of the Annual Conference of the Cognitive Science Society.*

Habets, B., Kita, S., Shao, Z., Özyurek, A., & Hagoort, P. (2011). The Role of Synchrony and Ambiguity in Speech–Gesture Integration during Comprehension. *J Cogn Neuroscience*, *23*(8), 1845–1854.

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two Sides of the Same Coin Speech and Gesture Mutually Interact to Enhance Comprehension. *Psychological Science*, *21*(2), 260–267.

McGurk, H., & MacDonald, J. (1976). Hearing Lips and Seeing Voices.

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

Posner, R., Noll, T., Krüger, R., & Serenari, M. (1999). Berliner Lexikon der Alltagsgesten. *Berlin.*

Willems, R. M., Özyürek, A., & Hagoort, P. (2007). When Language Meets Action: The Neural Integration of Gesture and Speech. *Cerebral Cortex*, *17*(10), 2322–2333.

Investigating non-verbal behaviors conveying interpersonal stances

Mathieu Chollet[1] and Magalie Ochs[2] and Catherine Pelachaud[2]

[1]Institut Mines-Telecom ; Telecom Paristech , mchollet@telecom-paristech.fr
[2]Institut Mines-Telecom ; Telecom Paristech ; CNRS-LTCI {mochs, cpelachau}@telecom-paristech.fr

Abstract. The study of the expression of affects and their expression by Embodied Conversational Agents is complex. This is because affects are expressed by non-verbal behaviors on a variety of different modalities, and that these behaviors are influenced by the context of the interaction and other interactants' behaviors. To overcome these challenges, we present a multi-layer framework and apply it to the study interpersonal stance dynamics. For this purpose, we built a corpus of non-verbal behavior annotations and interpersonal stance traces for job interviews.

Introduction - Embodied Conversational Agents (ECAs) are increasingly used in training and serious games. In the TARDIS project[1], we aim to develop an ECA that acts as a virtual recruiter to train youngsters to improve their social skills. Such a virtual recruiter should be able to convey different interpersonal stances, "*spontaneous or strategically employed affective styles that colour interpersonal exchanges* [1]": our goal is to find out how interpersonal stance is expressed through non-verbal behavior, and to implement the expression of interpersonal stance in an ECA. As a representation for interpersonal stance, we use Argyle's attitude dimensions [2], friendliness (also called warmth or affiliation) and dominance (also called agency).

Challenges in non-verbal behavior interpretation - A challenge when interpreting non-verbal behavior is that every non-verbal signal can be interpreted with different perspectives: for instance, a smile is a sign of friendliness [3]; however, a smile followed by a gaze and head aversion conveys embarassment [4]. Non-verbal signals of a person in an interaction should also be put in perspective to non-verbal signals of the other participants of the interaction: an example is posture mimicry, which can convey friendliness [5]. Finally, the global behavior tendencies of a person are important when interpreting their stance [6]. Another challenge is that, as Scherer points out [1], all kinds of affect dont happen in the same span of time. For instance, emotions have a very short duration, and to assess a persons emotion, one should only look at this persons very recent displays of emotion in their non verbal behavior. For moods, one has to look at a persons non-verbal behavior on a longer time span. It might get even longer to get a good sense of someones interpersonal stance. These different perspectives and time spans have seldom been studied together, and this motivates the use of multimodal corpora in order to analyze their different impact in a systematic fashion. To handle these challenges, we introduced a multi-layer model and a multimodal corpus we collected to study it. This model will be used to drive a computational model of interpersonal stance expression for an ECA.

Interpretation of signals using a multi-layer approach - In [7], we defined a multi-layer model to encompass the different non-verbal behavior interpretation perspectives. The *Signal* layer looks at one signal, its characteristics and their immediate interpretation. In the *Sentence* layer, we analyze the sequence of signals happening in a dialogue turn. The *Topic* layer focuses on the

[1] http://tardis.lip6.fr/

behavior patterns and tendencies happening in parts of the interactions where a topic is discussed (e.g. greetings, discussing job offer, saying goodbye...). Finally, the *Interaction* layer encompasses the whole interaction and looks at global behavior tendencies. These different layers allow to interpret interactants' interpersonal stances at every instant of the interaction, taking into account their behavior, their reactions to other interactants' behaviors, and their global behavior tendencies.

Multimodal corpus of interpersonal stance expression - In order to study how recruiters express interpersonal stance, we annotated three videos of job interview enactments, for a total of slighty more than 50 minutes. We consider full body non-verbal behavior, turn-taking, task and interpersonal stance [8]. Non-verbal behavior and interactional state annotations consist of labeled time intervals, while interpersonal stance was annotated as a trace. Trace data is prone to certain issues, such as scaling and reliability issues, thus we chose to follow Cowie's approach [9] by switching from an absolute perspective (i.e. what is the value of interpersonal stance at time t) to a relative perspective (i.e. is the interpersonal stance staying stable or is it changing at time t). This process allows to segment where the rater's perception of interpersonal stance varies or stays stable.

Investigating the perception of non-verbal behavior produced by an ECA - Our future work consists of investigating how behaviors contribute to the expression of interpersonal stance at every layer level, and how the four layers contribute to the perception of the interpersonal stance at every instant. For this, we intend to use our corpus as training data to learn the parameters of the model. The multi-layer model will then be implemented into our ECA as a behavior planning module which will receive target interpersonal stances as an input, and then compute which behaviors to express in order to reach these targets.

Conclusion - The complexity of non-verbal behavior expression and interpersonal stance perception in specific contexts motivates the use of a framework that considers all perspectives of behavior interpretation, and of a multimodal corpus as ground truth. We have proposed a multi-layer framework to handle the complexity of interpersonal stance expression, and we annotated videos of job interview enactments. Our future work consists of tuning our model using the multimodal corpus, and to then implement it as a behavior planner in our ECA platform before validation.

Acknowledgment - This research has been partially supported by the European Community Seventh Framework Program (FP7/2007-2013), under grant agreement no. 288578 (TARDIS).

References

1. Scherer, K.R.: What are emotions? and how can they be measured? Social Science Information **44** (2005) 695–729
2. Argyle, M.: Bodily Communication. University paperbacks. Methuen (1988)
3. Burgoon, J.K., Buller, D.B., Hale, J.L., de Turck, M.A.: Relational Messages Associated with Nonverbal Behaviors. Human Communication Research **10**(3) (1984) 351–378
4. Keltner, D.: Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. Journal of Personality and Social Psychology **68** (1995) 441–454
5. LaFrance, M.: Posture mirroring and rapport. In Davis, M., ed.: Interaction Rhythms: Periodicity in Communicative Behavior, New York: Human Sciences Press (1982) 279–299
6. Escalera, S., Pujol, O., Radeva, P., Vitria, J., Anguera, M.: Automatic detection of dominance and expected interest. EURASIP Journal on Advances in Signal Processing **2010**(1) (2010) 12
7. Chollet, M., Ochs, M., Pelachaud, C.: Interpersonal stance recognition using non-verbal signals on several time windows. In: Workshop Affect, Compagnon Artificiel, Interaction. (November 2012)
8. Chollet, M., Ochs, M., Pelachaud, C.: A multimodal corpus approach to the design of virtual recruiters. In: Workshop Multimodal Corpora, Intelligent Virtual Agents. (2013) 36–41
9. Cowie, R., Cox, C., Martin, J.C., Batliner, A., Heylen, D., Karpouzis, K.: Issues in Data Labelling. Springer-Verlag Berlin Heidelberg (2011)

Classifying the form of iconic hand gestures from the linguistic categorization of co-occurring verbs

Magdalena Lis, Costanza Navarretta

Scope

This paper deals with the relation between speech and form of co-occurring iconic hand gestures - a research issue relevant for understanding human communication and for modelling multimodal behaviours in applied systems. To investigate this relation, we apply supervised machine learning to a Polish multimodal corpus. In the corpus, speech is annotated with information extracted from a linguistic taxonomy (plWordNet) and form features of gestures are coded.

Background

In face-to-face interaction, verbal descriptions of events tend to be accompanied by iconic hand gestures (McNeill 1992). A single event can be gesturally represented in multiple ways, for instance from different perspectives: in character viewpoint (C-vpt), an event is shown from the perspective of an agent, in observer viewpoint (O-vpt), the gesturer sees the event as an observer (McNeill 1992). Furthermore, the gesturer has to choose which hand will perform the gesture (Handedness), the configuration of palm and fingers (Handshape), whether the gesture will be static or dynamic and the motion pattern single or repeated (Iteration), the shape of the motion (Movement) and the plane on which it will be performed (Direction), etc. Knowledge on factors influencing these choices is important for understanding cognitive-semiotic processes in face-to-face communication and can inform models of gesture generation.

Parrill (2010) has demonstrated that the choice of viewpoint is influenced by the events' structure. She has shown that events involving trajectories as a more salient element elicit O-vpt gesture, while events focused on handling evoke C-vpt gestures. Duncan (2002) has studied the relationship between handedness and verbal aspect in Chinese and English data. On the level of speech, events are expressed by verbs; verbal aspect concerns the events' relation to time. Duncan has found that symmetrical bi-handed gesture more often occur with perfective verbs than with imperfective ones; the latter are mostly accompanied by two handed non-symmetrical gestures. Parrill and colleagues (2013) have investigated the relationship between verbal aspect and gesture iteration. They have suggested that descriptions in progressive aspect are more often accompanied by iterated gestures, but only when the events were originally presented to the speakers in that aspect. Becker and colleagues (2011) have conducted a study on the relation between gesture and verb's Aktionsart. Aktionsart is a lexicalization of various 'manners of action' according to distinctions between static and dynamic, telic and atelic, durative and punctual (Vendler 1967). They have found differences in the speech-gesture temporal relationship between different Aktionsart categories.

Lis (2012) has proposed to investigate the relationship between gestures and verbs by using information extracted from a wordnet. Wordnets are electronic taxonomies, which group lexical units into sets of synonyms and link them via, among others, the relation of hyponymy, i.e. 'lS A SUBTYPE OF' relation. Hyponymy encodings in plWordNet contain information on verbs' aspect, Aktionsart and semantic domain. Based on the domains, Lis distinguished between different Semantic subtypes of events (inspired by Parrill's distinction in event structure) and she showed their correlation with gestural representation. The present study further builds up on this framework. Differing from preceding studies of the relation between verb and form of the co-occurring hand gesture, we test the hypotheses by applying supervised machine learning on information extracted from the corpus. In doing this, we follow the strategy proposed by inter alia Jokinen et al. (2008) who apply machine learning algorithms to multimodal annotated data in order to discover dependencies between shape and functions of head movements and facial expressions.

Data

The corpus consists of manually annotated audio-video recordings of narrations by 5 male and 5 female adult native Polish speakers. Speech is transcribed with word stamps and gestures are coded for Viewpoint, Handedness, Handshape, lteration, Movement and Direction in ANVIL tool developed by Kipp (2004). The sample used in this study consists of 269 gestures. Inter-coder agreement (κ scores) for the annotated attributes range from substantial (0.7) to almost perfect (0.95) agreement. Annotations for verbs are extracted

from the Polish WordNet, plWordNet 1.5, [i] and consist of information about Semantic subtype, Aspect and Aktionsart. Details on annotation (attributes, values, inter-coder agreement) can be found in (Lis 2012).

Experiments and results

Classification experiments are performed in WEKA (Witten and Frank 2005) applying a support vector classifier, SMO, to various combinations of features. Ten-fold cross-validation is used for testing and the baseline is given by the results obtained applying a ZeroR classifier to the data. ZeroR always chooses the most common nominal class.

The first experiment aims at predicting viewpoint from the annotations of the verb obtained via wordnet. The results confirm that there is a strong correlation between Viewpoint and Semantic subtype (F-score improvement with respect to the baseline: 0.35). We also find a correlation between Viewpoint and Aktionsart (F-score improvement with respect to the baseline: 0.16).

Next, we classify Viewpoint from gestures' form features. The results demonstrate a strong correlation between the form of a gesture and the gesturer's viewpoint. Handshape and Handedness are the features most strongly correlated to Viewpoint.

In the last experiment, we test whether it is possible to predict the form of the gesture from the wordnet annotations of the verb. The results indicate that Semantic subtype and Aktionsart influence Handshape and Handedness, while Aspect does not affect the classification. All features affect classification of the Direction but none contributes to the prediction of Iteration and Movement in gesture.

Discussion

The results of the classification experiments confirm the hypothesis that the semantic subtype of event, as identified using wordnet's categorization of verbs, is related to the form of co-occurring iconic hand gestures (Lis 2012). More specifically it is related to handshape, handedness and viewpoint. We also found that Aktionsart correlates with those features, but to a smaller degree. The observation that aspect is related to handedness and iteration is not reflected in our data. Since such a relation was noticed in other languages (Duncan 2002, Parrill et al. 2013), it needs to be tested whether this is a case of cross-linguistic variation.

The results also indicate that it is possible to some extent to predict the form of hand gestures from the linguistic categorization of the co-occurring lexical units. This is a promising result for gesture modelling. We hope that after further development, the framework presented in the paper may enable proposing a set of preference rules for gesture production.

Our results are interesting, but since our data is not large, the experiments should also be applied to more data and different types of interactions. We have recently started investigating gestures referring to entities of other semantic (sub)types and we are also looking at other articulators, such as head movements and posture changes.

References

Becker, R., Cienki, A., Bennett, A., Cudina, C., Debras, C, Fleischer, Z., M. Haaheim, T. Mueller, K. Stec, and A. Zarcone. Aktionsarten, speech and gesture. In Proceedings of GESPIN (2011).

Duncan, S. Gesture, verb aspect, and the nature of iconic imagery in natural discourse. Gesture, 2(2):183-206 (2002).

Fellbaum, Ch. (ed.), WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA, (1998).

Jokinen, K., C. Navarretta and P.Paggio. Distinguishing the communicative functions of gestures. In Proceedings of the 5th Joint Workshop on MLMI (2008).

Kendon, A. Gesture: Visible Action As Utterance. Cambridge University Press, Cambridge (2004).

Kipp, M. Gesture Generation by Imitation – From Human Behavior to Computer Character Animation. Boca Raton, Florida (2004).

Lis, M. Influencing gestural representation of eventualities: Insights from ontology. In Proceedings of ICMI (2012).

McNeill, D. Hand and Mind: What Gestures Reveal About Thought. University of Chicago Press, Chicago (1992).

Parrill, F. Viewpoint in speech and gesture integration: Linguistic structure, discourse structure, and event structure. Language and Cognitive Processes, 25(5):650-668 (2010).

Parrill, F., Bergen, B. and P. Lichtenstein. Grammatical aspect, gesture, and conceptualization: Using co-speech gesture to reveal event representations. In Cognitive Linguistics, 24(1): 135-158 (2013).

Vendler, Z. Linguistics in Philosophy. Cornell University Press, Ithaca, NY (1967).

Witten, J. and Frank, E. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco (2005).

[i] Polish wordnet, plWordNet1.5., http://plwordnet.pwr.wroc.pl/wordnet/

Up, down, in & out:
Following the Path in speech and gesture in Danish and Italian

Bjørn Wessel-Tolvig
Centre for Language Technology, University of Copenhagen
bwt@hum.ku.dk

Languages vary crucially in how lexical and syntactic elements are mapped onto linguistic form. This variation is particularly evident within the semantic domain of motion (Talmy 1985, Slobin 2004). The variations in lexicalization are also reflected in the content of co-speech gestures (Kita 2009). Research shows how speakers of satellite framed languages express Path of motion in verb particles (e.g. up, down, in, out) in tight clausal packaging with Manner of motion in the verb root. This pattern is often accompanied by *one* gesture representing either manner or path information, or conflating the two components into a single gesture. Speakers of verb framed languages, where Path of motion is frequently expressed in verb roots (e.g. ascend, descend, enter, exit) and Manner of motion in subordinate clauses (e.g. gerunds), are more likely to produce one gesture *per clause* reflecting the division of Manner and Path in speech (Hickmann et al. 2011, Kita & Özyürek 2003, Stam 2006).

Danish (satellite framed) and Italian (verb framed) should therefore express motion in speech and gesture in quite different ways. Danes typically express Path through an elaborate system of satellites (e.g. *op*, *ned*, up, down) whereas Italians, although verb framed, have multiple possibilities for expressing Path in either verb roots (e.g. *salire*, *scendere*, ascend, descend) or in verb particle constructions (e.g. *andare su*, go up) (Folli 2008, Iacobini & Masini 2006). This variety of possibilities in a verb framed language show properties of a 'split system' typology (Talmy 2000). Very little attention has been paid to Danish and Italian in speech-gesture patterns. Rossini (2005) found that Italians can, and do, express Path in satellites to the verb and synchronize gestures with either the lexical item or verb + satellite, but fails to mention the semantic content of the co-expressive gesture. Cavicchio & Kita (2013a, 2013b) found significant differences in gesture rate and gesture space between Italian and English speakers narrating motion, but do not explain whether these differences are due to habitual cultural differences alone or to linguistic packaging of lexical items as well.

This study investigates how different strategies for expressing Path in Danish and Italian influence the content of co-speech gestures. Since speech and gesture are increasingly seen as integrated in production we expect inter-typological and intra-typological variations in lexicalization are reflected in gesture patterns.

Ten Danish and eleven Italian native speakers narrated eight video animations each to a confederate listener. All animations contained motion events. Four scenes depicted an animate figure ascending or descending a hill (The Tomato Man Movies; Özyürek, Kita & Allen 2001) and the other four scenes illustrated an animate figure entering and exiting a house in various ways (Wessel-Tolvig 2013). The participants were all video recorded for speech-gesture analysis.

Results show that on-line choice of lexicalization pattern is reflected in gestures. All Danish speakers expressed Path in satellites and Manner in verb roots, and produced *one* gesture representing either path (46.9%) or manner-path conflated (50%) information. Italian shows evidence for 'split system' strategies by expressing Path either in verb roots (65.9%) or in verb particles (34.1%). Italian gesture data illustrate the influence of lexicalization patterns on gesture content: 61.5% of all path-only gestures were used when Path was expressed in verb roots and Manner of motion omitted (verb framed). On the other hand 52.9% of all manner-path conflated gestures were produced in constructions where Path was expressed in the verb particle and Manner in the verb root (satellite framed).

The results for Italian expressions of speech and gesture show a higher tendency for conflating gestures when Path is expressed in verb particle constructions opposed to producing path only gestures with Path-only expressions in the verb root. The results support the idea that the language you speak, and particularly the way Path of motion is mapped onto linguistic form in that language, influence the content of co-speech iconic gestures (Kita et al. 2007).

References:

Cavicchio & Kita (2013a). Bilinguals Switch Gesture Production Parameters when they Switch Languages. Proceedings of the Tilburg Gesture Research Meeting (TiGeR 2013). June 2013, Tilburg, The Netherlands.

Cavicchio & Kita (2003b). English/Italian Bilinguals Switch Gesture Parameters when they Switch Languages. Proceedings of Annual meeting of the cognitive science society (CogSci 2013). July/August 2013, Berlin, Germany

Folli, R. (2008). Complex PPs in Italian. In: Asbury, A., Dotlacil, J., Gehrke, B. & Nouwen, R. (Eds.). Syntax and Semantics of Spatial P. 197-221. John Benjamins Publishing Company.

Hickmann, M., Hendriks, H., & Gullberg, M. (2011). Developmental perspectives on the expression of motion in speech and gesture: A comparison of French and English. Language, Interaction and Acquisition/Language, Interaction et Acquisition, 2(1), 129-156.

Iacobini, C. & Masini, F. (2006). The emergence of verb-particle constructions in Italian: locative and actional meanings. Morphology 16(2), 155-188.

Kita, S. (2009). Cross-cultural variation of speech-accompanying gesture: A review. Language and Cognitive Processes, 24(2), 145-167.

Kita, S., & Ozyurek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal? Evidence for an interface representation of spatial thinking and speaking. Journal of Memory and Language, 48(1), 16-32.

Kita, S., Ozyurek, A., Allen, S., Brown, A., Furman, R., & Ishizuka, T. (2007). Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production. Language and Cognitive Processes, 22(8), 1212-1236.

Özyürek, A., Kita, S., & Allen, S. (2001). Tomato Man movies: Stimulus kit designed to elicit manner, path and causal constructions in motion events with regard to speech and gestures. Nijmegen, The Netherlands: Max Planck Institute for Psycholinguistics, Language and Cognition group

Slobin, D. I. (2004). The many ways to search for a frog. In: S. Strömqvist & L. Verhoeven (Eds), Relating events in narrative: Vol. 2. Typological and contextual perspectives. 219-257. Hillsdale, NJ:Erlbaum

Rossini, N. (2005). Phrasal verbs or words? Towards the analysis of gesture and prosody as indexes of lexicalization. On-line Proceedings of the 2nd ISGS Conference "Interacting Bodies". Lyon, France.

Stam, G. (2006). Thinking for speaking about motion: L1 and L2 speech and gesture. IRAL 44(2), 143-169.

Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In: Shopen, T. (Ed.). Grammatical categories and the lexicon. Vol III. 57–149. Cambridge: Cambridge University Press.

Talmy, L. (2000). Toward a Cognitive Semantics: Typology and process in concept structuring. Vol. II. Cambridge: MIT Press.

Wessel-Tolvig, B. (2013). Boundary Ball: An animated stimulus designed to elicit motion with boundary crossing situations. University of Copenhagen

Referent type and its verbal and gestural representation:
A test on English multimodal corpus and WordNet® 3.1.

Magdalena Lis[1], Fey Parrill[2]

[1]*Centre for Language Technology, University of Copenhagen, Copenhagen, Denmark*
[2]*Department of Cognitive Science, Case Western Reserve University, Cleveland, USA*

Aim

The present paper builds upon a framework that employs a linguistic taxonomy to investigate factors influencing how referents are represented in gesture. The framework was developed on Polish multimodal corpus (Lis 2012a). We present a pilot study on applying the framework to English. We investigate whether the linguistic taxonomy can predict gestural viewpoint and representation technique in English speech and gesture data.

Introduction

Referring to entities is one of the crucial elements of communication. In face-to-face interaction, speakers use not only speech but also gestures to make reference to an object or a concept.
One and the same referent can be represented gesturally in various ways. For example, speakers can adopt different viewpoints. With character viewpoint (C-vpt) a referent is depicted from the perspective of an agent: The speaker uses her body as the referent's body. With observer viewpoint (O-vpt), a referent is presented from a perspective of an external observer. Dual viewpoint (D-vpt) combines the two perspectives simulatenously (McNeill 1992). Another way of conceptualizing gestural representation is the notion of representation technique. With Depicting, the hands trace an outline or sculpt a shape, with Acting, they mime an action (Müller 1998). O-vpt is likely to be correlated with the Depicting Technique, while C-vpt and Acting are likely to be correlated, but this relationship has not been demonstrated.
What factors determine a speaker's choice of gestural representation? Understanding the factors influencing this choice can provide insight into the cognitive-semiotic processes in human communication and can inform models of gesture production.

Background

Poggi (2008) has suggested that gestural representation differs depending on 'the type of semantic entity [the referent] constitutes.' She distinguished four such types, one of them being events. However, events themselves can be represented in various ways, e.g. from different viewpoints (McNeill 1992). Parrill (2010) has demonstrated a correlation between the viewpoint a speaker used in gesture and the structure of an event. She proposed that events with a trajectory as the more salient element elicit O-vpt gesture, while events in which handling is more prominent tend to evoke C-vpt gestures.
Lis (2012a) used wordnet to combine and formalise Poggi and Parrill's insights. Wordnet is an electronic semantic-lexical database. It constitutes a large taxonomy, grouping lexemes into sets of synonyms and encoding relations between them. In wordnet, domains denote segments of reality symbolized by a set of related lexemes, which all have a common semantic property (Fellbaum 1998). In Lis' framework, verbs are annotated with information derived from wordnet domains to provide an identification of event subtype.
Lis looked at the verbs that speakers used to describe events in a multimodal corpus of Polish, using Polish wordnet (plWordNet 1.5). She found a correlation between event subtype and gestural viewpoint (2012b) and event subtype and representation technique (2013). 'Trajectory' events were accompanied by O-vpt, Depicting gestures, while 'Handling' events were accompanied by C-vpt, Acting gestures.
Lis suggested that the framework can be extended to other types of entities than events and can be applied cross-linguistically. Wordnets exist for numerous languages, and there are both similarities and differences in the taxonomies due to methodological decisions and differences between languages themselves. In the

present paper, we test Lis' claim on English. We apply the framework to English WordNet® 3.1 and to an English multimodal corpus of speech and gesture.

Data

The corpus consists of audio-videorecordings of 6 native American English speakers (3 women, 3 men) performing narration tasks. Speech was transcribed and gestures representing events were coded for viewpoint (Parrill 2010). Gestures were also annotated with representation techniques. Event subtypes were semi-automatically assigned based on WordNet® 3.1. The sample used in the current study consists of 91 hand gestures and con-current verbs.

Results

D-vpt was very infrequent in the data (0.01%). In all other instances Depicting Technique overlapped with O-vpt, and Acting Technique with C-vpt.

We found a significant relationship between viewpoint and representation technique in gesture and event subtype as identified using WordNet® 3.1. Handling events were likely to be accompanied by C-vpt and Acting Technique gestures, Trajectory events were accompanied mainly by O-vpt, Depicting gestures (χ^2= 66.92, df=1, p<.0001). The relationship found on Polish data in the previous study is, thus, confirmed on a language from a different language group, and on a different corpus.

Discussion

These results support the viability of the framework (i.e. application of wordnets for investigation of speech-gesture ensembles) and for its cross-linguistic application. Using wordnets as an external source of annotation ensures coding reliability and enables automatic assignment of values. Applied cross-linguistically, it can help to identify universal versus language-specific patterns in multimodal communication. In our future work, we plan to investigate the relationship between viewpoint and technique on the one hand and concrete aspects of gesture form (handedness, handshape, repetition) on the other. We hope that the results will be a step towards proposing a set of preference rules for modeling gesture production.

References

Fellbaum, Ch. (ed.), WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA, (1998).
Kendon, A. Gesture: Visible Action As Utterance. Cambridge University Press, Cambridge (2004).
Lis, M. Annotation scheme for multimodal communication: Employing plWordNet 1.5. In: Proceedings of the Formal and Computational Approaches to Multimodal Communication (2012a).
Lis, M. Influencing gestural representation of events: Insights from ontology. In: Proceedings of ICMI (2012b).
Lis, M. Gestural representation in the domain of animates' physical appearance. In: Proceedings of the Tilburg Gesture Research Meeting (2013).
McNeill, D. Hand and Mind: What Gestures Reveal About Thought. University of Chicago Press, Chicago (1992).
Müller, C. (1998). Redebegleitende gesten. Kulturgeschichte. Theorie. Sprachvergleich (Vol. 1). Berlin: Verlag Spitz.
Parrill, F. Viewpoint in speech and gesture integration: Linguistic structure, discourse structure, and event structure. Language and Cognitive Processes, 25(5):650-668 (2010).
Poggi, I. Iconicity in different types of gestures. Gesture, 8(1), 45-61 (2008).

Deep Learning for Multimodal Feature Extraction

Héctor P. Martínez Georgios N. Yannakakis
Institute of Digital Games
University of Malta
{hector.p.martinez, georgios.yannakakis}@um.edu.mt

I. Introduction

A large amount of information is generated while interacting with computers. This information is by nature temporal and multimodal as it spans for a period of time and it is spread over several dissimilar modalities such as speech, gestures, physiology and mouse clicks. The analysis and modeling of the interrelation among the different modalities and along time is central to the advancement of human-computer communication, for instance to create affect-sensitive computers [1]. However, the immense complexity arising from long temporal dimensions and particularities of each modality has promoted analyses based on reduced versions of each modality. This process, namely *feature extraction*, generally involves a domain expert who chooses a set of statistical features (e.g. average values) in an attempt to capture all the relevant qualitative characteristics of the modalities. While in single-modality datasets some information is inevitably lost, this loss is increased in multimodal data as the included signals are often reduced independently (e.g. [2], [3] among others). In particular, this process hinders the low-level interactions among modalities.

This paper introduces *deep learning* [4], [5] approaches for the automatic extraction of multimodal features. These approaches can produce features that potentially reveal relevant dependencies among multiple modalities which could be otherwise missed by expert-designed features. Deep-learned features can not only offer new views for the analysis of multimodal interactions but also yield more accurate computational models (e.g. when used as inputs to predictors of user's psychological state, as in [1], [5]). With the proposed approach, the different signals can be combined before building a model (data-level fusion [6]) achieving deeper fusion than feature-level or decision-level fusion that facilitates a fine-grained interpretation of the interrelation among modalities and its effect on the target output.

Data-level fusion of dissimilar modalities has been achieved utilizing frequent sequence mining [1]; however, this approach requires transforming *continuous signals* into sequences of discrete events (i.e. *nominal signals*). For instance, in [1] a *skin conductance* signal (SC) is converted into a sequence of significant increments and decrements of the signal. This transformation requires ad-hoc preprocessing which discards part of the information present in the raw signal. Deep learning approaches on the other hand are defined for continuous signals and therefore do not require ad-hoc preprocessing.

Deep learning has been already applied to a number of human-computer interaction tasks such as facial expressions recognition [7], speech recognition [4] and modeling of physiological signals [5]; however, it has not been employed — to the best of the authors' knowledge — to fuse signals from different modalities. We investigate a particular deep learning algorithm to create the multimodal features, namely *convolutional neural networks* [4] (CNNs). CNNs present advantages over other deep structures as they reduce the signals using a hierarchy of simple local (translation-invariant) features, which enable a simple analysis of the multimodal feature extraction process and reduce *segmentation problems* (e.g. defining the exact chunks of the input signals that define a feature).

The modality combination investigated in this paper includes physiological signals and context information as found in the game dataset *Maze-Ball* [2]. This dataset contains skin conductance and *blood volume pulse* (BVP) recordings along with *game context information* such as the sequence of moments when the player is hit. For this particular combination, ad-hoc multimodal features are practically non-existent in the literature. The relations between physiology and context is often studied through feature-level fusions [2], [3]. Some studies have scarcely explored the reactions of physiology to few selected sudden events (e.g. [8]). However, we expect that an automatic method to fuse physiology and context at data-level can outperform hand-crafted analysis which are limited to the creativity of the expert. In the following sections, we describe the basic structure and training algorithm for CNNs and the deviations required to deal with multimodal data.

II. Standard Convolutional Networks

CNNs are feed-forward neural networks designed to deal with large input spaces as those seen in image classification tasks. CNNs are constructed by stacking alternatively *convolutional layers* and *pooling layers*. A convolutional layer consists of a number of neurons that process sequentially consecutive patches of the input signal, i.e. this layer convolves a set of neurons along the temporal dimension of the input signal. Each neuron defines one local feature which is extracted at every position of the input signal; the resulting values create a new signal referred to as *feature map*. A pooling layer reduces the dimensionality of the feature maps generated by a convolutional layer. It applies a simple statistical function (e.g. average or maximum value) to non-overlapping patches of the feature maps. By stacking several convolutional and pooling layers, CNNs can effectively reduce the input signals

to a small set of features. The experimenter must define the topology of the network including the number of neurons and the number of inputs for each neuron in each convolutional layer, and the number of inputs (window) and function used by the pooling layer. Once those parameters are fixed, the neurons can be automatically trained in order to minimize the loss of information in the feature extraction process.

We train each convolutional layer using *auto-encoders* [5]. This approach consists of feeding the outputs of the convolutional layer into a *decoder* that reconstructs the original input signal; by means of gradient-descent, the weights of each neuron are adjusted iteratively to achieve a minimal reconstruction error, i.e. a minimal discrepancy between the signal reconstructed by the decoder and the original input signal [5].

III. Multimodal Convolutional Networks

CNNs have been extensively and successfully applied to image classification tasks, which can be considered multimodal tasks when 3 color channels are present (one modality per color). However, real multimodal tasks present a number of properties which create difficulties to the application of standard CNNs:

- **Different sampling rates** the neurons of a convolutional layer scan the input signals sample-by-sample along time; when the signals present different number of samples per unit of time, this is not applicable as the scanning process requires a different pace for each modality.
- **Strongly entangled components** a convolutional layer attempts to identify the main (entangled) components of its input signal; when several complex modalities are presented simultaneously, the number of dissimilar relevant components can be humongous and the relations among modalities hard to discern.
- **Misaligned modalities** CNNs generate local features that find particular components at any point in time; however, this property only holds along time and not across modalities. Therefore, standard CNNs cannot capture multimodal patterns when a variable lag exists across modalities. This issue can be regarded as the *correspondence problem* [9], as the lag hinders the time correlation from one modality to the others.

We propose a number of modifications to the standard CNN definition in order to overcome these challenges.

- The difference in sampling rates can be easily overcome by feeding the different modalities at different layers, i.e. the signals with a higher sampling rate are processed by a convolutional and pooling layers which transform them into feature maps with the same time resolution as the signals with lower sampling rates. Alternatively, one could simply undersample all the signals to the same sampling rate which can be seen as a form of pooling layer. However, using also a convolutional layer can potentially reduce the information loss because consecutive outputs of a convolutional layer are very similar, which produces a small loss of information in the pooling process.
- A solution for the strong entanglement across modalities consists of, first, processing each signal individually with a convolutional layer and then, fusing the resulting feature maps in the following layers. Once the components of each signal have been disentangled (in the individual convolution), the relations across modalities can be more easily captured (in the convolution over the feature maps).
- While a pooling layer might not be necessary after the individual convolutional layer, it can serve to counteract signal misalignments. Note that the feature maps generated by a pooling layer conform representations of the input signal at lower resolution; if the reduction is greater than the lag among modalities, then the fusion of the resulting pooled feature maps is, to a large degree, unaffected by the lag.

In all, we propose a simple CNN architecture to deal with the particularities of multimodal data. First, each modality is processed by one convolutional and one pooling layer that extracts its main components and reduces its temporal resolution. Then, the resulting feature maps are fused at different layers where the time resolution is similar. The training algorithm would not require any modification as only the global structure of the CNN has been altered while keeping the standard definition of each of the components (i.e. convolutional and pooling layers).

The depth of the fusion and the complexity of the extracted features can be directly regulated through the parameters of the CNN. The number of neurons and their inputs limit the components that can be extracted from each modality while the window of the pooling layers has a direct impact to the depth of the fusion and the size of the network. Note that small windows generate a timid reduction of the time resolution, allowing for a deep fusion of the modalities while requiring a large number of layers to generate a small number of features; on the other hand, a large window reduces the resolution aggressively, resulting in a lower number of layers but forcing the fusion of modalities at a low time resolution.

References

[1] H. P. Martínez and G. N. Yannakakis, "Mining multimodal sequential patterns: a case study on affect detection," in *Proceedings of International Conference on Multimodal Interfaces*. ACM, 2011, pp. 3–10.

[2] G. N. Yannakakis, H. P. Martínez, and A. Jhala, "Towards affective camera control in games," *User Modeling and User-Adapted Interaction*, vol. 20, pp. 313–340, 2010.

[3] S. Mcquiggan, B. Mott, and J. Lester, "Modeling self-efficacy in intelligent tutoring systems: An inductive approach," *User Modeling and User-Adapted Interaction*, vol. 18, no. 1, pp. 81–123, 2008.

[4] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," in *The handbook of brain theory and neural networks*. Cambridge, MA: MIT Press, 1995, vol. 3361.

[5] H. P. Martínez, Y. Bengio, and G. N. Yannakakis, "Learning deep physiological models of affect," *Computational Intelligence Magazine, IEEE*, vol. 9, no. 1, pp. 20–33, 2013.

[6] M. Pantic and L. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE*, 2003.

[7] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza, "Disentangling factors of variation for facial expression recognition," in *European Conference on Computer Vision*, 2012.

[8] N. Ravaja, T. Saari, M. Salminen, J. Laarni, and K. Kallinen, "Phasic emotional reactions to video game events: A psychophysiological investigation," *Media Psychology*, vol. 8, no. 4, pp. 343–367, 2006.

[9] R. K. Srihari, "Computational models for integrating linguistic and visual information: A survey," *Artificial Intelligence Review*, vol. 8, no. 5-6, pp. 349–369, 1994.

Overlaps in Maltese: a comparison between task-oriented and conversational data

Alexandra Vella
University of Malta
alexandra.vella@um.edu.mt

Patrizia Paggio
University of Malta
University of Copenhagen
patrizia.paggio@um.edu.mt

The fact that spontaneous speech does not consist of neatly arranged and separated turns, and that speakers, on the contrary, speak over each other and interrupt each other, has been observed by many (e.g. Schegloff (2000), Cetin and Shriberg (2006), Adda-Decker M. et al. (2008), Campbell et al. (2010)). As shown in these studies, the amount and use of overlap varies depending on a number of parameters, especially the presence of pre-established roles and speakers' familiarity. This paper presents a comparison between two different corpora of spoken Maltese with specific reference to the issue of overlap, thereby testing the hypothesis that overlaps are used to different degrees and for different purposes in different communicative situations. It also provides an analysis of overlaps in a type of situation, first acquaintance dialogues, which has not been studied earlier with respect to this topic.

The first corpus consists of eight Maltese Map Task dialogues which form part of the MalToBI corpus (Vella and Farrugia 2006). The Maltese Map Task design is based on that of the Map Task used for the collection of data in the HCRC Map Task corpus (Anderson et al. 1991). Two participants engage in a communication gap activity. The aim is for the participant in the Leader role to describe the route on the Leader Map to the participant in the Follower role, whose task is to draw the route in accordance with the information provided by the Leader. The Maps are not identical, thus necessitating an element of negotiation. Contrary to what happens in other Map Task collections, the two participants can see each other. The dialogues involve 16 speakers (8 females and 8 males): half the females fulfil the Leader role and the other half the Follower role, and similarly in the case of the male speakers.

The second collection is the multimodal corpus of Maltese MAMCO, which consists of twelve video-recorded first encounter conversations between pairs of Maltese speakers. Twelve speakers participated (6 females and 6 males), each taking part in two different short conversations that took place in a recording studio. The setting and general organisation of the collection replicate those used to develop the Nordic NOMCO corpus (Paggio et al. 2010).

In both corpora the speech was transcribed using Praat (Boersma and Weenink, 2009) and following the guidelines described in Vella et al. 2010. In both types of data, overlaps are defined as temporal segments in which the speakers speak at the same time.

Our hypothesis, tentatively confirmed in a preliminary investigation (Paggio and Vella, 2013), is that overlaps serve different purposes in the two corpora. This assumption is based on the fact that the two corpora differ in at least three important respects: i. the Map Task dialogues are task-oriented, while the MAMCO conversations involve free face-to-face interaction; ii. in the Map Task dialogues there is a clear role distinction between the Leader and the Follower, whereas in MAMCO the participants all have equal status; iii. finally, familiarity is not an issue in the Map Task corpus, while it is an important parameter in MAMCO, since the participants do not know each other and become acquainted during the interaction.

Based on the findings by the authors quoted earlier, as well as the results presented in our previous paper, we would expect a greater degree of overlap in the MAMCO conversations because neither of the speakers has a predetermined leading role. In other words, both have to negotiate the floor, and there is no reason to expect that one of the speakers should overlap more than the other, not, at least, as a consequence of the communication situation. This expectation is

indeed confirmed by the analysis of the data. On the other hand, the main function of overlaps in the Map Task dialogues is to assure the Leader, who gives the instructions, that an instruction has been understood (or the opposite) and to maintain continuity with a view to task completion. Therefore, we expect the Leader mostly to keep the turn at the end of an overlap. .

An interesting question is also whether the use of overlaps changes in the course of the conversation. We know that there are situations in the MapTask dialogues in which the participants have to correct misunderstandings due to the differences between the two maps: we would expect more overlap to occur around these situations. In MAMCO, on the other hand, if familiarity has an effect as we know from the literature, we would expect the two participants to overlap more as the conversation develops. Conversely, it could be that the initial phase, in which they greet each other and attempt to establish a pace for the dialogue, as well as some common ground to base their interaction on, contains more overlap.

These expectations have been verified by extracting the overlaps and carrying out a quantitative comparison across the two corpora. In addition to presenting the results of the quantitative analysis, we will also discuss typical examples to illustrate the different types of function that overlaps have in the two corpora.

References

Adda-Decker M., Barras, C., Adda, G., Paroubek, P., Boula de Mareüil P. and B. Habert. 2008. "Annotation and Analysis of Overlapping Speech in Political Interviews", in *Proceedings of the 6th International Language Resources and Evaluation (LREC '08)*.

Anderson, A. H., Bader, M., Bard, E.G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. and R. Weinert. (1991). "The HCRC Map Task Corpus". *Language and Speech* 34: pp. 351-366.

Boersma, P. and D. Weenink (2009). Praat: doing phonetics by computer (Version 5.1.05) [Computer program].

Campbell, N. and S. Scherer, (2010). "Comparing Measures of Synchrony and Alignment in Dialogue Speech Timing with Respect to Turn-Taking Activity", in *Proceedings of Interspeech*, pp. 2546-2549.

Cetin O. and E.E. Shriberg. (2006). "Overlap in Meetings: ASR Effects and Analysis by Dialog Factors, Speakers, and Collection Site". MLMI06 (3rd Joint Workshop on Multimodal and Related Machine Learning Algorithms), Washington DC.

Paggio, P., J. Allwood, E. Ahlsén, K. Jokinen and C. Navarretta (2010). "The NOMCO Multimodal Nordic Resource - Goals and Characteristics", in Calzolari et al. (eds.) *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC '10)*, pp. 2968–2974, Valletta, Malta.

Paggio, P. and Vella, A. (2013). Overlaps in Maltese: a comparison between Map Task dialogues and multimodal conversational data. In Postproceedings of the 4th Symposium on Multimodal Communication. Gothenburg.

Schegloff, E. A. (2000). "Overlapping Talk and the Organization of Turn-Taking for Conversation". *Language in Society*, 29:1, 1-63.

Vella, A. and P-J. Farrugia. (2006). "MalToBI – Building an Annotated Corpus of Spoken Maltese". *Speech Prosody 2006*, Dresden.

Vella, A., Chetcuti, F., Grech, S. and M. Spagnol. (2010) "Integrating Annotated Spoken Maltese Data into Corpora of Written Maltese", in *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC '10)*, Workshop on Language Resources and Human Language Technologies for Semitic Languages, pp. 83-90, Valletta, Malta

Quantification of Self and Mechanical Turk: Two Future Pillars for Multimodal Corpora Research?

Jean-Claude MARTIN (LIMSI-CNRS), MARTIN@LIMSI.FR

ABSTRACT

Research in Multimodal Corpora has been quite active in the last ten years (Kipp et al. 2010). In my presentation, I will discuss about two potential future pillars of multimodal corpora research which interestingly seat on two borders of the individual participant: 1) the individual and the quantification of self, and 2) collective and third party annotations (e.g. Mechanical Turk).

I will illustrate these two perspectives with multimodal data collection situations which might have a strong impact on the future of multimodal corpora research. They might change the way we collect data, the type of data we collect, the way we collect annotations and the way we consider privacy issues. They might have an impact on the aim of multimodal corpora and on the methodology that is used in multimodal corpora research.

The individual and the quantification of self. Affective computing applications and experimental protocols involve more and more frequently the psychological and physiological self of participants, for example in the case of job interviews (Hoque et al. 2013, Giraud et al. 2013). Both the form and the content of this collected data need to be considered. More precisely, "Quantification of self" refers to the recent possibilities for everyone to log information about ones' general features (e.g. weight), activity (e.g. walking steps per day using , running time ; speed using RunKeeper iPhone app), and physiological signals (e.g. heartbeats, skin conductance using the mybasics.com watch). Several companies are selling recent devices that aim at tracking individual data for monitoring health, sport and other activities, often combined with location data. People are also collecting video of their everyday life. The design and evaluation of such equipment and use could benefit from psychological perspectives on personality and the self (Faur et al. 2013, Skowronski 2012, Morf 2006). Multiple devices enable to collect data on users' activities (e.g. vision processing by a robot companion) and use (eg logs of individual and familly use of a TV box) (Buisine et al. 2010).

Collective and third party annotations (e.g. crowdsourcing). These initiative tend to distribute the collection and/or annotation of data and hire (paid) annotators over the Internet. For example, Amazon's mechanical Turk service is used more and more frequently in recent studies for the annotation of multimedia data (Park et al. 2012). Researchers also consider means for validating these subjective annotations.

I will compare these two extreme cases in terms of goals and uses, lifetime, collected signals and data, collected participants, annotators, manual vs. automatic annotation, data sharing, ethics and privacy.

Related issues that I will also consider in my presentation include: personality questionnaires which are becoming more and more used, annotations which are made by different entities (self vs. friends vs. expert vs. anybody hired or not), quantified self that becomes shared with others for social reasons, and finally a shift going from in lab data to in the field data.

Finally, I will consider the impact of these two strands, beyond the mere multimodal corpora area of research, on the design of multimodal interfaces such as companions (Faur et al. 2013).

References

Buisine, S., Fouladi, K., Nelson, J., Turner, W. (2010). Optimiser le processus d'innovation grâce aux traces informatiques d'usages. IC'2010 Journées francophones d'Ingénierie des Connaissances, pp. 145-156.

Faur, C., Clavel, C., Pesty, S., Martin, J.-C. PERSEED: a Self-based Model of Personality for Virtual Agents Inspired by Socio-cognitive Theories. 5th biannual Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII 2013). Geneva, Switzerland, September 2-5, 2013. Published by IEEE Computer Society. (acceptance rate = 31% for oral presentations)

Giraud, T., Soury, M., Hua, J., Delaborde, A., Tahon, M., Gomez D.A., Eyharabide, V., Filaire, E., Le Scanff, C., Devillers, L., Isableu, B., and Martin, JC. Multimodal Expressions of Stress during a Public Speaking Task. 5th biannual Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII 2013). Geneva, Switzerland, September 2-5, 2013. Published by IEEE Computer Society. (acceptance rate = 31% for oral presentations)

Gomez, D., L. Philip, C. Clavel, S. Padovani, M. Bailly and J.-C. Martin (2013). Video Analysis of Approach-Avoidance Behaviors of Teenagers Speaking with Virtual Agents. ACM International Conference on Multimodal Interaction (ICMI 2013). Sydney, Australia, ACM.

Hoque, M. E., M. Courgeon, B. Mutlu, J-C. Martin, R. W. Picard, MACH: My Automated Conversation coacH, To appear in the 15th International Conference on Ubiquitous Computing (Ubicomp), September 2013. (Acceptance rate: 23.4%). Best Paper Award.

Kipp, M., J.-C. Martin, P. Paggio and D. Heylen (2010). Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality. Workshop on "Multimodal Corpora", in conjunction with the 7th International Conference for Language Resources and Evaluation (LREC 2010). Malta.

Morf, C.C. “Personality reflected in a coherent idiosyncratic interplay of intra-and interpersonal self-regulatory processes,” J. Pers., vol. 74, no. 6, pp. 1527–1556, 2006.

Park, S., G. Mohammadi, R. Artstein and L.-P. Morency (2012). Crowdsourcing Micro-Level Multimedia Annotations: The Challenges of Evaluation and Interface. International ACM Workshop on Crowdsourcing for Multimedia held in conjunction with ACM Multimedia 2012 (CrowdMM 2012). Nara, Japan.

Skowronski, J. J. “I, Me, Mine: Variations on the Theme of Selfness,” Soc. Cognition, vol. 30, no. 4, pp. 396–414, 2012.

The Parody of politicians: a multimodal tool for political satire

Isabella Poggi and Francesca D'Errico

Discredit, ridicule, and parody in political communication

Within research on political communication, various strategies have been studied as to how a political party may try to overcome another, among which fair argumentation, fallacies, and the appeal to emotions. A specific strategy recently investigated is one of casting discredit over political opponents, and within this, beside serious attacks a quite effective one is to make fun of them. This may be done not only by a politician, but also by a reporter, a journalist, a film director, an actor, a comedian; and the most typical way in which a comedian can definitely cast discredit over a political character is through parody.

Casting discredit over someone means to spoil his image, and a politician's image is generally spoiled by attacking him concerning the features of *benevolence* (caring the electors' goals, working on behalf of their interest, being trustworthy, honest, ethical), *competence* (expertise, skill, knowledge, planning and reasoning capacity), and *dominance* (capacity of winning in contests, of influencing others and imposing one's will) (D'Errico et al., 2012).

Making fun of somebody is a particular way to discredit someone, by casting a negative evaluation of impotence over him: a communicative act through which a Sender S remarks, in front of some Audience A, a feature of a victim V that shows his lack of power, that contrasts with V's pretence of superiority, but is seen as not threatening for S and A, so much so as to elicit their relief and laughter. When S makes fun of V before A, S and A feel superior to V (Bergson, 1901), because immune from his inadequacy and not threatened by it; this common superiority strengthens their social bond, through the emotion of laughing together, of feeling similar to each other and different from V, and a sense of complicity; V, with his image and self-image strongly attacked and spoiled, feels abasement, shame, humiliation, feelings of rejection, solitude, isolation from the group. Thus mocking and teasing are used as "moralistic aggression" (Bishof, 1985) to stigmatize members of a group and force conformism to group norms.

A particular way to make fun of something or someone is parody. In this paper we analyze some cases of parody in Italian political shows and propose a first sketch of the cognitive, communicative and motor processes implied in making parodies with a function of political satire.

A cognitive model of parody

Parody is a communicative act – a text or a multimodal communicative behavior (discourse, song, film, fiction) – that performs a distorted imitation of another text or multimodal behavior, with the aim of amusing and eliciting laughter, most often to make fun of the author or sender of that text or behavior. In political satire, to make fun of a politician the parodist generally imitates his/her communicative or non-communicative behavior, but performing a distorted imitation to enhance the Victim's flaws.

Generally the Parodist must single out the most characterizing features of V's physical traits or behaviors, and imitate them, that is, produce some trait or behavior Y while soliciting A to recognize it as similar to, or evoking, another internal or external trait or behavior X of V; but P must distort that trait or behavior, for example by exaggerating it, so as to make it appear ridicule. Sometimes, though, the Parodist "extracts" a deep, abstract feature of the Victim – a general attitude or a personality trait – and freely invents and performs the multimodal behaviors that may plausibly display it, even if they were never actually performed by the Victim himself. A key element to have the Audience connect the Parodist's behavior to the Victim is therefore allusion, in that P wants A to infer that P refers to X, but without explicitly mentioning it, yet simply making reference to it in an indirect way.

Based on this description, any parody contains the ingredients of

1. Imitation of or similarity with the Victim's traits or behaviors
2. distortion of the similarity, aimed at highlighting the Victim's ridicule features
3. allusion to real events in which the politician displayed his trait or behavior
4. humor, aimed at eliciting laughter

while political parody contains in addition

5. political criticism, that makes the parody a case of political satire.

Therefore, in each political parody three aspects can be highlighted:

a. the "allusion points": the contents in the background knowledge that are supposedly shared with the audience, and to which the Parodist alludes in his parody;
b. the "humor points", those in which the parodist aims at eliciting laughter
c. the inferences that can be drawn from the allusions
d. the resulting evaluation of the parodized politician in terms of the three politically relevant aspects above: *competence*, *benevolence* and *dominance*.

Multimodal communication in Parody

In a qualitative observational study we analyzed the multimodal communication in the parody of the Mayor of Rome Gianni Alemanno, where the Italian comedian Max Pajella makes fun of how he behaved during a rare event: the snow in Rome. In that situation Alemanno and his administration proved ignorant in meteorology and very inefficient, and to defend himself he accused the Civil protection for having left him alone in managing the event.
To analyze parodies we devised the annotation scheme of Table 1. Pajella's parody was annotated by two independent coders, who finally discussed and agreed about divergent annotations.
In columns 1. we write the words and in 2. the body traits or behaviors displayed, in col. 3 the "allusion points", in 4. the inference induced by the allusion, and in 5. a classification of the example in terms of the aspects made fun of, *competence*, *benevolence* or *dominance*. The "humor points" are simply highlighted in bold italics.

Table 1. Fragments in Pajella's parody of Alemanno

1.WORDS	*2.BODY TRAITS AND BEHAVIOR*	*3.BELIEF ALLUDED TO*	*4.MEANING TO BE INFERRED*	*5.RIDICULED FEATURE*
1	Alemanno dressed as a **Roman centurion**	men playing centurions with tourists are uneducated people from Roman slums, waiting for a tip after a picture.	Alemanno is a lout, a buffon,	**COMPETENCE** **DOMINANCE**
2	A. has a **shovel** in his right hand	he distributed shovels to Romans to help themselves, due to lack of snowplows he showed himself shoveling, to project an image of a willing boy doing everything to help.	Alemanno was unable to organize Rome for the emergency. Only the display of benevolence	**COMPETENCE** **BENEVOLENCE**
3	A. holds a **sheet of paper** in his left hand	the **newsletter** from the Civil Protection about the centimeters of water expected	A.is ignorant	**COMPETENCE**
4 *Sono stato lasciato solo…chiamo l'esercito, sono stato lasciato solo, chiamo l'esercito.* I have been left alone::: I'll call the army, I have been left alone, I call the army.	High voice intensity Frown and oblique eyebrows	A.often said he would call the army	He plans solemn actions But He is impotent	**DOMINANCE**
5 *Dovevano dircelo che l'acqua ghiacciava a zero* ***centigradi centimetri*** They should have told us that water would freeze at zero centigrade centimeters	*shows the paper* with the forecast, *stutters, cries, tightened eyebrows*	A.and his staff proved very ignorant about meteorological facts	A. is ignorant	**COMPETENCE**
6 *La Protezione civile aveva detto che sul Campidoglio* ***sarebbero piovuti soltanto 35 millimetri di mmerda.*** Civil Protection had said that only 35 millimeters **shit** would fall over Capitolium.		The snow was cause of a great loss of face for A.	A.was impotent in managing the event of snow in Rome	**DOMINANCE**

This annotation scheme allows to finally compute the number of allusion points and humor points; their ratio evidences the extent to which allusions are functional to making fun of the Victim, and the number of attacks to the three criteria of evaluation of the politician. In Pajella's fragment, for instance, the attacks to Alemanno as a politician are quite balanced, since he is made fun of 6 times as to Competence, 6 for Benevolence and 5 for Dominance.
After analyzing Pajellas fragment, we compared Pajella's parody with videos of the real Alemanno, to find out the specific traits of the Mayor's communicative and non-communicative behavior that were picked up and imitated by the parodist: for example, the trait of Alemanno's prosody, with *talk slowing down and then hurrying up*, is mimicked by Pajella. In other cases, what is taken up by the parody is not a single prosodic or gestural trait, but rather a whole attitude – for example, a didactic attitude – that the Parodist renders by bodily behaviors that are not exactly the same as the Victim's, but through other means represent the same "deep abstract feature": for example, Alemanno makes the *ring gesture* as to convey precision, *puts on glasses* to read a sheet of paper, *looks at the audience before reading on;* Pajella interprets and categorizes these behaviors as a didactic attitude and transposes them into other behaviors like reading the sheet of paper while *repeatedly pointing at it*. Again, the self victimization of Alemanno in accusing the Civil Protection to leave him alone with the problems of snow in Rome, which is expressed in a long discourse by the real Alemanno, is evoked by Pajella simply by his *oblique eyebrows* showing sadness as if imploring help.

This comparative analysis allows to assess which aspects of the Victim's traits and behaviors are selected by the Parodist to be simply mimicked and which instead are distorted in order to make fun of them.
While this work concerns the analysis of one parodist's behavior and a comparison with his victim's behavior, in future work various pairs of parodist and victims will be analysis in order to highlight the process leading to make a parody. Here is a quick sketch of the steps required:

1. goal of moralistic aggression toward a Victim, to be pursued through ridicule.
2. "sense of humor" (Ruch, 1998), i.e., the capacity for grasping scenarios feasible to being the object of fun or laughter, while distinguishing them from those that cannot or should not.
3. singling out one or more traits or behaviors of the Victim that are worth being ridiculed.
4. a Model of the Addressee: a representation of what knowledge about the Victim is shared, and whether it may be subject to negative evaluation by the Addressee too
5. capacity of imitating the traits, body behavior, verbal text or discourse singled out as potentially ridicule
6. capacity of exaggerating or distorting the features by still keeping them recognizable
7. using exaggeration or other cues as a signal of allusion to induce the Audience to search its memory for similar traits or behaviors in the Victim and attach a tag of ridicule to it

References

1. Bergson, H. (1900) Le rire. Essai sur la signification du comique. Paris: Éditions Alcan.
2. Bischof, N.: On the phylogeny of Human Morality. In: Stent, G.S. (ed.) Morality as a biological phenomenon. The Presuppositions of Sociobiological Research. University of California Press, Berkeley, CA (1980).
3. D'Errico F., Poggi I., Vincze L., Discrediting signals. A model of social evaluation to study discrediting moves in political debates. Special issue in "Social signal processing". Journal on Multimodal User Interfaces. Vol.6 (3-4),163-178 (2012)
4. Hulstijn, J., Nijholt, A. (eds.). Proceedings of the International Workshop on Computational Humour *(TWLT 12)*, University of Twente, Enschede, Netherlands (1996).
5. Ruch, W. (ed.). The Sense of Humor: Explorations of a Personality Characteristic. Mouton-de Gruyter, The Hague-Berlin (1998)

Svetlana L. Mishlanova
Anna E. Khokhlova
Ekaterina V. Morozova
Perm State National Research University

Verbal and gestural representation of the space-time relation in the oral narrative

This research deals with the consideration of verbal and gestural representation of space-time relation in multimodal communication.

The communication is regarded as multimodal in case of including both modalities: spatial-visual (gestural) and oral-aural (speaking).

The aim of this research is to define the way space and time relate in verbal and gestural forms in oral narrative of Russian-speaking students.

As the material for this research we used
1. the videos with the interviews of the Perm state national research university students,
2. the media files viewing and editing program "Sony Vegas Pro 11.0", "VLC Media Player".

Our research is based on the works of foreign and Russian researchers in the field of cognitive linguistics such as: Alan Cienki, Cornelia Müller, Daniel Casasanto, E.A.Grishina, E.S.Kubryakova, N. D. Arutyunova, G.E.Kreydlin, etc.

In this particular research 14 students' interviews were analyzed. Among the students there are 8 female and 6 male participants.

The chosen subject is relevant because there are very few Russian researches on multimodal communication. In Russia studying of gesticulation is still regarded as the area which is more likely to be interesting to psychologists. That is the reason of bringing out of consideration the gestural component of multimodal communication in linguistic researches.

During multimodal communication in oral narrative along with a verbal component there is a gestural component.

Chenki and Müller research shows that in the mind of English-speakers time is represented along a peculiar temporal arrow-shaped axis, or a time vector.

The Casasanto research shows that when the English-speakers gesticulate deliberately the time axis model is directed vertically (sagittal model). In the case when gesticulation is represented spontaneously, the model of the time axis appears to be horizontal (lateral).

During our analysis of the video records we used the following procedure:
1. we divided the narrative into two modalities;
2. the next step is the division of the oral modality into events of the past, present and future; then division of events of the past, the present and the future into thematic groups

(criteria of this classification:

- general semantics of the sentences,
- words and expressions having temporary semantics, marking this or that time, for example adverbs (long ago, soon, now, earlier, later, then)
- the direct nomination of this or that time);

3. calculation of gestures in each episode, their division into three groups: right-handed (16%) , two-handed (65%) and left-handed (19%) gestures;
4. comparison of verbal and gestural representation of the space-time relation.

According to the results of the research the activity of gesticulation depends on gender accessory. The number of gestures the female informants made surpasses the number of gestures of the male informants by several times.

The biggest number of gestures (65%) was revealed in past events.

The biggest number of gestures in all episodes (65,5%) was made with two hands.

Speaking about the events of the past informants gesticulated with their left hands more often, whereas speaking about the events of the future they used the right-handed gestures more frequently.

On the basis of the obtained data we made the assumption that the concept of the lateral time axis in oral narrative can be also applied in Russian narrative, which means that the space-time model might be general for both Russian and English languages.

Relationship between home position-formation and storytelling.

Ryosaku Makino[1], Nobuhiro Furuyama[1,2]
[1] Department of Informatics, School of Multidisciplinary Science, The Graduate University for Advanced studies
[2] Information and Society Research Division, National Institute of Informatics
{ryosaku, furuyama}@nii.ac.jp

This presentation reports on the relationship between what we call "home position-formation" and storytelling. Home position-formation (henceforth referred to as HP-formation) refers to a combination of multiple participants' home positions. "Home position" is defined as the place where gestures depart from and return to. It manifests itself in various forms and people sometimes change it during conversation (Sacks & Schegloff, 2001). "Formation" as in HP-formation is meant to echo the spirit of Kendon's "F-formation (Kendon, 1990)". That is a pattern of mutual standing positions of multiple persons in and around the conversational event and these position indexes whether each one of them participates joins the conversation and, more specifically, their roles in the participation structure (Goffman, 1981). We suppose that HP-formation contribute to establishing the foundation for communication like F-formation. While F-formation relates to participation structure, in our conversation data (described in more detail below), all participants were recruited as the legitimate participants. Accordingly, we are not in a position to say much about how these factors contribute to communication. The present analysis thus addresses another possibility by examining HP-formation, and attempt to see whether it relates to communication, especially storytelling. Storytelling refers to a situation in which one of the participants tell a story to the other(s) (Mandelbaum, 2013). To enter a storytelling phase, there should be a mutual understanding between the teller and the listener(s) about whether or not storytelling is initiated. For this, the participants set up the "environment" for storytelling by giving a preface to the story or displaying a certain body behavior. The hypothesis of this study is that particular HP-formation predicts a storytelling phase to follow. In the following two analyses, we used twelve conversation data form Chiba university three participants corpus (Den &Enomoto, 2007). The conversations each engaged by three participants lasted for about 10 minutes. The participants were each seated in a chair surrounding three round tables as shown in Fig.1. Analysis I addresses what types of HP-formation were observed. We used a criterion to classify home position of individual participants. The criterion was the relationship between the participant's hands and the table. This criterion was used to separate home position status into two types. Type I is "ON", where the participant's both hands are on the table (Fig.2). Type II is "UNDER", where both hands are under the table (Fig.2). If all the participants' body status were simultaneously home position, this counts as "HP-formation." HP-formation can be considered as a combination of two types of individual home position, i.e., "ON" and "UNDER." Using these classifications, we examined what types of HP-formation there were.

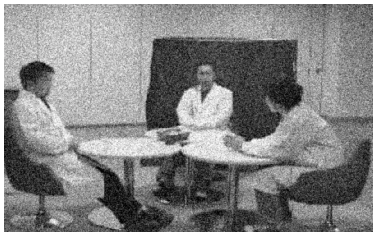

Fig. 1. The exemplary shot of the Chiba three-party conversation: The figure tessellated for protecting the participant privacy

Fig. 2 "ON" home position and "UNDER" home position

The result of analysis I suggests that HP-formation types, All UNDER (Both hands of all participants were under the table) and Solo ON (Only one participant's both hands were on the table and the others' both hands were under the table), appeared more frequently than others and appeared in multiple data sets. Thus, we decided to focus on All UNDER and Solo ON HP-formations in the analyses II. Analysis II addresses how different HP-formation types relate to storytelling. We analyzed the time difference (TD) between the onset of HP-formation and that of storytelling and examined whether HP-formation started before the onsets of storytelling or after them. Fig.4 shows frequency of TD between the onset of All UNDER and them of storytelling. Fig.5 shows frequency of TD between the onset of Solo ON and them of storytelling. What are common in the two types of HP-formation is that there is a frequency spike of HP-formation in the range of -10 sec and 10 sec from the onset of storytelling phase. What differentiate them is the following two points: (a) Looking at the distribution of HP-formation before the onset of storytelling, the frequency of All UNDER gradually increases as the time difference becomes closer to zero, while there were only a few, if any, instances of Solo ON in that range; (b) Looking at the distribution of HP-formation after the onset of storytelling phase, All UNDER gradually decreases after the large spike when the time difference is in the range of 0 sec and 10 sec, with only one exception when the time difference is over 100 sec. Contrastively, with regard to Solo ON, although there is a large spike of frequency when the time difference is in the range of 0 and 10 sec just like All UNDER, there is a sharp drop of frequency after that

range. However, that does not mean that there is no Solo ONs there; there are a few cases of Solo ON constantly when the time difference is more than 10 sec.

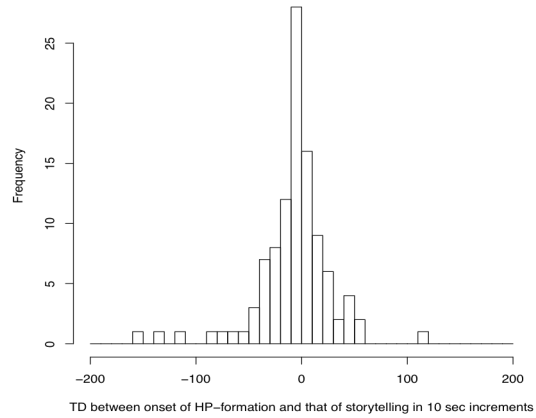

Fig. 3. Frequency of time-difference between the onset of All UNDER HP-formation and storytelling.

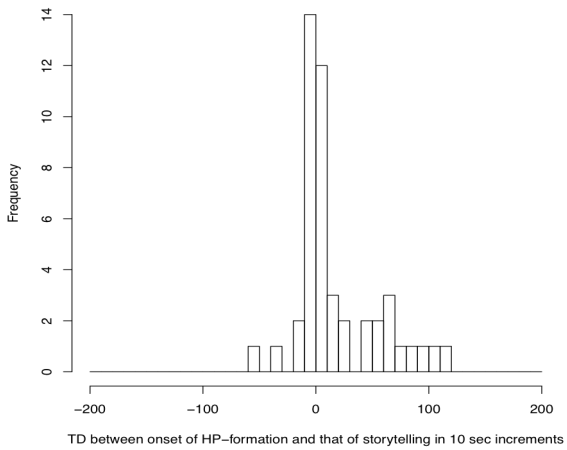

Fig. 4. Frequency of time-difference between the onset of Solo ON HP-formation and storytelling.

Provided the foregoing, let us discuss the relationship between the two types of HP-formation, All UNDER and Solo ON. As for Solo ON, we can at least identify two features: i) Solo ON seldom started when the TD was less than -10 sec, and many storytelling phases started 10 sec after when Solo ON started; ii) Solo ON can start when the TD is more than 50 sec. The former suggests that Solo ON, possibly, projects storytelling phase to begin soon after (though, it is necessary to take a closer look at what is happening before and after the appearance of Solo ON in the conversation to determine whether or not the notion of "projection" is really applicable here.) The latter feature of Solo ON, given the average duration of storytelling phase, suggests that Solo ON can occur in the middle of storytelling phase. What could this possibly mean? For one thing, Solo ON can be taken to mean that all three conversational participants agree that the storyteller can keep telling the story. The results (a) and (b) above regarding All UNDER that there is a certain correlation between it and the transition into storytelling phase may suggest that other information available to the participants, such as a preface to the story being offered by one of them, may also play a role in enhancing the correlation between the them. Although further studies are required to determine different possibilities suggested here, the analyses do imply the possible correlation between the HP-formation and conversation phase. In addition, we examined only one particular situation in which participants were seated in chairs surrounding desks. The types of home positions, i.e., UNDER and ON, depend on this situation. However, we think that HP-formation plays an important role in daily communication by considering that Solo ON HP-formation is treated as one participant's form different from the others and All UNDER HP-formation is treated as all participants' forms being the same. Therefore, what we take as critical is not weather or not the participants' hands are ON/UNDER the table, but weather or not they constitute patterns (All UNDER, Solo ON etc.) and weather or not these patterns predict or project what to come next in the conversation. Finally, the importance of the idea of HP-formation newly proposed in this very paper should considered in order to understand more fully how different kinds of multimodal information, i.e., speech, gestures, home positions, and other non-gestural and/or non-verbal behaviors, are organized to regulate the flow of conversation.

Acknowledgments

We are grateful to Professor Yasuharu Den at Chiba University for providing us with the Chiba three-party conversation corpus data.

Multimodal Metaphors of Health: an Intercultural Study

In light of the cognitive paradigm, metaphor is one of the keys to understanding how people categorize reality, what processes govern the conceptualization of the world. The study is aimed at revealing metaphorical representation of the universal concept of health in the individual perspectives on meaning regarding the issues of conceptualization and categorization of medical knowledge in Russian and American participants.

One of the methodological principles that we use in the research is the notion of "mode of discourse" (A Kibrik, 2008). Discourse is a "bi-unity of the communication process and its result, leading to texts", and thus it covers "all forms of language use " and types of discourse may be contrasted "by modus or transmission channel of information"(A Kibrik, 2003, 2008). Provided that the hierarchy of discourse modes is "mental, oral, written or graphic/visual," it includes multimodal types of discourse (submodes) e.g., oral and written, combining the features of spoken, written and graphic discourse, for example, in electronic communication. This view complies with a wide range of studies of multimodality (G. Kress, T. Van Leeuwen 2001; J. Bezemer, A. Cope., G. Kress, R. Kneebone 2011; A. Burn and D. Parker 2003, Kay L. O'Halloran 2008; J. Steen 2011, Alan J. Cienki, C. Muller 2008)

Medical discourse presents all kinds of discourse: mental, oral, written, electronic and visual modes. In all modes of discourse, a generalized representation of medical knowledge takes the form of the concept. In our study, the main issue is the problem of the conceptualization of medical knowledge through the multimodal metaphor.

The guiding method of the study is metaphorical simulation (modeling). (S. Mishlanova, T.Utkina 2008; S. Mishlanova, S, Polyakova 2011). A metaphorical scheme can be used as an instrument to describe a metaphorical representation of the concept in different types of discourse. The scheme consists of four basic metaphor models: **Man as a Human Being, Animate Nature, Inanimate Nature, and Man as a Social Subject**. Each of the basic metaphor models is structured as follows: Taxons (e.g.: *Professional Areas, Politics, War etc*.), Generic Taxons (e.g. *Army, Law, Weapons* etc.), Sub generic Taxons (*guns, steel* etc.), Specific metaphor (guilty allergen, shooting pain etc.) All the levels in the hierarchical structure are related to each other in a generic-sub-generic relationship.

In order to obtain a comprehensive view on multimodal metaphorical models in medical discourse, we carried out two sets of empirical studies - Associative Experiment and Free Drawing Project in groups of Russian and American respondents.

Respondents and Procedure: A psycholinguistic associative experiment was conducted in Perm, Russia and two U.S. states (Iowa and Kentucky) from 2008 to 2010. It included 209 participants: 108 Russians (58 women and 50 men) and 101

Americans (56 women and 45 men) aged 25-65, at average 41 yrs. The main criterion for choosing participants for the experiment was the lack of special medical knowledge and no experience in the medical sphere. The participants were requested to give answers in writing to the following statement which was given orally: «Health is like ....». 175 associations in Russian and 130 associations in English were chosen. The examples of associations in Russian and in English are as follows:

Russian: health is like *the sun; dawn at the sea; clean wood; an apple; an expensive car.* **English:** health is like *a clear stream of water; a sunny fall afternoon; blooming nature in the summer; a bank account; a jewel; helicopter control.* The metaphors obtained in both groups were assigned to the models of the metaphorical schemes.

The Free Drawing study included 64 Russian and 50 American participants. They were requested to draw a picture of health in response to the same oral instruction. Some respondents added a brief description to the drawings. 104 drawings of health were analyzed via metaphor models approach described above. The examples of the Russian respondents' drawings are: *the smiling sun, the flow, the tree with roots and apples, people walking in the woods.* The examples of the American respondents' drawings are: *the apple, the hills, the sun, the carrot, the dancing girl, the running man.* Every picture was divided into separate elements and each element was referred to a particular category.

The results of the first study demonstrate that metaphoric schemes of Associative fields of the Russians and the Americans are presented in the following way: Man as A Human Being (Russians - 4%, Americans - 15%), Animate Nature (Russians -34%, Americans - 4%) Inanimate nature (Russians-22%, Americans – 24%), Man as a social subject (Russians – 40%, Americans – 58%).

The findings of the second study show that the 4 basic models of visual metaphors are represented as: Man as A Human Being (Russians - 14%, Americans- 28%), Animate Nature (Russians -38%, Americans - 26%) Inanimate nature (Russians - 15%, Americans – 14%), Man as a social subject (Russians – 33%, Americans – 32%).

It is seen that the most active metaphorical model is the Man as a Social Subject both in the mode of associations (Russians - 40% Americans - 58%) and in the visual metaphors (Russians - 33 %, Americans – 32%), whereas the Animate Nature is the second significantly active model in both modes (34% and 38% respectively). The least active metaphorical model for the Russians is Man as Human Being (4%), while the least active metaphorical model for the Americans is Animate Nature (4%).

In the study of multimodal metaphor of health in the associations and the drawings, we revealed that in the visual representation of the concept of health, there is a general similarity between the metaphorical schemes of the drawings in two different cultural groups, while in the metaphorical schemes of written

representations of health we found some differences in the Russians and the Americans.

Based on this interdisciplinary project, the paper will highlight the differences and similarities in verbal responses and drawings and discuss a further application of multimodal metaphor of health in different cultures and discourses with some implications for multimodal communication.

Precision gestures in oral examinations and political debates

1. Introduction

When we talk to other people, besides informing on our topic, we also inform about the level of certainty, uncertainty, vagueness, approximation, precision or specificity of what we are saying.

These are all concepts highly related but nonetheless distinct.

Vagueness is defined by Poggi & Vincze (2011) as a property of the knowledge assumed about a certain topic: a lack of detail in what one knows about something. As we are vague, we do not have a detailed knowledge of the topic, but only general beliefs, and not ones on particular aspects of it.

Vagueness and approximation are both antagonistic of precision, defined as the fact of having beliefs on each specific aspect of a topic. But while vagueness has to do with describing, approximation involves measuring. We can say therefore that while the first concerns qualitative aspects, the second focuses on quantitative ones.

Vagueness was also distinguished from uncertainty, since we may have a vague knowledge, a vague idea, a vague remembering of something, but still be certain of it.

Hence, we can be certain and vague at the same time, but we cannot be approximate and certain at the same time, as approximation contains an underlying quantitative imprecision due to uncertainty.

For the purpose of the present paper, it is important to differentiate between the concepts of *precision* and *specificity*. Poggi & Vincze (2011) see precision both as a property of knowledge and as a property of communication. In the former case we may say we have a precise idea or a precise memory of something when, while thinking of that topic, we have beliefs about each single aspect of it; here what is precise is the way in which a concept or notion is represented in our own mind (for instance, with clear-cut boundaries and vivid details). In the latter case, precision refers to how a topic, concept or notion is phrased in our discourse (for instance, by distinguishing sub-topics and making one's discourse structure explicit). Precision differs from *specificity* because specificity dwells in the field of definition, of recognizing entities and assigning them to classes, while precision holds in the field of description, that is, it pertains to the goal of having (and consequently, possibly providing) a more thorough knowledge of a single entity. In terms of cognitive structures, definition and specificity regard semantic memory, while description and precision have to do with episodic memory.

The need to inform about our level of certainty, uncertainty, approximation, precision or vagueness stems from Grice's (1975) maxims of quality and quantity, that imposes not to say what one believes to be false and that for which one lacks adequate evidence (maxim of quality) or not to say more nor less than what is required by the circumstance (maxim of quantity). These norms determine the threshold of precision we should stick to. When we keep below the required level of information, either because our knowledge is in itself vague or because, though we could go into details, we do not want to, we may meta-communicate that we are being vague, by acknowledging our vagueness through verbal or bodily signals of vagueness (Vincze & Poggi, 2011).

But if when vague we remark our vagueness, in the same vein, when we consider that details are particularly important to our discourse, we may keep ourselves above the required precision and meta-communicate our goal of being more precise. We may do so

by verbal expressions like *"precisely"*, but also by bodily signals: for example I may squint eyes to convey that I am picking up a very specific detail of the topic dealt with, or I may perform various types of precision gestures. .

2. Types of precision gestures

Previous works have already dealt with the issue of gestures conveying precision. Already Morris (1977) noticed that the *precision grip* gesture is typically used when "the speaker wants to express himself with delicacy and with great exactness, while his hand emphasizes the fineness of the points he is stressing".

Kendon (2004), in his analysis of pragmatic gestures, investigates two gesture families which in certain contexts of occurrence convey the concept of precision: *grappolo* family (G-family) and ring family (R-family). Calbris (2003), in her analysis of Lionel Jospin's gestures, dedicates a chapter to the precision gestures made by the leftist politician. According to Calbris (2003), as for Kendon (2004), the *grappolo* (called pyramid by Calbris), conveys the condensed, the quintessence of something that the opening of fingers is going to set free so that the listener can discover it. But in addition to Kendon's view, according to Calbris, the pyramid can also emphasize the particular character of the mentioned thing and it is evocative of precision thanks to the very position of the palm: closed and hiding a secret content about to be revealed.

Ring gestures are used in a context where the speaker is indicating that he means to be very precise about something, that what he is saying is 'exact' in some way, and that it demands special attention for this reason. (Kendon 2004, 228). In fact, Calbris (2003) associates ring with insistence from the sender's side. According to Lempert (2011), ring-precision grips indicate the focus of discourse, but they have also undergone a degree of conventionalization and acquired the performative meaning of "making a 'sharp', effective point", thus finally reflecting this image to the speaker who starts to be perceived by the audience as "being argumentatively sharp".

Other precision gestures analysed by Calbris (2003) are: the frame (le cadre), i.e. open hands with palms facing each other and fingertips forward as if holding and touching the sides of a box; the pincers (*les pinces*), i.e. thumb and index touching, with the rest of fingers closed in the palm. According to Calbris, even the raised index finger can contribute to precision by signalling it.

3. Precision gestures in debates and oral exams: an observational study

In this paper we aim to analyze multimodal signals of precision (reversed pyramids, rings, pincers) that speakers employ when wanting to convey to interlocutors that they are being precise because precision is important to the aims of the discourse. In most cases, precision is multimodally communicated: concomitant to adverbs of precision, the speaker reiterates his goal of being precise through the performance of concomitant body signals conveying precision

3.1. Method

We analysed the precision gestures occurring in two different types of corpora, oral examinations in Psychology and political debates. To single out and analyse these gestures we used two methods, "verbal-to-body" and "body-to-verbal". In the former, we transcribed the whole interaction, then we looked for adverbs of precision in the verbal transcription, and finally described and analysed the co-occurring body signals. In the latter, without listening to the verbal content, we looked for body signals that to our communicative intuition looked as "signals of precision", and then checked the

plausibility of their interpretation in the verbal transcription. For each case, we analysed the body signals of precision in terms of their physical production, their meaning, and the reason for use in their context.

3.2. Results:

We noticed that while the *ring* gesture is very recurrent in the corpus of political debates, it is not at all employed in the oral examination corpus. Due to the ring's being also associated with insistence from the sender's side, it might be perceived by both sender and addressee as an inappropriate signal during non-peer interaction. Debates are a context of argumentation within a peer interaction, while exams are one of simple exposition within an asymmetrical interaction (Orletti 2000), finally aimed at the student's evaluation. When contradicted, the student rarely insists on supporting his thesis with argumentation but easily gives up. This could be the reason why we have not found any *ring* gestures in our corpus. The *ring* might be a gesture more typically devoted to convey precision in argumentative contexts in which conversationalists have the same interactional power.

Furthermore, we found some systematic differences between gestures of precision versus gestures of vagueness.

The gestures of precision are characterized by features that, interestingly enough, quite systematically contrast with those of vagueness, found by Vincze & Poggi (2012).

- contact between fingers, as opposed to open hand in gestures of vagueness. This physical feature conveys a morpho-semantic feature of grasping, of picking up a single object, worth being caught because relevant; (as Calbris notices, all these gestures in their parameters: hand orientation, movement, finger position, direction of movement, have something that recalls concepts related to precision: precise borders, as the gesture of the *frame* (fr. *cadre*), or pinpointing precise details, (gesture of *pincers*).
- high muscular tension, as opposed to loose hands in vagueness gestures. This conveys a meaning of concentration and focused attention of the gesturer, that is transmitted to the Interlocutor as a request for attention
- straight and precisely targeted direction of movement, as opposed to wavy movement in vagueness gestures. This physical feature conveys a morpho-semantic feature of punctuation, of aiming at single points – and small points, details – within the object of discourse.

This contrast recalls Darwin's (1872) principle of opposition.

From these results, though, we have not yet drawn significant differences between gestures of precision and gesture of specificity. Further research will focus on this possible difference and will find methods to operationalize this difference and test it in experimental settings.

References

Calbris, G.: L'expression Gestuelle de la Pensée d'un Homme Politique. Paris: CNRS Editions (2003)

Darwin, C.: The expression of the emotions in man and animals (3rd ed.; P. Ekman, Ed.). London: HarperCollins (1998). (Original work published 1872)

Grice, P., H.: Logic and Conversation. In: Cole Peter, Morgan Jerry, L. (eds.) Syntax and Semantics, 3, Speech Acts, 41–58, New York: Academic Press (1975)

Kendon, A.:. Gesture. Visible Action as Utterance. Cambridge: Cambridge University Press (2004)
Morris, D.: Manwatching: A Field Guide to Human Behaviour, London: Jonathan Cape Ltd (1997)
Orletti, F.: La conversazione diseguale, Roma: Carocci Editore (2000)
Poggi, I.: La mano a borsa: analisi semantica di un gesto emblematico olofrastico. In Comunicare senza parole. La comunicazione non-verbale nel bambino e nell'interazione sociale tra adulti, ed. by Grazia, Attili e P.E.Ricci Bitti, 219-238. Roma: Bulzoni (1983)
Poggi, I., Vincze, L.: Communicating vagueness by hands and face. In Proceedings of the International Conference on Multimodal Interaction, Workshop on Multimodal Corpora, Alicante, Spain (2011)
Vincze, L., Poggi, I., D'Errico, F.: Vagueness and dreams. Analysis of body signals in vague dream telling. In Human Behaviour Understanding. Lecture Notes in Computer Science (7559), 77-89 (DOI) 10.1007/978-3-642-34014-7_7 (2012)

Multimodal Representation of the Concept of Happiness in Russian Students' Narrative

This research is aimed at the analysis of multimodal representation of Russian everyday understanding of happiness. We study this layer of the concept of HAPPINESS in Russian students' narrative. The data we analyse is not only texts but hand gestures representing this abstract notion, so the representation we study is verbal and gestural.

To analyse multimodal representation of the concept of HAPPINESS, first of all, we recorded interviews and made video transcripts. According to the study design respondents were asked to describe a state of happiness they have ever been in. The interviews averaged less than five minutes and took place over a one-week period. In total we recorded 25 interviews.

Secondly, we divided interviews into events. *An event* is a relatively short narrative about one precise situation in which the interviewee felt happy. We made the division on the basis of spoken data with special regard for *"the conceptual formula of happiness"* developed by S. Vorkachev [Vorkachev 2001, 2003]. It is a logical formula showing the evaluative relation (including the basis of evaluation) between the subject of cognition and the objective world[1] [Vorkachev 2001]. The formula is based on the conception of happiness as subjective appreciation of life as a whole [Argyle 1987]. Thus the concept of HAPPINESS has two major components: *a state of affairs* and *a state of mind* [Vorkachev 2003]. A state of affairs is referred to as *an objective component* and a state of mind is referred to as *a subjective component*. In the semantics of HAPPINESS evaluation (both intellectual and emotional) links the two components. In other words, various facts of the objective world are evaluated by a subject of cognition as making him/her happy. Such facts of the objective world are called *sources of happiness*.

We divided 25 interviews into 76 events. The structure of each event corresponds to the formula of happiness: the interviewee describes the situation when he/she felt happy (an objective component is represented) and tells about his/her emotions at that moment (a subjective component is represented). We applied content analysis to the parts of events which contain representation of an objective component to find exact sources of happiness for each event. We classified the sources of happiness (83 in total; 100%) according to the topic discussed. Thematic groups include relations (30; 36%), studies and work (16; 19%), attainment of goals (12; 15%), material objects (9; 11%), nature (4; 5%), vacations and tourism (4; 5%), memorable occasions (4; 5%), pets (2; 2%) and creative work (2; 2%).

After we had analysed the spoken data we proceeded to the analysis of the visual one. As it has been mentioned above, we study hand gestures. We apply *the Method of Gesture Analysis* developed by C. Müller, E. Fricke, H. Lausberg, and K. Liebal [Cienki 2010]. The method includes three steps: gesture identification, gesture form analysis and gesture interpretation. The respondents produced 422 gestures in total, 76 of which were adopting (we do not work with adopting gestures in this research). In the rest 346 gestures a particular group of gestures drew our attention. This group consists of gestures of the following forms: 2H PO, 2H PFlatVert, 2H PFlatHoriz, 2H Fist, 2H ZipClosed, 2H ZipOpen, 2H Closed. All these gestures are upward movements of hands shaping a circle or a semicircle. Gestures of this group occurred 120 times in 72 events out of 76. It is notable that an upward movement of hands shaping a circle or a semicircle was detected at least once in all the events apart from those where interviewees produced either no gestures at all (3 events in one interview) or adopting gestures only (1 event). The very character of the movement suggests that it has positive implications (consider a well-known conceptual structure GOOD IS UP [Lakoff, Johnson 1980]). Thus we refer to the gestures of this group as *"gestures of happiness"* used when one is talking about happiness in its everyday sense. We classify these gestures as referential.

[1] Here and further the translation of quotations from Vorkachev's works from Russian into English is mine – M. Suvorova

Since events in this research are regarded as complex unities of the linguistic and paralinguistic we studied the interrelation between spoken data and gestures as well. This part of the research was based on studying *referents* of "gestures of happiness". In gesture studies the term *reference* is being used based on an interpretation of what the speaker may have had in mind when producing the gesture in light of the speech with and around it [Cienki 2010]. To all the referents (120 in total) we applied thematic classification as we had done to the sources of happiness before. Feelings and emotions (37) amounted to 31% of all referents which provided additional proof that the gesture in question is closely related to positive emotions. These feelings and emotions described by the interviewees form the subjective component of the concept of HAPPINESS. When spoken data represent the subjective component of the concept, gestures represent this component as well. ***Thus, the subjective component of the concept of HAPPINESS has multimodal representation in Russian students' narrative***.

The rest 69% of the referents of "gestures of happiness" turned out to be the sources of happiness named by the interviewees. In such cases the objective component of the concept of HAPPINESS is represented verbally, while its subjective component is represented in a special gesture simultaneously[2]. ***Therefore when a subject of cognition evaluates various facts of the objective world as making him/her happy this person is likely to produce "a gesture of happiness" while talking about these facts expressing both components of the concept at the same time.***

Both objective and subjective components cannot be represented in speech simultaneously due to its linear character. It means that analysing one's speech we are unable to detect whether certain facts of the objective world are appreciated or depreciated by the speaker unless it is not expressed implicitly or explicitly in words. But speech is only one mode of expression. When we analyse narrative from the point of view of its multimodality we can grasp a deeper sense in it. ***Thus when we study multimodal representation of the concept of HAPPINESS we can see appreciation on the part of the speaker of certain facts of the objective world on the basis of gesture analysis before any kind of positive emotions are expressed in words.***

References

Argyle, M. (1987) The Psychology of Happiness. Great Britain: Methuen Publishing.

Cienki, A. (2010) Multimodal Metaphor Analysis // Metaphor Analysis: Research Practice in Applied Linguistics, Social Sciences and the Humanities. Ed. by Lynne Cameron, Robert Maslen. Great Britain: Equinox Publising.

Gesture studies. Vol. 3 *Metaphor and gesture* (2008). Ed. by A. Cienki & C. Müller. Amsterdam / Philadelphia: John Benjamins Publishing Company.

Lakoff, G. and Johnson, M. (1980) Metaphors We Live By. Chicago: University of Chicago Press.

Воркачев С. Г. (Vorkachev S.) Концепт счастья: понятийный и образный компоненты // Известия РАН. Серия лит-ры и языка, 2001. Т. 60, № 6. С. 47-58.

Воркачев С. Г. (Vorkachev S.) Сопоставительная этносемантика телеономных концептов «любовь» и «счастье» (русско-английские параллели): Монография. Волгоград: Перемена, 2003. 164 с.

[2] We did not analyse cases when the objective component is represented multimodally, i.e. in speech and referential gestures other than "gestures of happiness".