Proceedings from the 1st European Symposium on Multimodal Communication

University of Malta, Valletta, October 17–18, 2013

Edited by

Patrizia Paggio and Bjørn Wessel-Tolvig

Copyright

Linköping Electronic Conference Proceedings, No. 101
ISBN: 978-91-7519-266-6
ISSN: 1650-3686
eISSN: 1650-3740
NEALT Proceedings Series, 21
URL: http://www.ep.liu.se/ecp_home/index.en.aspx?issue=101
Linköping University Electronic Press
Linköping, Sweden, 2014

Preface

This volume of proceedings collects a choice of the papers that were presented at the 1st European Symposium on Multimodal Communication held at the University of Malta in Valletta, Malta, on 17-18 October 2013. The purpose of the symposium was to bring together researchers who study multimodality in human communication as well as human-computer interaction, and to provide a multidisciplinary forum for scholars from different fields. The symposium continued and further developed a tradition of similar scientific events held previously in the Nordic countries, by aiming at a more international audience.

The collection of papers presented here reflects both goals. The range of topics is broad, the authors represent a large number of different nationalities, and the data examined concern a range of different language communities. The papers are listed in alphabetical order in the table of contents. In what follows, we mention them in a slightly different order in an attempt to tease out some commonalities of topic or method.

The first paper, by Allwood, Lanzini and Ahlsén, investigates the way in which the audio and visual modalities contribute to the perception of affective-epistemic states (happiness, interest, disinterest, understanding, etc.) in Swedish dyadic interactions. The authors claim that the two modalities may inhibit or reinforce each other depending on the nature of the epistemic state. In the second paper, Chollet, Ochs and Pelachaud look at a related phenomenon, stance, in particular the two dimensions of friendliness and dominance, which they want to model in view of generating situation-sensitive behaviour in embodied agents. They propose a data mining technique to extract non-verbal sequences expressing stances in a corpus of job interviews in French.

The question of the role played by audio and visual modalities in the perception of multimodal behaviour is also taken up in the paper by Gilmartin, Hennig, Chellali and Campbell, which deals with the perception of laughter. It is argued that while recognition based on the audio signal works for stereotypical laughter, the visual modality is necessary to capture other kinds. The authors investigate laughter in task-based and social interaction, using different types of corpora. A similar concern for the way different communicative situations influence speaker behaviour can be found in the paper by Paggio and Vella, which investigates the use of overlapping in two different corpora of Maltese. It is found that overlaps occur more frequently in free conversations than in task-oriented dialogues, and also that they are of different lengths and serve different functions in the two types of communication.

Three of the papers in the collection deal with hand gestures. The paper by Lis and Navarretta focuses on referential hand gestures, in particular the issue of how form and meaning are related in the multimodal expressions of events in Polish narratives. They formalise event meaning in terms of ontological categories from the Polish WordNet, and apply machine learning to the data to predict gesture features based on the semantics of the corresponding verbs. The paper by Wessel-Tolvig addresses a topic which many other gesture linguists have discussed, namely the way speakers of different languages conceptualise and express motion events differently both in terms of spoken language and accompanying gestures. The pair of languages examined here, Danish and Italian, has not been compared before. Especially for Italian, the paper claims that several multimodal strategies are available to speakers. The semantic and cognitive interpretation of gestures is also the subject of the paper by Mishlanova, Khokhlova and Morozova, which focuses on the relation between features of hand gestures and the temporal dimension of narrated events in a corpus of Russian narratives.

The claim put forward by the authors is that there is a correlation between handedness and the future-past dichotomy, and that this correlation presumably holds for European languages in general.

The paper by Poggi and D'Errico deals with the expression of parody. The authors propose a cognitive model of how parody is constructed, and exemplify the various expressive components of their model by considering a number of concrete examples from a corpus of parodies of Italian politicians by well-known comedians. The method used by the authors is mainly qualitative, in contrast to the other studies, and thus contributes to the intended interdisciplinary flavour of this collection.

Finally, the paper by Jokinen provides an overview of eye-tracking technology and how it is used to investigate the use of eye-gaze in multimodal communication. In particular, the author presents two studies in which eye-tracking has been instrumental in discovering the role played by eye-gaze in turn taking coordination, and in studying the behaviour of silent partners in multiparty conversations.

Patrizia Paggio and Bjørn Nicola Wessel-Tolvig

Organising Committee

Patrizia Paggio, University of Copnhagen and University of Malta
Jens Allwood, University of Gothenburg
Elisabeth Ahlsèn, University of Gothenburg
Kristiina Jokinen, University of Helsinki and University of Tartu
Costanza Navarretta, University of Copenhagen

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

Table of Contents

Contributions of Different Modalities to the Attribution of Affective-epistemic States
Jens Allwood, Stefano Lanzini and Elisabeth Ahlsén 1

Investigating Non-Verbal Behaviors Conveying Interpersonal Stances
Mathieu Chollet, Magalie Ochs and Catherine Pelachaud 7

Exploring Sounded and Silent Laughter in Multiparty Task-based and Social Interaction - Audio, Video and Biometric Signals
Emer Gilmartin, Shannon Hennig, Ryad Chellali and Nick Campbell 17

Eye-trackers and Multimodal Communication Studies
Kristiina Jokinen 29

Classifying the Form of Iconic Hand Gestures from the Linguistic Categorization of Co-occurring Verbs
Magdalena Lis and Costanza Navarretta 41

Verbal and Gestural Representation of the Space-time Relation in Multimodal Communication
Svetlana Mishlanova, Anna Khokhlova and Ekaterina Morozova 51

Overlaps in Maltese Conversational and Task-Oriented Dialogues
Patrizia Paggio and Alexandra Vella 55

Cognitive processes and multimodal communication in the parody of politicians
Isabella Poggi and Francesca D'Errico 65

Up, down, in & out: Following the Path in speech and gesture in Danish and Italian
Bjørn Wessel-Tolvig 73

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

Jens Allwood
SCCIIL Center
University of Gothenburg
jens@ling.gu.se

Stefano Lanzini
SCCIIL Center
University of Gothenburg
lanzhbk@hotmail.it

Elisabeth Ahlsén
SCCIIL Center
University of Gothenburg
eliza@ling.gu.se

Abstract

The focus of this study is the relation between multimodal and unimodal perception of emotions and attitudes. A point of departure for the study is the claim that multimodal presentation increases redundancy and often thereby also the correctness of interpretation. A study was carried out in order to investigate this claim by examining the relative role of unimodal versus multimodal visual and auditory perception for interpreting affective-epistemic states (AES). The abbreviation AES will be used both for the singular form "affective-epistemic state" and the plural form "affective-epistemic states". Clips from video-recorded dyadic interactions were presented to 12 subjects using three types of presentation, Audio only, Video only and Audio+Video. The task was to interpret the affective-epistemic states of one of the two persons in the clip. The results indicated differences concerning the role of different sensory modalities for different affective-epistemic states. In some cases there was a "filtering" effect, rendering fewer interpretations in a multimodal presentation than in a unimodal one for a specific AES. This occurred for *happiness*, *disinterest* and *understanding*, whereas "mutual reinforcement", rendering more interpretations for multimodal presentation than for unimodal video or audio presentation, occurred for *nervousness*, *interest* and *thoughtfulness*. Finally, for one AES, *confidence*, audio and video seem to have mutually restrictive roles.

1 Introduction and background

1.1 Different perceptual modalities and affective-epistemic states (AES)

It is often claimed that multimodal presentation, of information, i.e., information presented to more than one sensory modality, provides more redundancy and is therefore easier to interpret correctly than unimodal presentation. Support for this comes, for example, from studies showing that gestures enhance the comprehension and memory of a spoken message (Beattie and Shovelton, 2011). However, the relative contribution of cues from different perceptual modalities, in relation to each other, for the interpretation of AES has not been extensively researched. To some extent, it has been addressed in studies aiming at the automatic recognition and generation of emotional cues in Embodied Communicative Agents (ECA) (see, for example, Abrilian et al. 2005). Other examples include a database of multimodal videorecorded emotional interactions, established by Douglas-Cowie et al. (2000) and the complex emotions from videorecorded interactions that were analyzed by Buisine et al. (2006). The analysis of Buisine et al. 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. Also research concerning visual face recognition and auditory speech analysis for emotions and attitudes, aiming at automatic signal processing and generation, is, to some extent, although less ecologically valid, relevant in this context (cf., for example, Vinciarelli et al. 2012, Cunningham et al. 2005, Cohn and de La Torre, forthcoming).

By affective-epistemic states we mean internal (mental) states that simultaneously involve cognition, perception and emotion (Allwood, Chindamo and Ahlsén, 2012). The term is chosen in order to capture not only the major categories of emotions, like *happy*, *angry* and *sad*, some of which do not occur very often in most recorded everyday interaction data, but also to include more commonly occurring affective-epistemic attitudes, like *surprise*, *boredom*, interest etc., since these are important for understanding what goes on in ordinary face-to-face interaction.

1.2 Some conceptual issues and distinctions concerning AES

Some of the theoretical issues concerning AES are related to the following conceptual distinctions:

(i) Simple (pure) AES (e.g. happiness) versus complex (mixed) AES (e.g. nervousness) (cf. also Buisine et al. (2006), as well as

(ii) Simple (pure) versus complex (mixed) behavioral cues for AES. Compare smiling, which is visual and therefore simpler than laughing, which is visual and auditive.

These distinctions can be helpful, since, in general, we cannot expect one-to-one relations between AES and behavioral cues, but rather many-to many relations. For example, the behavioral cues of smiling and laughing can both be related to the AES of happiness, nervousness, fearfulness, politeness, tension release and subordination. They can also be related to complex combinations of these, such as "nervous happiness" and "happy nervousness".

Other relevant distinctions obtain between different physiological and behavioral cues, such as facial gestures and other visible body movements, voice quality and other audible vocal characteristics, skin conductance, heart-beat, blood pressure, digestive type and rate, and brain activity, indicating particular AES in particular ways.

In the light of this potential complexity, we can ask whether all cues are equally important for all affective-epistemic states or whether AES in general as well as more specific AES are more oriented to some modalities than others? For example, are cues for happiness more visual than auditive?

Other features of interest are "noticeability", which plays a role, on the one hand, for self awareness of behavioral reactions and, on the other hand, for whether such reactions are hard to notice for other persons.

Furthermore, we can ask whether features of behavior related to AES are hard or easy to control (which features are hard, which are easy to control) and, related to this, whether they are automatic or more intentional. We can also ask if the different behavioral features related to AES are separately controllable or interdependent. We intend the investigation reported below to contribute to improving our understanding of some of the above issues.

The paper is structured as follows: section 1 gives an introduction and background. Section 2 presents our method in terms of participants, material, procedure and analysis, section 3 presents our results concerning the sensory modality orientation of different AES and section 4 discusses the results. Section 5, finally, provides our conclusions.

2 Method

A study was carried out with the purpose of examining how affective-epistemic states (i.e. emotions, like "happy" or "sad" and more epistemic states, like "surprised") in a dialog were perceived when data was presented either unimodally as Video only and Audio only or multimodally in Video+Audio format (cf. also Lanzini, 2013).

2.1 Participants

There were 12 participants, 6 men and 6 women, all native speakers of Swedish, at least 20 years old.

2.2 Material

Three recordings of "First encounters" from the Swedish NOMCO database were presented to each subject. The NOMCO project (Multimodal Corpora for the Nordic Languages) 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 now available for research (Paggio et al. 2010). The "First encounter" interactions in the corpora were recorded in a studio setting, where the two participants were standing in front of a light background, so that automatic registration of body movements was 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) transcription and annotation tools. Functional annotation was mainly related to communicative feedback (Allwood, Nivre & Ahlsén 1992) and other interaction phenomena (cf. Paggio et al. 2010).

The clips from the recordings that were shown to the participants in the study were about 2 minutes long (with an average of 2 min, 7 sec).

Figure 1. Example of NOMCO "First encounter" recording

2.3 Procedure

The clips were presented to each subject individually. Each clip was presented in only one of the three modes: 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 try 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 session, the goal of the study was explained to the subjects and they were given instructions. The meaning of the term "affective-epistemic states" was briefly explained and the subjects were told to identify states of this sort (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), describe how it was expressed (i.e. through what behavioral trait), and provide a motivation. The subjects were free to choose their own words both for affective epistemic states and behavioral traits. The session of the experiment lasted around 75 minutes per subject. This procedure was chosen, not to disturb sequential contextual dependence between the points of invited interpretation. This would have been disturbed with a pre-selected presentation of shorter clips. A priority in the study was to ensure a high ecological validity, which requires the presentation of a context and non-artificial stimuli. (i.e. not to short, decontextualized or artificial). Our main interest was the contribution of the visual and auditory modalities to the interpretation of AES, whereas the contribution of specific features within each modality was not the main focus and is therefore mainly considered in the discussion section.

2.4 Analysis

The responses were transcribed and coded manually. Around 200 different labels and descriptions of the affective-epistemic states and the behavior used to express them were obtained from the participants in the study. Figure 2 shows some examples of responses in English translation.

Video + Audio 8602: 1.52-1.59

Participant 1		**Participant 2**		**Participant 3**	
Person left	Person right	Person left	Person right	Person left	Person right
Interested Asking questions, eye contact, gives feedback	**Relaxed** (moving more, moving hands)	**Enthusiastic, happy** Looking into the camera, smiling	**Glad** – Tone of voice, linght voice, smiling	**Happy** smiling	

Audio 8602: 1.52-1.59

Participant 4		**Participant 5**		**Participant 6**	
Person left	Person right	Person left	Person right	Person left	Person right
Interested Asking for more information	**Confident** – Loud voice	**Neutral** Just sharing information	**Neutral** Just sharing information	**Listening** – agreeing	**Confident** Loud voice No hesitations

Video 8602:

Participant 7		**Participant 8**		**Participant 9**	
Person left	Person right	Person left	Person right	Person left	Person right
Curious Smiling and pointing to the camera	**Nervous** Nodding, not smiling	**Nervous** Moving, Doesn't know what the other participant is talking about	**Secure,** listening Eye contact	**Relaxed,** interest-ed Moving foot and arms to one side, head forward	**Relaxed,** interested Arms to the side, Nodding

Figure 2. Examples of responses from 9 participants for one of the stimulus clips. (Attributed AES in boldface, behaviors in regular print.)

Using intuitive semantic analysis, the responses were then grouped into semantic fields of types of AES, by the authors, who after discussion arrived at consensus concerning which terms to group together for all semantic fields.

Examples of the semantic groupings of the AES responses are:

Happiness: contentment, gladness, joy and enjoyment in doing something.

Interest: curiosity, listening, seeking more information and paying attention to something.
Nervousness: uncomfortableness, insecurity, uneasiness, tension, embarrassment, shyness.
Confidence: relaxation, comfortableness, security.
Disinterest: indifference, being tired, bored or being annoyed about something.

The types of AES were then compared with regard to the three different conditions of presentation (Audio only, Video only and Video + Audio).

The obtained AES terms were the material for the semantic analysis, while the obtained behavioral terms were the basis for the summarizing discussion of specific AES in section 4.

3 Results

The results show differences concerning which affective-epistemic states were attributed to the different behaviors, depending on the mode of presentation. In this study, we focus on the seven most frequently attributed types of affective-epistemic states, which are: *happiness, interest, nervousness, confidence, disinterest, thoughtfulness and understanding*. The effect on the frequency of attribution connected with different affective-epistemic states in each of the presentation modes (audio, video or multimodal) is given in table 1.

Affective Epistemic States	**Video only**	**Audio only**	**Multimodal**	**Total**
Nervousness	80	73	142	305
Happiness	65	36	46	147
Interest	94	81	96	271
Confidence	82	103	75	250
Disinterest	8	24	19	50
Thoughtfulness	4	1	5	10
Understanding	6	1	2	9

Table 1. Effect of unimodal and multimodal presentation on the frequency of attribution of different affective-epistemic states.

The results show differences concerning the frequency with which affective-epistemic states were attributed in unimodal or multimodal mode.

When video images and sounds are perceived together, one sensory modality can affect the perception of the other modality. For example, an attribution of *happiness* was more often provided for laughing and smiling in the unimodal video mode than when they were presented in multimodal video + audio mode. Similarly, *confidence* was more frequently attributed in the unimodal audio mode than in the unimodal video or multimodal video + audio mode. In table 1 below, we present a summary of the ways in which unimodality and multimodality adds or decreases likelihood of attribution for different affective-epistemic states.

If we only consider the unimodal presentations, we see that only *confidence* and *disinterest* get a larger number of attributions in unimodal audio than in unimodal video, all the rest get more attributions in unimodal video.

For the two AES where unimodal video has most attributions (*happiness* and *understanding*), the number of attributions is lower for audio only than for multimodal presentation; this indicates that auditory information seems to filter out some of the attributions of the AES.

For *nervousness*, *interest* and *thoughtfulness*, multimodal presentation gets most attributions and video only second most: this, rather than a filtering role for audio information, indicates that audio and video mutually reinforce each other, leading to the highest number of attributions for multimodal presentation. This interaction is rather complex. Thus, for example, *nervousness* was more easily attributed in the multimodal video + audio mode than in the unimodal audio or video 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. In the case of *interest* and *thoughtfulness*, however, the total number of attributions was about the same in the video and video + audio modes, so in this case, the reinforcing effect of multimodality seems fairly weak.

For *confidence* and *disinterest*, unimodal audio presentation leads to the highest number of attributions, indicating that both these AES have a strong presence in audible features of speech. In the case of *disinterest*, multimodal presentation gets more attributions than unimodal video, indicating a filtering function for video, similar to the filtering function for audio noted above for happiness and understanding. In the case of *confidence*, however, unimodal video gets more attributions than multimodal presentation; this seems to indicate that the two modalities act as

restrictions on each other, thus a case of mutual restriction, rather than mutual reinforcement.

4 Discussion

Below, we now also consider the results with regard to attributed behaviors for the seven studied AES, in the form of a list that summarizes what specific auditory and visual cues we have found to influence interpretation concerning sensory orientation.

Nervousness
Respondents interpret this *AES* on the basis of both *audio* and *video* cues.
Auditory cues: Respondents say that *auditory cues*, like *tone of voice, prosody, hesitations* and *repetitions* have a central role in the interpretation of nervousness.
Visual cues: Respondents mention *avoidance of eye contact* and *frequent body movements* (hands, arms, feet) as important cues for nervousness.

Happiness
Respondents interpret this *AES more* on the basis of *video* cues than on *audio* cues. *Laughing* and *smiling* have a central role in the interpretation of happiness. We have regarded smiling as being visual and laughter as involving both visual and auditory cues.

Interest
Respondents seem to interpret this *AES* both on the basis of *auditory* and *visual* cues.
Auditory cues are *vocal feedback, asking questions* and *tone of voice.*
Visual cues are eye contact and nodding.

Confidence
Respondents interpret this *AES slightly more* on the basis of *audio* cues than on *video* cues.
Auditory cues are *showing verbal competence, laughing, tone* and *volume of voice.*
Visual cues are *posture, eye contact, laughing* and *smiling.*

Disinterest
Respondents seem to interpret this *AES* much more on the basis of *audio* cues than on *video* cues.
Auditory cues are *tone of voice, prosody* and *giving vocal feedback*

Thoughtfulness
Respondents attribute this AES more on the basis of *video* cues than *audio* cues, Thoughtfulness seems to be strongly related to specific visible movements, namely *looking up* and *looking away*.

Understanding
Respondents attribute this *AES* more on the basis of *video* than *audio* cues.
Visual cues are *smiling, looking around, looking into the camera* and *nodding.*

The behavioral descriptions, by and large, reinforce the impression given by the interpretation and attribution data reported in section 3. The mode of presentation connected with most attributions also receives the most well articulated behavioral descriptions. The AES have a differentiated orientation to the two sensory modalities that are investigated in this study.

This, in turn, can be related to one of the main findings of our study, namely that multimodality does not simply provide redundant information that merely reinforces information that is given unimodally. Rather the multimodal interaction between sensory modalities is more complex, sometimes restricting and filtering and sometimes complementing and adding to unimodally given information.

5 Conclusions

In summary, we have found that two AES, *happiness* and *understanding*, have a visual orientation and that the addition of audio information has a filtering effect when they are presented multimodally. Two AES, *confidence* and *disinterest*, have an auditory orientation, where for *disinterest*, but not for *confidence*, the video mode has a filtering function. Three AES, *nervousness, interest* and *thoughtfulness,* are most frequently attributed in multimodal presentation mode and audio and video cues here seem to have a mutually reinforcing role. In contrast, in one case – *confidence* – audio and video cues seem to have a mutually restrictive role.

Multimodal interaction between sensory modalities is complex, sometimes subtracting and sometimes complementing and adding to information given unimodally. It does not simply provide redundant information that merely reinforces unimodal information.

The role of the sensory modalities vision and hearing is, thus, not the same for all AES. This means that each state has to be studied with regard to what specific sensory cues are important concerning its specific sensory orientation.

Acknowledgments

The research that has led to this work has been supported by the NOMCO project, which is funded by the NORDCORP program under the Nordic Research Councils for the Humanities and the Social Sciences (NOS-HS) and The SSPNet (European Community's Seventh Framework Programme (FP7/2007-2013), under grant agreement no. 231287).

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.

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.

Cohn, Jeffrey F., & De la Torre, Fernando. (In press). Automated face analysis for affective computing. In Calvo, R.A., D'Mello, S.K, Gratch, J. & Kappas, A. (Eds.), Handbook of affective computing. New York, NY: Oxford.
Inget av detta är egentligen vad vi gör.

Cunningham, D. W., Kleiner, M., Vallraven C. & Bülthoff, H. H. (2005). Manipulating video sequences to determine the components of conversational facial expressions. *ACM Transactions on Applied Perception (TAP)* Volume 2 Issue 3, July 05:251 - 269

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.iscaspeech.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.

Vinciarelli, A. Pantic, M., Heylen, D., Pelachaud, C.,Poggi, I. D'Errico, F. & Schroeder, M. (2012). Bridging the gap between social animal and unsocial machine: a survey of Social Signal Processing, *IEEE Transactions on Affective Computing,* Vol. 3:1, pp. 69-87.

Investigating Non-Verbal Behaviors Conveying Interpersonal Stances

Mathieu Chollet[1], **Magalie Ochs**[2] **and Catherine Pelachaud**[2]
[1]Institut Telecom; Telecom Paristech; CNRS-LTCI
[2]CNRS LTCI; Telecom ParisTech
37 rue Dareau, 75014, Paris, France
{mchollet,ochs,pelachaud}@telecom-paristech.fr

Abstract

Interpersonal stances are expressed by non-verbal behaviors on a variety of different modalities. The perception of these behaviors is influenced by the context of the interaction, how they are sequenced with other behaviors from the same person and behaviors from other interactants. In this paper, we introduce a framework considering the expressions of stances on different layers during an interaction. This framework enables one to reason on the non-verbal signals that an Embodied Conversational Agent should express to convey different stances. To identify more precisely humans' non-verbal signals conveying dominance and friendliness attitudes, we propose in this paper a methodology to automatically extract the sequences of non-verbal signals conveying stances. The methodology is illustrated on an annotated corpus of job interviews.

Keywords

Interpersonal stance; Non-verbal behaviors; Sequence mining

1 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, that can be defined as "*spontaneous or strategically employed affective styles that colour interpersonal exchanges* (Scherer, 2005)". Our goal is to find out how interpersonal stances are expressed through non-verbal behavior, and to implement the expression of interpersonal stances in an ECA.

Most modalities of the body are involved when conveying interpersonal stances (Burgoon et al., 1984). Smiles can be signs of friendliness (Burgoon et al., 1984), performing large gestures may be a sign of dominance, and a head directed upwards can be interpreted with a dominant stance (Carney et al., 2005). A common representation for interpersonal stance is Argyle's bi-dimensional model of attitudes (Argyle, 1988), with an affiliation dimension ranging from hostile to friendly, and a status dimension ranging from submissive to dominant (see Figure 1).

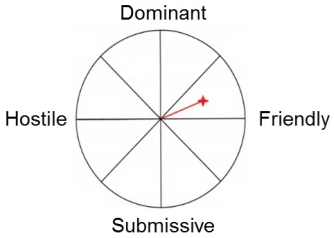

Figure 1: The Interpersonal Circumplex, with Argyle's attitude dimensions. The sample coordinate represents a friendly and slightly dominant interpersonal stance.

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 (Burgoon et al., 1984); however, a smile followed by a gaze and head aversion conveys embarassment (Keltner, 1995). 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

[1]http://http://www.tardis-project.eu/

can convey friendliness (LaFrance, 1982). Finally, the global behavior tendencies of a person, such as performing large gestures in general, are important when interpreting their stance (Escalera et al., 2010). These different perspectives have seldom been studied together, and this motivates the use of multimodal corpora of interpersonal interactions in order to analyze their impact in a systematic fashion.

We propose a model for non-verbal behavior analysis, composed of multiple layers analyzing a particular perspective of non-verbal behavior interpretation on time windows of different lengths. To build this model, we annotated a corpus of job interview enactment videos with non-verbal behavior annotations and interpersonal stance annotations. In this paper, we focus on a layer of the model which deals with how sequences of non-verbal signals displayed while speaking can be interpreted as the expression of dominance and friendliness stances. While it has been proved that the sequencing of non-verbal signals influences how they are perceived (With and Kaiser, 2011), the literature on the topic is limited. To gather knowledge about this layer, we use a data mining technique to extract sequences of non-verbal signals from the corpus.

The paper is organized as follows. In section 2, we present related models of interpersonal stances for ECAs and their limits. We then introduce our multi-layer model. Section 4 describes the multimodal corpus and how it was annotated. Section 5 details a data mining method we propose to gather knowledge about how sequences of non-verbal behavior are perceived.

2 Related work

Models of interpersonal stances expression for virtual agents have already been proposed. For instance, in (Ballin et al., 2004), postures corresponding to a given stance were automatically generated for a dyad of agents. Lee and Marsella used Argyle's attitude dimensions, along with other factors such as conversational roles and communicative acts, to analyze and model behaviors of side participants and bystanders (Lee and Marsella, 2011). Cafaro *et al.* (Cafaro et al., 2012) conducted a study on how smile, gaze and proximity cues displayed by an agent influence the first impressions that the users form on the agent's interpersonal stance and personality. Ravenet *et al.* (Ravenet et al., 2013) proposed a user-created corpus-based methodology for choosing the behaviors of an agent conveying a stance along with a communicative intention. These models, however, only consider the expression of a few signals at a given time, and do not consider longer time spans or sequencing of signals.

Other works have gone further by also considering global behavior tendencies and reactions to the interactants' behaviors: the *Laura* agent (Bickmore and Picard, 2005) was used to develop long term relationships with users, and would adapt the frequency of gestures and facial signals as the relationship with the user grew. However, dominance was not investigated, and the users' behaviors were not taken into account as they used a menu-based interface. Prepin *et al.* (Prepin et al., 2013) have investigated how smile alignment and synchronisation can contribute to stance building in a dyad of agents. Although not directly related to dominance or friendliness, Sensitive Artificial Listeners designed in the Semaine project (Bevacqua et al., 2012) produce feedback and backchannels depending of the personality of an agent, defined by extraversion and emotional stability.

Even though different perspectives of interpretation of non-verbal behavior we mentioned have been integrated in models of ECAs, the existing models of interpersonal stances expression consider only consider one perspective at a time, with a limited number of modalities. Moreover, no model of stance expression seems to consider how non-verbal signals are sequenced. In the next section, we present a theoretical model to the integration of these different perspectives.

3 A multi-layer approach to the expression of interpersonal stances

In (Chollet et al., 2012), we defined a multi-layer model to encompass the different non-verbal behavior interpretation perspectives (See figure 2). The *Signal* layer looks at the interpretation of signals in terms of communicative intentions (*e.g.* a hand wave means greeting someone). In the *Sentence* layer, we analyze the sequence of signals happening in a dialogue turn (*e.g.* a smile followed by a head aversion means embarrassment). The *Topic* layer focuses on the inter-personal behavior patterns and tendencies (*e.g.* adopting the

same posture as the interlocutor is a sign of friendliness). Finally, the *Interaction* layer encompasses the whole interaction and looks at global behavior tendencies (*e.g.* smiling often is a sign of friendliness). 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.

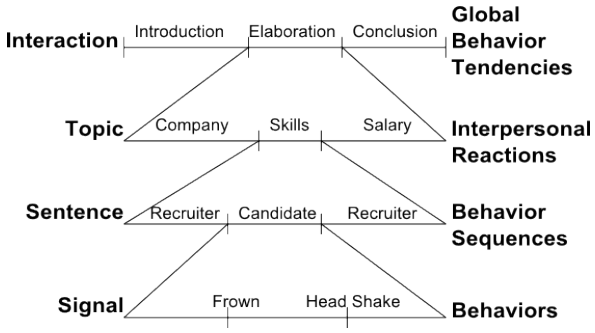

Figure 2: This figure illustrates the multi-layer model in a job interview setting. On the left are represented the layers of the model, and on the right which behavioral features they analyze.

Here is an example of how the different layers work in an interaction. Imagine a recruiter who is annoyed by a candidate because he thinks his foreign language skills do not meet the requirements for a job. The recruiter spreads his right hand towards the candidate while asking the question "You claim to be proficient in English. Can you prove it to me?". The candidate looks down for a while, thinking and hesitating. He looks up at the recruiter and tries an answer with a faint smile, then moving his head to the side. While the candidate is speaking, the recruiter frowns, and then shakes his head as the candidate finishes. All this time, the recruiter kept looking at the candidate.

In the example, the gesture performed by the recruiter is used to show a question is asked and that he gives the speaking floor to the candidate. These two communicative functions are handled by the *Signal* layer. When replying, the candidate smiles and then averts his head away from the recruiter. In that case, the *Sentence* layer considers the sequencing of signals: the smile could have been interpreted as a sign of friendliness at first, however followed by a head aversion it is a sign of submissiveness. The recruiter behavioral replies to the candidate's answer, the frown and head shake, are analyzed by the *Topic* layer as sign of dominance and hostility. Finally, the fact that the recruiter barely averted gaze during the interaction is a sign of *dominance* revealed by the *Interaction* layer.

In order to build a model for each layer, our approach consists of automatically extracting knowledge from a multimodal corpus of interactions during which interpersonal stances are expressed. In this paper, we focus on the *Sentence* layer: it is known that the sequencing of non-verbal signals influence how these behaviors are perceived (With and Kaiser, 2011), however since relatively little accounts exist on this phenomenon, automated methods of knowledge extraction are particularly relevant for this layer. In the next section, we present our multimodal corpus and its annotation process.

4 Multimodal corpus of interpersonal stance expression

As part of the TARDIS project, a study was conducted with practitioners and youngsters from the Mission Locale Val d'Oise Est, a French job coaching association. The study consisted in creating a situation of job interviews between 5 practitioners and 9 youngsters. The setting was the same in all videos (see Figure 3). The recruiter and the youngster sat on each side of a table. A single camera embracing the whole scene recorded the dyad from the side. From this study was gathered a corpus of 9 videos of job interview lasting approximately 20 minutes each. We decided to use these videos to investigate the sequences of non-verbal signals the recruiters use when conveying interpersonal stances. In order to study how recruiters express interpersonal stances, 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.

Numerous coding schemes exist to annotate non-verbal behavior in multimodal corpora. A widely used system for facial expressions is the Facial Action Coding System (Ekman and Friesen, 1977). A very exhaustive coding scheme for multimodal behavior is the MUMIN multimodal coding scheme, that was used for the analysis of turn-taking and feedback mechanisms (Allwood et al., 2007). For the non-verbal behavior annotation, we adapted the MUMIN multimodal coding scheme

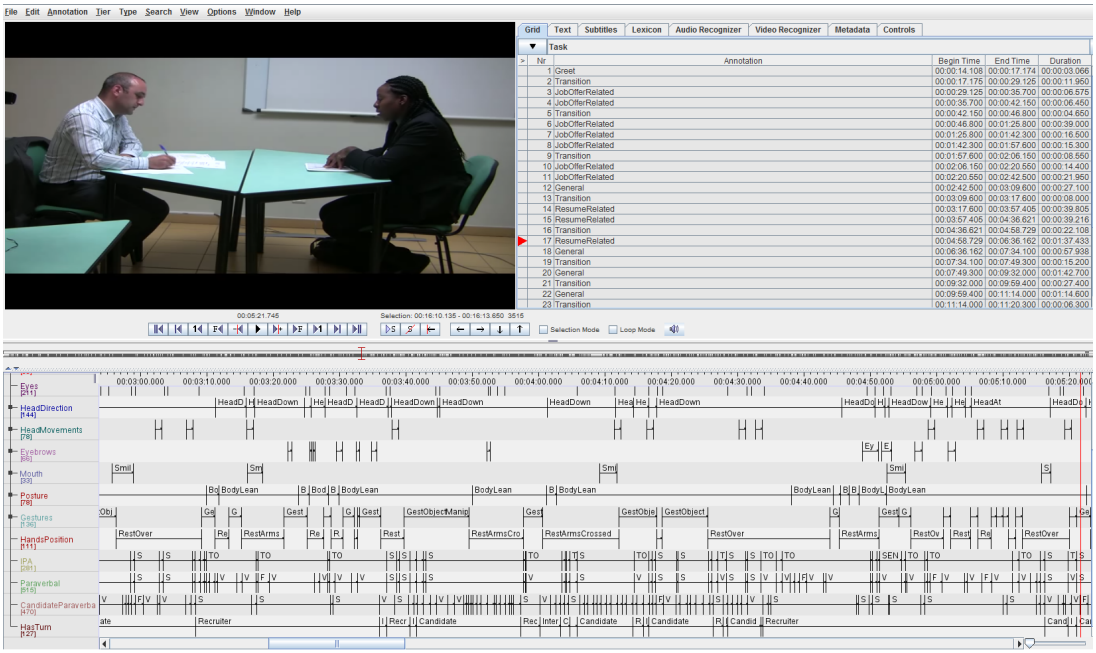

Figure 3: Video of the study in the Elan (Wittenburg et al., 2006) annotation environment

to our task and our corpus (*e.g.* by removing any types of annotations we cannot extract from the videos, such as subtle facial expressions). We used Praat (Boersma and Weenink, 2001) for the annotation of the audio stream and the Elan annotation tool (Wittenburg et al., 2006) for the visual annotations. A single annotator annotated the three videos. To measure the reliability of the coding, three minutes of video were randomly chosen and annotated a second time one month after the first annotation effort, and we computed Cohen's kappa score between the two annotations. It was found to be satisfactory for all modalities ($\kappa >= 0.70$), except for the eyebrow movements ($\kappa >= 0.62$), which low score can be explained by the high camera-dyad distance making detection difficult. The highest scores were for gaze ($\kappa >= 0.95$), posture ($\kappa >= 0.93$) and gestures ($\kappa >= 0.80$). This annotation processes amounted to 8012 annotations for the 3 videos. The para-verbal category has the highest count of annotations, between 483 to 1088 per video. On non-verbal annotations, there were 836 annotations of gaze direction, 658 head directions, 313 gestures, 281 head movements, 245 hands positions, 156 eyebrow movements and 91 smiles. Important differences in behavior tendencies exist between recruiters: for instance the first recruiter performed many posture shifts: 5.6 per minute, to compare with 2.2 for the second recruiter and 0.6 for the third one. The second recruiter smiles much less than the others: 0.4 smiles per minute versus 2.4 per minute for both the first and third recruiters.

As the interpersonal stance of the recruiters varies through the videos, we chose to use GTrace, successor to FeelTrace (Cowie et al., 2011). GTrace is a tool that allows for the annotation of continuous dimensions over time. Users have control over a cursor displayed on an appropriate scale alongside a playing video. The position of the cursor is sampled over time, and the resulting sequence of cursor positions is known as trace data. We adapted the software for the interpersonal stance dimensions we considered. Though the software allows for the annotation of two dimensions at a time using a bi-dimensional space, we constrained it to a single dimension to make the annotation task slightly easier. We asked 12 persons to annotate the videos. Each annotator had the task of annotating one dimension for one video, though some volunteered to annotate more videos. As the videos are quite long, we allowed them to pause whenever they felt the need to. With this process, we collected two to three annotation files per attitude dimension per video. While evaluating inter-rater agreement is a simple task when analyzing discrete labels (*e.g.* two people assign the same class to an item), it is not as straight-

forward when dealing with trace data (Metallinou and Narayanan, 2013), though recently new approaches to this problem have been proposed (Cowie and McKeown, 2010). Similarly to previous experiments on trace data annotation of emotions, we found that raters agreed more on how attitude values varied (*i.e.* when attitudes raises or falls), than on actual absolute values.

Similarly to (Cowie and McKeown, 2010), we averaged attitude values in bins of 3 seconds. We then computed the reliability of different annotations by computing Cronbach's α, using the variation values from one bin to the next. Cronbach's α value was found to be generally average ($\alpha = 0.489$), with the highest video scoring $\alpha = 0.646$. We believe these values to be acceptable for our purposes, considering Cronbach's α is likely to produce lower scores on annotations continuous both in time and in value, and that the sequence mining process we propose (described in Section 5) provides a natural way of discarding the time segments where annotators where not agreeing. Indeed the non-verbal signals sequences contained in these segments will be distributed for different types of attitude variations, and therefore will not be very frequent before any particular attitude variation. However, the sequence mining algorithm we use relies on frequence to extract meaningful non-verbal signals sequence, which means that the time segments where annotators do not agree will not contribute to making some non-verbal signals sequences more frequent.

In a nutshell, the corpus has been annotated at two levels: the non-verbal behavior of the recruiters and their expressed stances. Our next step was to identify the correlations between the non-verbal behaviors and the interpersonal stances. As a first step, we have focused on the non-verbal signals sequences expressed by the recruiters when they are speaking (*i.e.* at the *Sentence* level, Section 3). In the next section, we describe a method for extracting knowledge about non-verbal behavior sequences from the multimodal corpora.

5 Investigating non-verbal behavior sequences

A number of tools and techniques exist for the systematic analysis of sequences of events in sequential data. Traditional sequence analysis (Bakeman and Quera, 2011) techniques typically revolve around the computation of simple contigency tables measuring the occurence of one type event of event after another one. Such methods are not well suited to longer sequences of events (*i.e.* made of more than 2 events) and to cases where noise can happen (*i.e.* behaviors unrelevant to a particular sequence that can happen in the middle of it). Magnusson proposed the concept of *T-patterns* (Magnusson, 2000), sequences of events occuring in the same order with "relatively invariant" temporal patterns between events. The THEME software automatically detects *T-patterns* and was used in (With and Kaiser, 2011) to detect characteristic sequences of signals for emotion expression. Finally, *sequence mining* techniques have been widely used in task such as protein classification (Ferreira and Azevedo, 2005), and recent work has used this technique to find sequences correlated with video game players' emotions such as frustration (Martínez and Yannakakis, 2011).

In order to extract significant sequences of non-verbal signals conveying interpersonal stances from our corpus, we use a *frequent sequence mining* technique. To the best of our knowledge, this technique has not yet been applied to analyse sequences of non-verbal signals. In the following part, we describe the procedure used to mine frequent sequences in our corpus, and we then describe the result of applying this procedure on our data.

5.1 Applying sequence mining to our multimodal corpus

To apply the frequent sequence mining technique to our data, we proceed through the following six steps.

The first step consists of parsing the non-verbal annotations files, coded in the ELAN format, filtering the annotation modalities and time segments to investigate (*e.g.* we only consider here behavior sequences while speaking, therefore we discard the segments when the recruiter is listening) and converting every interaction's annotations into a list containing all the non-verbal behaviors in a sequence.

The second step's objective is to find events to segment the interactions: indeed, frequent sequence mining techniques require a dataset of sequences. In our case, our data consists of 3 continuous interactions. Since we investigate which

sequences of signals convey stances, we decide to segment the full interactions with attitude variation events: *attitude variation events* are the timestamps where an attitude dimension begins to vary. To this end, we parse the attitude annotations files, smoothe them and find the timestamps where the annotated attitude dimension starts to vary. More details can be found in (Chollet et al., 2013).

We found that the attitude variation events in our data came with a wide range of values, *i.e.* in some cases the annotators moved the cursor a lot, indicating he annotators perceived a strong change in the recruiters' stance from the recruiter's behavior, while sometimes the cursor movements were more subtle. We chose to differentiate between small and strong attitude dimension variations, therefore we used a clustering technique to identify the 4 clusters corresponding to small increases, strong increases, small decreases and strong decreases. To this end, we used a K-means clustering algorithm with $k = 4$.

The fourth step consists of segmenting the full interaction sequences with the attitude variations events obtained from step 2. Following this procedure, we obtain 219 segments preceding dominance variations and 245 preceding friendliness variations. We found dominance segments to be longer in duration, averaging at 12.7 seconds against 8.3 for friendliness segments. These two sets are split further depending on which cluster the attitude variation event belongs to. For instance, we have 79 segments leading to a large drop in friendliness, and 45 segments leading to a large increase in friendliness (see Table 1).

Step five consists of applying the frequent sequence mining algorithm to each set of segments. We used the commonly used Generalized Sequence Pattern (GSP) frequent sequence mining algorithm described in (Srikant and Agrawal, 1996). The GSP algorithm requires as an input a minimum support, *i.e.* the minimal number of times that a sequence has to be present to be considered frequent, and its output is a set of sequences along with their support. For instance, using a minimum support of 3, every sequence that is present at least 3 times in the data will be extracted. The GSP algorithm based on the *Apriori* algorithm (Agrawal and Srikant, 1994): first, it identifies the frequent individual items in the data and then extends them into larger sequences iteratively, pruning out the sequences that are not frequent enough anymore.

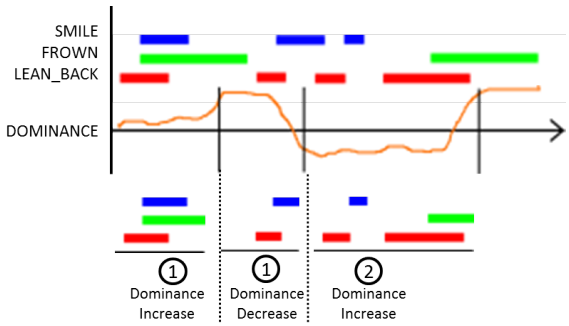

Figure 4: Step 1 through 4 consist of preprocessing the data before performing sequence mining. Attitude variations events are detected and used to segment the non-verbal behavior stream. The result is a set of non-verbal behavior segments for each type of attitude variation event.

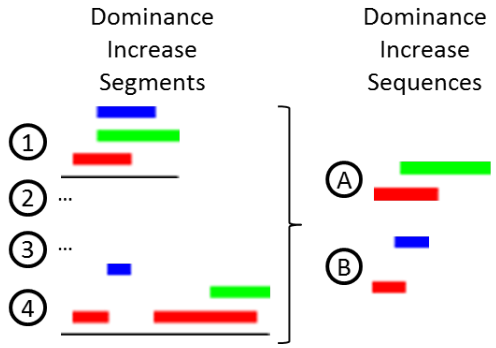

Figure 5: This figure illustrates the data mining process. All the segments for a given type of attitude variation event (here, an increase in dominance) are gathered. The result of the *GSP* algorithm is the set of sequences along with their support

However, the support is an insufficient measure to analyse how a sequence is characteristic of a type of attitude variation event. For instance, having the gaze move away and back to the interlocutor happens very regularly in an interaction. Thus it will happen very often before all types of attitude variation events (*i.e.* it will have a high support), even though it is not sure that it characteristic of any of them. The objective of step 6 is to compute *quality measures* to assess whether a sequence is really characteristic of a type of attitude variation events. Based on (Tan et al., 2005), we choose to compute *confidence* and *lift* quality mea-

Variation type	Cluster Center	Segment Count	Frequent Sequences
Friendliness Large Increase	0.34	68	86
Friendliness Small Increase	0.12	66	72
Friendliness Small Decrease	-0.11	77	104
Friendliness Large Decrease	-0.32	36	67
Friendliness Total		247	329
Dominance Large Increase	0.23	49	141
Dominance Small Increase	0.09	66	244
Dominance Small Decrease	-0.13	80	134
Dominance Large Decrease	-0.34	24	361
Dominance Total		219	879

Table 1: Description of results for each attitude variation type

sures for every sequence. The confidence represents how frequently a sequence is found before a particular type of attitude variation event. The lift represents how more frequently the sequence occurs before a type of attitude variation event than in other cases (the higher the value, the more likely it is that there is dependence between the sequence and the attitude variation).

In the next part, we describe the sequences we extracted when applying this procedure to our corpus.

5.2 Results

As a first step, we wanted to get a glimpse at which kinds of non-verbal signals were more frequent in the extracted sequences of the different attitude variation types. For this purpose, we performed Student T-tests, comparing the number of occurrences of each signal type for the different types of attitude variations. Note that this is not meant as a complete analysis of the extracted sequences, but rather as an exploration of the types of signals most present in these sequences.

We found smiles to be significantly more common before large increases in friendliness than in all other cases (Small increase: $p = 0.005 < 0.05$, small decreases $p = 0.001 < 0.05$, large decreases $p = 0.011 < 0.05$). Head nods happened significantly more often before large increases in friendliness than large decreases ($p = 0.026 < 0.05$). The same was found for head shakes, which appeared more before large increases in friendliness than small decreases ($p = 0.023 < 0.05$) or large decreases ($p = 0.024 < 0.05$). Leaning towards the candidate was found to be more common before small increases in dominance than large decreases ($p = 0.013 < 0.05$). Similarly, adopting a straight posture was more common before small increases in dominance, compared to small decreases ($p = 0.040 < 0.05$) and large decreases ($p = 0.001 < 0.05$). A head averted sideways was found to be more common before small increases in dominance than before large decreases ($p = 0.019 < 0.05$). The same was found for crossing the arms ($p = 0.044 < 0.05$).

To obtain a reasonable number of potentially relevant sequences, we have chosen to only identify the sequences present in our corpus at least 10 times (using a large minimum support would yield very few sequences, while a small minimum support would yield a very large number of sequences). The output of the GSP algorithm with a minimal support of 10 occurrences is a set of 879 sequences for dominance variations, and a set of 329 sequences for friendliness variations (see table 1). In average we found friendliness sequences to contain 2,91 signals, and dominance sequences to contain 3,58 signals.

In table 2 we show the top scoring (*i.e.* highest *Lift* score) extracted sequences for every attitude variation type found using this process. The Sup column corresponds to the support of the sequence and the $Conf$ column to the confidence of the sequence. We have integrated the extracted sequences in an animation module for our ECA platform. Our next step consists of conducting user perceptive tests to validate that the sequences displayed by the virtual agent convey the expected attitude.

6 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 stances expression and we annotated videos of job interview enactments. We presented a knowledge extraction method for non-verbal behavior sequences based on a data mining technique. Our future work consists of validating that the extracted sequences convey the appropriate interpersonal stance when

Sequence	Attitude Variation	Sup	$Conf$	$Lift$
BodyStraight -> HeadDown	Friendliness Large Decrease	0.016	0.4	2.74
HeadDown -> HeadAt -> GestComm -> HandsTogether	Friendliness Small Decrease	0.032	0.72	2.33
HeadAt -> HeadSide	Friendliness Small Increase	0.028	0.54	2.02
Smile	Friendliness Large Increase	0.061	0.52	1.88
GestComm -> HeadDown -> HeadAt -> HeadDown	Dominance Large Decrease	0.028	0.42	3.80
HeadDown -> HeadAt -> HeadDown -> HandsTogether	Dominance Small Decrease	0.041	0.75	2.05
HeadAt -> ObjectManipulation -> HandsOverTable	Dominance Small Increase	0.037	0.67	2.21
HeadDown -> EyebrowUp	Dominance Large Increase	0.022	0.45	2.03

Table 2: Top scoring sequences for each attitude variation event

expressed by a virtual agent.

Acknowledgment

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

References

R. Agrawal and R. Srikant. 1994. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

J. Allwood, S. Kopp, K. Grammer, E. Ahlsen, E. Oberzaucher, and M Koppensteiner. 2007. The analysis of embodied communicative feedback in multimodal corpora: a prerequisite for behavior simulation. *Language Resources and Evaluation*, 41:255–272.

M. Argyle. 1988. *Bodily Communication*. University paperbacks. Methuen.

R. Bakeman and V. Quera. 2011. *Sequential Analysis and Observational Methods for the Behavioral Sciences*. Cambridge University Press.

D. Ballin, M. Gillies, and B. Crabtree. 2004. A framework for interpersonal attitude and non-verbal communication in improvisational visual media production. In *1st European Conference on Visual Media Production*.

E. Bevacqua, E. Sevin, S. J. Hyniewska, and C. Pelachaud. 2012. A listener model: introducing personality traits. *Journal on Multimodal User Interfaces*, 6(1-2):27–38.

T. W. Bickmore and R. W. Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput.-Hum. Interact.*, 12(2):293–327, June.

P. Boersma and D. Weenink. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

J. K. Burgoon, D. B. Buller, J. L. Hale, and M. A. de Turck. 1984. Relational Messages Associated with Nonverbal Behaviors. *Human Communication Research*, 10(3):351–378.

A. Cafaro, H. H. Vilhjálmsson, T. Bickmore, D. Heylen, K. R. Jóhannsdóttir, and G. S. Valgarðsson. 2012. First impressions: users' judgments of virtual agents' personality and interpersonal attitude in first encounters. In *Proceedings of the 12th international conference on Intelligent Virtual Agents*, IVA'12, pages 67–80, Berlin, Heidelberg. Springer-Verlag.

D. R. Carney, J. A. Hall, and L. S. LeBeau. 2005. Beliefs about the nonverbal expression of social power. *Journal of Nonverbal Behavior*, 29(2):105–123.

M. Chollet, M. Ochs, and C. Pelachaud. 2012. Interpersonal stance recognition using non-verbal signals on several time windows. In *Workshop Affect, Compagnon Artificiel, Interaction*.

M. Chollet, M. Ochs, and C. Pelachaud. 2013. A multimodal corpus approach to the design of virtual recruiters. In *Workshop Multimodal Corpora, Intelligent Virtual Agents*, pages 36–41.

R. Cowie and G. McKeown. 2010. Statistical analysis of data from initial labelled database and recommendations for an economical coding scheme.

R. Cowie, C. Cox, J.-C. Martin, A. Batliner, D. Heylen, and K. Karpouzis. 2011. *Issues in Data Labelling*. Springer-Verlag Berlin Heidelberg.

P. Ekman and V. Friesen. 1977. *Manual for the Facial Action Coding System*. Palo Alto: Consulting Psychologists Press.

S. Escalera, O. Pujol, P. Radeva, J. Vitria, and M. Anguera. 2010. Automatic detection of dominance and expected interest. *EURASIP Journal on Advances in Signal Processing*, 2010(1):12.

P. G. Ferreira and P. J. Azevedo. 2005. Protein sequence classification through relevant sequence mining and bayes classifiers. *Progress in Artificial Intelligence*, 3808:236–247.

D. Keltner. 1995. Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of Personality and Social Psychology*, 68:441–454.

M. LaFrance. 1982. Posture mirroring and rapport. In M. Davis, editor, *Interaction Rhythms: Periodicity in Communicative Behavior*, pages 279–299. New York: Human Sciences Press.

J. Lee and S. Marsella. 2011. Modeling side participants and bystanders: The importance of being a laugh track. In *Proceedings of the 10th International Conference on Intelligent Virtual Agents*, IVA'11, pages 240–247, Berlin, Heidelberg. Springer-Verlag.

M. S. Magnusson. 2000. Discovering hidden time patterns in behavior: T-patterns and their detection. *Behavior Research Methods, Instruments, Computers*, 32:93–110.

H. P. Martínez and G. N. Yannakakis. 2011. Mining multimodal sequential patterns: a case study on affect detection. In *Proceedings of the 13th international conference on multimodal interfaces*, ICMI '11, pages 3–10, New York, NY, USA. ACM.

A. Metallinou and S. Narayanan. 2013. Annotation and processing of continuous emotional attributes: Challenges and opportunities. In *Automatic Face and Gesture Recognition*, pages 1–8.

K. Prepin, M. Ochs, and C. Pelachaud. 2013. Beyond backchannels: co-construction of dyadic stancce by reciprocal reinforcement of smiles between virtual agents. In *International Conference CogSci (Annual Conference of the Cognitive Science Society)*.

B. Ravenet, M. Ochs, and C. Pelachaud. 2013. From a user-created corpus of virtual agent's non-verbal behaviour to a computational model of interpersonal attitudes. In *International Conference on Intelligent Virtual Agent (IVA2013)*.

K. R. Scherer. 2005. What are emotions? and how can they be measured? *Social Science Information*, 44:695–729.

R. Srikant and R. Agrawal. 1996. Mining sequential patterns: Generalizations and performance improvements. *Advances in Database Technology*, 1057:1–17.

P. Tan, M. Steinbach, and V. Kumar. 2005. *Introduction to Data Mining,(First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

S. With and W. S. Kaiser. 2011. Sequential patterning of facial actions in the production and perception of emotional expressions. *Swiss Journal of Psychology*, 70(4):241–252.

P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. 2006. Elan: a professional framework for multimodality research. In *In Proceedings of Language Resources and Evaluation Conference (LREC)*.

Exploring Sounded and Silent Laughter in Multiparty Task-based and Social 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

Abstract

We report on our explorations of laughter in multiparty spoken interaction. Laughter is universally observed in human interaction. It is multimodal in nature: a stereotyped exhalation from the mouth in conjunction with rhythmic head and body movement. Predominantly occurring in company rather than solo, it is believed to aid social bonding. Spoken interaction is widely studied through corpus analysis, often concentrating on 'task-based' interactions such as information gap activities and real or staged business meetings. Task-based interactions rely heavily on verbal information exchange while the immediate task in natural conversation is social bonding. We investigate laughter *in situ* in task-based and social interaction, using corpora of non-scripted (spontaneous) multiparty interaction: the task-oriented AMI meetings corpus, and the conversational TableTalk, and D-ANS corpora. We outline extension of previous work on laughter and topic change, describe the collection of natural social spoken interaction in the D-ANS corpus including audio-visual and biosignals, and describe an annotation experiment on multimodal aspects of laughter. We discuss the results and signal current and future research directions.

Keywords: laughter, conversation analysis, social conversation, multimodal communication

1 Introduction

In our work, we investigate natural face-to-face social and task-based spoken interaction. 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. From this information, the interlocutors make inferences on the speaker's literal and pragmatic meaning, intentions, and affective state. This paper reports some of our recent work on laughter, an intrinsic thread of the conversational information bundle. Below we outline the background to our work, describe the corpora we use in our investigations, report problems encountered with laughter annotations in earlier work and describe an experiment in the multimodal annotation of laughter and the replication of an earlier study on laughter around topic change.

1.1 Task-based and Social Conversation

Communication is situated, and its characteristics vary with the type of interaction or 'speech-exchange system' (Sacks, Schegloff, and Jefferson 1974) participants are engaged in. While types or genres of written communication are fairly clear, categorization is not regarded as straightforward for speech genres. The range of interactions involving speech which humans engage in is enormous, with the problem of categorizing different types of speech exchange

into genres labeled as 'notorious' (Bakhtine 1986). Although there are many genres of spoken interaction, much attention has been paid to 'task-based' or 'transactional' dialogue. Indeed, dialogue system technology is based on task-based dialogue (Allen et al. 2001) for reasons of tractability. However, in real-life conversation there is often no obvious short term task to be accomplished through speech and the purpose of the interaction is better described as a short or longer term goal of building and maintaining social bonds (Malinowski 1923; Dunbar 1998). In these situations, the transfer of verbal information is not the main goal, and non-verbal information may carry more importance in attaining the goals of social bonding. Conversation is 'interactional' rather than 'transactional'. As an example, a tenant's short chat about the weather with the concierge of an apartment block is not intended to transfer important meteorological data but rather to build a relationship which may serve either of the participants in the future. The study of different types of interaction, and indeed stretches of social and task-based communication within the same interaction sessions, can help discriminate which phenomena in one type of speech exchange system are generalizable and which are situation or genre dependent. In addition to extending understanding of human communication, such knowledge will be useful in human machine interaction design, particularly in the field of 'companion' robots or 'relational agents', as the very notion of a companion application entails understanding of social spoken interaction (Gilmartin and Campbell 2014). 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.

1.2 Laughter in Spoken Interaction

Laughter is a universally observed element of human interaction, part of the gesture call system (Burling 2007), appearing in human behavior before language both in evolutionary terms and in child development, and believed to have evolved from the primate social bonding (Glenn 2003). Laughter is predominantly a shared rather than solo activity, with most incidences reported as occurring in the company of others (Provine 2004). Laughter occurs widely in spoken interaction, but punctuates rather than interrupts speech (Provine 1993). Laughter episodes take a range of forms – from loud bouts to short, often quiet chuckles. Laughter is described as a stereotyped exhalation of air from the mouth in conjunction with rhythmic head and body movement (Mehu and Dunbar 2008), and is thus multimodal, with visual and acoustic elements available to the interlocutor's senses.

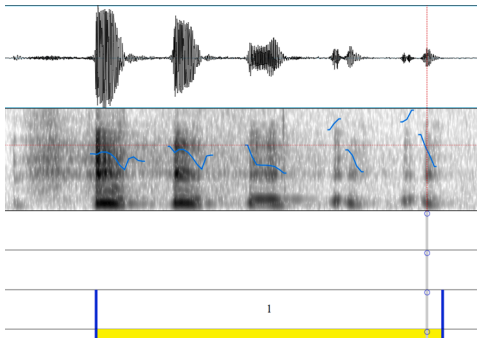

Figure 1 Stereotypical laughter from D-ANS, with alternation of aspiration and vocalic sounds.

Laughter itself is not clearly defined as a discrete activity in common parlance as evidenced by the plethora of terms used in English to describe it – from short shrieks, through snorts, chuckles, and peals. In the scientific literature, there is also a wide range of terminology although there is broad consensus that a 'stereotypical' laugh consists of breathy aspirations alternating with vocalic segments of decreasing power (Bachorowski and Owren 1995; Black 1984); an example of such laughter from the D-ANS corpus is shown in Figure 1. However, while this stereotypical laugh is generally produced upon asking an informant to laugh, it has been shown to be only one of several manifestations of laughter present in social interaction, and often not the most prevalent (Trouvain 2003). Also, laughter can occur within speech rather than around it, although these speech laughs, and smiling speech are not regarded as laughter by all researchers. In our investigations we study

laughter in spoken interaction using multimodal corpora of non-scripted (spontaneous) multiparty interaction: the task-oriented AMI meetings corpus (McCowan et al. 2005), and the conversational TableTalk (Jokinen 2009), and D-ANS (Hennig, Chellali, and Campbell 2014) corpora.

1.3 Laughter and topic change

Previous work explored relationships between laughter and topic change in the AMI and TableTalk corpora (Bonin, Campbell, and Vogel 2012), showing greater incidence of laughter in the social TableTalk corpus than in the more task-based AMI corpus. Shared laughter was more prevalent in both corpora than solo laughter. Laughter, and particularly shared laughter, was also seen to be likely immediately before topic change, with a more marked effect in social talk (Gilmartin et al. 2013). These results on multiparty interaction were in line with observations from Conversation Analysis on the distribution of laughter around topic change in two-party conversations (Holt 2010), and strengthened the hypothesis that laughter plays a role in topic transition in social dialogue.

1.4 Obtaining Interaction Data

The information bundle in conversation includes audio (vocal) and visual signals and cues. Vocal information comprises verbal and non-verbal phenomena, including speech, pauses, prosodic contours, voice quality, laughter, coughing, and aspiration noise. The visual channel contains gesture displays, posture variations, and other cues ranging from facial expression, hand and body movements, and eye-gaze to variations in skin colouring (blushing). Some phenomena are tightly linked to the linguistic content of utterances – examples include intonation, beat and iconic gestures, others contribute to the participants' estimation of attitudinal and affective information about each other. In recent years there has been increasing interest in insights into affective and cognitive states gleaned from biosignals – data collected from the body including heart rate, electrodermal activity (skin conductance), skin temperature, pupil dilation and blood pressure. With technological advances, such measurements can be made unobtrusively over prolonged periods of time using wearable sensors and monitors.

The various facets of conversational exchange are often 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, biosignal data, and data collected from technology used during the interaction such as mice, keyboards or shared whiteboards. Earlier linguistic fieldwork relied on careful written transcription and, later, audio recordings of speech, where a phrase or structure was often elicited directly from an informant in speech community of interest. These elicited pieces of language informed surveys of phonological, syntactic, and semantic aspects of a language or dialect. To gather data on spoken interaction, speech has been elicited by having subjects perform a task or take part in interactions with a clear short-term goal; these 'task-based' interactions are often information gap activities completed by participants exchanging verbal information. Recordings of these tasks then form corpora on which study of dialogue is based. Typical tasks include describing a route through a map as in the HCRC MapTask corpus (Anderson et al. 1991), spotting differences in two pictures as in the DiaPix task in the LUCID (Baker and Hazan 2010) and Wildcat (Van Engen et al. 2010) corpora, ranking items on a list – for example to decide which items would be useful in an emergency (Vinciarelli et al. 2012), and participating in real or staged business meetings as in the ICSI and AMI corpora (Janin et al. 2003; McCowan et al. 2005). While these data collection paradigms do result in corpora of non-scripted dialogue of great utility to researchers, the participants' motivation for taking part in an artificial task is not clear, and the activity is removed from a natural everyday context. It is not certain that tasks such as these can be used to make generalizations about natural conversation (Lemke 2012).

Conversational data has been collected by researchers in the ethnomethological and conversational analysis traditions by recording everyday home or work situations. Early studies used recordings of telephone calls, as in the suicide hotline data collected by Sacks, Schelgoff's emergency services data, or the conversational data collected by Jefferson (c.f. for example Sacks, Schegloff, & Jefferson, 1974). More domain independent natural telephonic speech data has also been gathered by recording large numbers of real phone conversations, as in the Switchboard corpus (Godfrey, Holliman, and McDaniel 1992), and the ESP-C collection of Japanese telephone conversations (Campbell 2007). Audio corpora of non-telephonic spoken interaction include the Santa Barbara Corpus (DuBois et al. 2000), sections of the ICE corpora (Greenbaum 1991) and of the British National Corpus (BNC-Consortium 2000). Researchers have focused on everyday 'chat' by having subjects wear a microphone throughout the course of their daily lives for extended periods (Campbell 2004), while others have amassed collections of recordings of different types of human activity as in the Gothenburg Corpus (Allwood et al. 2000) which also contains video recordings. With increasing access to video recording and motion capture equipment, and awareness of the multimodal nature of human spoken interaction, rich multimodal corpora are appearing comprising naturalistic encounters with no prescribed task or subject of discussion imposed on participants. These include collections of free-talk meetings, or 'first encounters' between strangers as in the Swedish Spontal, and the NOMCO and MOMCO Danish and Maltese corpora (Edlund et al. 2010; Paggio et al. 2010). In our work, we have studied laughter in task-based and social talk using multimodal multiparty recordings - (the AMI meetings corpus) and social talk (the TableTalk corpus. For our current explorations of laughter in conversation we use the D-ANS corpus of multiparty conversation. Earlier work on laughter in multimodal corpora had brought to light a number of issues in existing laughter annotations. These issues were explored by creating an experimental annotation scheme for D-ANS. Below we outline the D-ANS corpus in general and the session of interest to our current work, and then describe the experimental laughter annotation procedure we used on D-ANS.

2 Data collection and annotation – the D-ANS corpus

The Dublin Autonomous Nervous System (D-ANS) corpus comprises three sessions of informal English conversation recorded over three days in a living-room like setting with sofa and armchair.

Participant	Sex	Age	Origin
F1	F	30s	USA
F2	F	30s	France
M1	M	30s	Ireland
M2	M	50s	Ireland
M3	M	60s	UK

Table 1 Participants in the D-ANS corpus

There were five participants in total – three men and two women, as shown in Table 1. Two of the men were Irish native English speakers, while the third was a British native English speaker. One woman was a North American native English speaker while the other woman was a French national who had lived in Ireland for some years. She worked in English and her spoken and written command of the language was at the C1 or near-native level on the Common European Scale (Little 2006). Participants were free to speak about any topic they liked and to move around as they wished.

Session 1 consisted of informal chat between the American woman (F1) and the British man (M3). Session 2 contains three hours of conversation between the American woman (F1), the British

Figure 2 Camera Angles in session 3 of D-ANS

man (M3), and one of the Irish men (M2). Session 3, the focus of the analysis described below, contains informal chat between the American woman (F1), one of the Irish men (M1), and the French woman (F2). The camera setup for this session is shown in Figure 2. There was a camera focused on the sofa giving a clear view of speaker F1 sitting on left of sofa and speaker M1 sitting on the right of the sofa. Another camera gave a clear view of speaker F2 sitting on the armchair. Sound was recorded using microphones close to each speaker with an additional omnidirectional microphone on the coffee table between them. The corpus includes measurements of participants' electro-dermal activity (EDA), also known as skin conductance or galvanic skin response. This is a measure of perspiration in the skin, which is linked to arousal. All participants wore biological sensors (Q-sensors) on the underside of each wrist to measure EDA. Q-sensors (Poh, Swenson, and Picard 2010) are wristbands which unobtrusively measure electro-dermal activity (galvanic skin response) in the wrist as well as three degrees of acceleration. While the wrist, in comparison to the palm or fingers, is a less sensitive location for recording EDA, these sensors were selected to allow free hand movement for gesturing during conversation.

2.2 Annotation of laughter in D-ANS

Several corpora we use have been annotated previously for laughter. The use of existing annotations is attractive to researchers, but in the early stages of our laughter work it became apparent that existing laughter annotation in corpora of interest was not adequate for detailed study, as we encountered several problems including mixtures of point and interval annotation, 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. Our observations are in line with problems with laughter annotation outlined by other researchers (Truong and Trouvain 2012). Many of these problems stemmed from the fact that the original

annotation was not created with laughter studies in mind, and so laughter was often annotated as an 'extra' or non-verbal phenomenon, often by including a symbol for laughter in the transcription – thus indicating the presence of laughter but not giving information on when exactly it occurred. There is also the lack of agreement in the literature as to what should be considered as laughter. Therefore in our preliminary studies our annotation scheme included only one label for laughter, and annotators were instructed to mark any incidence of laughter that they noticed.

We addressed these problems in our earlier work on TableTalk by performing a new manual laughter annotation using ELAN (Wittenburg et al. 2006) with separate laugh tracks for each speaker, annotating laughter according to the MUMIN scheme (Allwood et al. 2004). During annotation we noticed that many laughs were not acoustically sounded, (a requirement in the MUMIN definition), or too quiet to be picked up by microphone, and therefore we performed extra passes over the data using Praat with the sound files and Elan with video files to capture as much of the laughter as possible. This was a time-consuming process and raised the question of how much laughter was solely identifiable by video or audio alone. To investigate this question we devised an experimental annotation scheme by adding sounded (acoustic) and silent (non-acoustic) laugh tracks. We used this scheme with D-ANS to investigate silent and sounded perception of laughter. Annotation of laughter in D-ANS was performed in three passes – video only, audio only, and video with sound.

2.3 Video only 'silent' annotation

The video annotation was performed by two annotators, one 'naïve' annotator, a member of the public who was not involved in the linguistics or communications field. The second annotator was a speech researcher. For the video only ('silent') passes the two annotators were provided with silent video of the data and asked to mark intervals where they saw laughter. The annotation was performed in ELAN. For each video, a linked annotation file was created containing a single tier in which to annotate laughter for one of the speakers in the video. The video was marked for laughter while annotators watched in real time. 'Hot keys' were set up in the annotation file so that annotators could press a key when they saw laughter and press again when the laughter ended. Participants were allowed to pause the video, but were discouraged from replaying sections and changing their markings except in cases where they recognized that they had forgotten to mark laughter endpoints. In real time annotation of this type, there is an issue of lag or reaction time in the button pressing by participants. While Elan does offer a facility to automatically factor in a lag time correction to annotations, this was not used as it would correct for a constant lag. In real time annotation, it seems more likely that the initial delays in reaction would be greater than those later in the process as the annotator became accustomed to the task and to the speaker they were annotating. To address this contingency, before annotating a particular speaker, annotators were given 'practice' ELAN files containing video of the speaker taken from a different section of the corpus. It was hoped that this would allow any lag to settle before the annotator started on the video of interest. In any case, the real time laughter annotation by naïve annotators was not regarded as a highly temporally exact segmentation of laughter but rather as an indication of the occurrence of laughter, and was used as such in the analysis.

2.4 Audio only 'sounded' annotation and standard annotation

The audio only annotation was performed in Praat (Boersma and Weenink 2010) by two annotators marking sound recordings of the data with intervals where they heard laughter. The third annotation was a standard annotation made using the video and audio tracks together in ELAN. We performed two analyses on the D-ANS laughter data – investigating laughter around topic change and exploring differences in silent video only and sounded audio only annotations of laughter.

P	C	A1	A2	Agreed	Agreed %
M 1	**A**	51	52	48	**87**
F1	**A**	73	72	70	**93**
F2	**A**	52	50	49	**92**
M 1	**V**	98	90	86	**86**
F1	**V**	69	70	64	**85**
F2	**V**	105	94	92	**86**

Table 2 Agreement between Annotators (A1 and A2) on Audio only (A) and Video only (V) annotations.

3 Results

In the data we noted that many laugh annotations were separated by a short silence, often while the participant was breathing, so for the purpose of our analyses we define a 'laugh event' in which we amalgamated annotations of laughter which were separated by less than 1 second.

3.1 Multimodal annotation of laughter

We analysed the relationship between the silent and sounded laughter annotations in categorical terms by looking at raters' agreement on the incidence rather than the duration of laughter. We counted the number of laugh events noted in each of the Audio only (sounded) and Video only (silent) annotations (Table 2). We found inter-rater agreement to be high for each condition between annotators annotating the same modality. In the silent video condition agreement between the annotators ranged from 85 to 86% depending on speaker, while for the audio only annotations agreement ranged from 87 to 93% agreement per speaker. We then discarded all cases of audio or video laugh events which were marked by only one annotator, leaving a dataset with only the 'agreed' laughs - laughter recorded by both annotators in a particular modality (Table 3).

P	A and V	A only	V only
M1	37	3	44
F1	56	7	5
F2	45	2	47

Table 3 'Agreed' laugh events, where all annotators in the modality recorded laughs

Table 3 shows the resulting per-speaker counts for agreed laughs appearing in the annotations for the three speakers F1, F2, and M2 as described above. The 'A and V' column shows laughs which were recorded by both the Audio only annotators and the Video only (silent video) annotators. The 'A only' column shows laughs only picked up by Audio only annotators, while the 'V only' column shows laughs picked up by silent video annotators but not by audio annotators.

We found that most cases where annotations were made on video but not on the audio (V only) involve a combination of head tilting (pitch) and a growing broad smile or wide or toothy grin. In annotations on the audio but not the video (A only), most involve laughter co-occurring with speech with a much smaller number of cases where the annotation was of a short phrase initial or final laugh or snort.

3.2 Laughter and topic change

We extended previous analyses of shared and solo laughter in relation to topic change to the D-ANS data in order to investigate whether earlier results on the likelihood of laughter in the vicinity of topic change (Bonin, Campbell, and Vogel 2012; Gilmartin et al. 2013) would generalize to the D-ANS corpus.

In Session 3 of D-ANS, there were a total of 80 shared laughter events and 21 topics discussed. The distance from each of the 20 topic change points to the last shared laugh, as shown in Figure 3, ranged from 0 to 10.2 seconds, with 90%

Figure 3 Histogram of distances from topic changes to the endpoint of the last shared laugh

of topic changes occurring within 5 seconds of the end of an interval of shared laughter. Modelling the situation as a binomial distribution with the probability of any point falling within 5 seconds of the end of a shared laugh equal to the ratio of SL to T, where SL is the number of seconds on the recording meeting the criterion of being within 5 seconds of the end point of a shared laugh event and T is the total number of seconds in the recording, we can reject the null hypothesis that topic change points are randomly placed with respect to shared laughter ($p < 0.001$).

4 Discussion

The results of the topic transition analysis on a section of the DANS corpus show more shared than solo laughter in multiparty social dialogue; in line with earlier results and reports in the literature on the social nature of laughter. The strong likelihood of laughter before topic change points found in our analysis of D-ANS echoes the results of earlier work on TableTalk and AMI. A possible explanation for this tendency for topic changes to occur in or soon after shared laughter stems from the fact that in social talk there is set structure to follow and no agenda of topics lined up as there is in a meeting, nor are roles predetermined (Cheepen 1988; Laver 1975; Thornbury and Slade 2006). Without set roles or tasks or indeed a chairperson to 'move things along', the management of topic falls back on the participants. In extended social talk, conversation evolves as a series of often narrative longer 'chunks' interspersed with short 'chat' exchanges. Thus participants share the burden of providing entertaining speech input to the group. As topics are exhausted, there is a danger of uncomfortable silence, and participants may avoid this by entering a buffer of shared laughter, reinforcing positive bonds, and providing space for a new topic to be produced and the baton of speaker passed to another participant. Laughter may thus function around topic change much as the 'idling' behavior noted in social talk, when there is nothing much to be said but the impetus seems to be to keep the conversational ball in the air (Schneider 1988).

The pilot study results on multimodality indicate that careful annotation on the audio channel picks up most stereotypical sounded laughter, but can result in false positives in the case of speech laughs, although this phenomenon was observed in only one of the three speakers examined. Naïve human annotators asked to mark laughter watching silent video picked up the vast bulk of audio laughter, but also identified smiles and head nods accompanied by a wide grin as laughter. While this could be viewed as misclassification, it happens on a large enough scale to beg the question of whether such behavior should be regarded as silent laughter and thus may point to the need for a clearer taxonomy of laughter. In terms of applications of our findings on multimodal aspects of laughter, video based automatic identification of laughter may be an attractive prospect. Automatic identification of laughter on the audio stream is possible for stereotypical laughter (Scherer et al. 2009) but requires clean audio signals from close-coupled microphones. This is a limitation for real-world use of audio-based technology for laughter detection. Video signals are more robust, and identification on video data is an attractive idea, however there is a need for clear definitions of the various phenomena identified as laughter outside of the narrow stereotypical description. 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.

5 Future work

We are currently investigating the interplay of laughter and biosignals in D-ANS. We are particularly interested in any variation in 'task' and. 'chat' dialogue, in terms of laughter and of measured electro-dermal activity (EDA). EDA has been linked to levels of emotional arousal (Dawson, Schell, and Filion 2007) and to cognitive load (Shi et al. 2007), with work in psychology observing lower cognitive load in social chat than in talk arising during tasks (Kahneman 2011). Laughter has been observed to be more frequent in social than in task-based dialogue, and to be active around topic changes. This knowledge may help distinguish whether stretches of conversation are transactional or social in nature. It may thus be possible to contribute towards technological extraction of important or content-rich sections of dialogue using insights gained in our work. To further investigate the multimodality of laughter, we are creating more detailed laughter annotations which will allow us to further explore whether the silent phenomena our naïve annotators marked as laughter are functionally different to sounded laughter in terms of where and in what capacity they occur in conversation – in the listener/speaker, as backchannels, before or after audio laughter, in solo or shared laughter. We are currently extending our investigations to the D64 corpus of conversational English to test the generalizability of our findings.

Acknowledgments

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

Allen, James F., Donna K. Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu, and Amanda Stent. 2001. 'Toward Conversational Human-Computer Interaction'. *AI Magazine* 22 (4): 27.

Allwood, J., L. Cerrato, L. Dybkær, and P. Paggio. 2004. 'The MUMIN Multimodal Coding Scheme'. In *Proc. Workshop on Multimodal Corpora and Annotation*.

Allwood, Jens, Maria Björnberg, Leif Grönqvist, Elisabeth Ahlsén, and Cajsa Ottesjö. 2000. 'The Spoken Language Corpus at the Department of Linguistics, Göteborg University'. In *FQS–Forum Qualitative Social Research*. Vol. 1.

Anderson, A.H., M. Bader, E.G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, et al. 1991. 'The HCRC Map Task Corpus'. *Language and Speech* 34 (4): 351–66.

Bachorowski, J. A, and M. J Owren. 1995. 'Vocal Expression of Emotion: Acoustic Properties of Speech Are Associated with Emotional Intensity and Context'. *Psychological Science*, 219–24.

Baker, Rachel, and Valerie Hazan. 2010. 'LUCID: A Corpus of Spontaneous and Read Clear Speech in British English'. In *Proceedings of the DiSS-LPSS Joint Workshop 2010*.

Bakhtine, Mikhail Mikhaïlovitch. 1986. *Speech Genres and Other Late Essays*. 8. University of Texas Press.

Black, Donald W. 1984. 'Laughter'. *JAMA: The Journal of the American Medical Association* 252 (21): 2995–98.

BNC-Consortium. 2000. 'British National Corpus'. *URL Http://www. Hcu. Ox. Ac. uk/BNC*.

Boersma, Paul, and David Weenink. 2010. *Praat: Doing Phonetics by Computer [Computer Program], Version 5.1. 44*.

Bonin, Francesca, Nick Campbell, and Carl Vogel. 2012. 'Laughter and Topic Changes: Temporal Distribution and Information Flow'. In *Cognitive Infocommunications (CogInfoCom), 2012 IEEE 3rd International Conference on*, 53–58.

Burling, R. 2007. *The Talking Ape: How Language Evolved*. Vol. 5. Oxford University Press, USA.

Campbell, Nick. 2004. ‘Speech & Expression; the Value of a Longitudinal Corpus.’ In *Proc. LREC 2004*.

———. 2007. ‘Approaches to Conversational Speech Rhythm: Speech Activity in Two-Person Telephone Dialogues’. In *Proc XVIth International Congress of the Phonetic Sciences, Saarbrucken, Germany*, 343–48.

Cheepen, Christine. 1988. *The Predictability of Informal Conversation*. Pinter London.

Dawson, Michael E., Anne M. Schell, and Diane L. Filion. 2007. ‘The Electrodermal System’. *Handbook of Psychophysiology*, 159.

DuBois, John W., W. L. Chafe, C. Meyer, and S. A. Thompson. 2000. *Santa Barbara Corpus of Spoken American English. CD-ROM. Philadelphia: Linguistic Data Consortium*.

Dunbar, R. 1998. *Grooming, Gossip, and the Evolution of Language*. Harvard Univ Press.

Edlund, Jens, Jonas Beskow, Kjell Elenius, Kahl Hellmer, Sofia Strömbergsson, and David House. 2010. ‘Spontal: A Swedish Spontaneous Dialogue Corpus of Audio, Video and Motion Capture.’ In *LREC*.

Gilmartin, Emer, Francesca Bonin, Carl Vogel, and Nick Campbell. 2013. ‘Laugher and Topic Transition in Multiparty Conversation’. In *Proceedings of the SIGDIAL 2013 Conference*, 304–8. Metz, France: Association for Computational Linguistics.

Gilmartin, Emer, and Nick Campbell. 2014. ‘More Than Just Words: Building a Chatty Robot’. In *Natural Interaction with Robots, Knowbots and Smartphones*,

Glenn, Phillip J. 2003. *Laughter in Interaction*. Cambridge University Press Cambridge.

Godfrey, John J., Edward C. Holliman, and Jane McDaniel. 1992. ‘SWITCHBOARD: Telephone Speech Corpus for Research and Development’. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, 1:517–20.

Greenbaum, Sidney. 1991. ‘ICE: The International Corpus of English’. *English Today* 28 (7.4): 3–7.

Hennig, Shannon, Ryad Chellali, and Nick Campbell. 2014. ‘The D-ANS Corpus: The Dublin-Autonomous Nervous System Corpus of Biosignal and Multimodal Recordings of Conversational Speech.’ In *Proc. LREC 2014*, Reykjavik, Iceland.

Holt, Elizabeth. 2010. ‘The Last Laugh: Shared Laughter and Topic Termination’. *Journal of Pragmatics* 42 (6): 1513–25.

Janin, Adam, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, and Andreas Stolcke. 2003. ‘The ICSI Meeting Corpus’. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP’03). 2003 IEEE International Conference on*, 1:I–364.

Jokinen, Kristiina. 2009. ‘Gaze and Gesture Activity in Communication’. In *Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments*, 537–46. Springer.

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

Laver, John. 1975. ‘Communicative Functions of Phatic Communion’. *Organization of Behavior in Face-to-Face Interaction*, 215–38.

Lemke, Jay L. 2012. ‘Analyzing Verbal Data: Principles, Methods, and Problems’. In *Second International Handbook of Science Education*, 1471–84. Springer.

Little, David. 2006. ‘The Common European Framework of Reference for Languages: Content, Purpose, Origin, Reception and Impact’. *Language Teaching* 39 (03): 167–90.

Malinowski, B. 1923. ‘The Problem of Meaning in Primitive Languages’. *Supplementary in the Meaning of Meaning*, 1–84.

McCowan, Iain, Jean Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, and V. Karaiskos. 2005. ‘The AMI Meeting Corpus’. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*. Vol. 88.

Mehu, Marc, and Robin IM Dunbar. 2008. ‘Relationship between Smiling and Laughter in Humans (Homo Sapiens): Testing the Power Asymmetry Hypothesis’. *Folia Primatologica* 79 (5): 269–80.

Paggio, Patrizia, Jens Allwood, Elisabeth Ahlsén, and Kristiina Jokinen. 2010. ‘The

NOMCO Multimodal Nordic Resource–goals and Characteristics'.

Poh, M. Z., N. C. Swenson, and R. W. Picard. 2010. 'A Wearable Sensor for Unobtrusive, Long-Term Assessment of Electrodermal Activity'. *Biomedical Engineering, IEEE Transactions on* 57 (5): 1243–52.

Provine, Robert R. 1993. 'Laughter Punctuates Speech: Linguistic, Social and Gender Contexts of Laughter'. *Ethology* 95 (4): 291–98.

———. 2004. 'Laughing, Tickling, and the Evolution of Speech and Self'. *Current Directions in Psychological Science* 13 (6): 215–18.

Sacks, H., E.A. Schegloff, and G. Jefferson. 1974. 'A Simplest Systematics for the Organization of Turn-Taking for Conversation'. *Language*, 696–735.

Scherer, Stefan, Friedhelm Schwenker, Nick Campbell, and Günther Palm. 2009. 'Multimodal Laughter Detection in Natural Discourses'. In *Human Centered Robot Systems*, 111–20. Springer.

Schneider, Klaus P. 1988. *Small Talk: Analysing Phatic Discourse*. Vol. 1. Hitzeroth Marburg.

Shi, Yu, Natalie Ruiz, Ronnie Taib, Eric Choi, and Fang Chen. 2007. 'Galvanic Skin Response (GSR) as an Index of Cognitive Load'. In *CHI'07 Extended Abstracts on Human Factors in Computing Systems*, 2651–56.

Thornbury, Scott, and Diana Slade. 2006. *Conversation: From Description to Pedagogy*. Cambridge University Press.

Trouvain, Jürgen. 2003. 'Segmenting Phonetic Units in Laughter'. In *Proceedings of the 15th International Congress of Phonetic Sciences. Barcelona: Universitat Autònoma de Barcelona*, 2793–96.

Truong, Khiet P., and Jurgen Trouvain. 2012. 'Laughter Annotations in Conversational Speech Corpora-Possibilities and Limitations for Phonetic Analysis'. *Proceedings of the 4th International Worskhop on Corpora for Research on Emotion Sentiment and Social Signals*, 20–24.

Van Engen, Kristin J., Melissa Baese-Berk, Rachel E. Baker, Arim Choi, Midam Kim, and Ann R. Bradlow. 2010. 'The Wildcat Corpus of Native-and Foreign-Accented English: Communicative Efficiency across Conversational Dyads with Varying Language Alignment Profiles'. *Language and Speech* 53 (4): 510–40.

Vinciarelli, Alessandro, Hugues Salamin, Anna Polychroniou, Gelareh Mohammadi, and Antonio Origlia. 2012. 'From Nonverbal Cues to Perception: Personality and Social Attractiveness'. In *Cognitive Behavioural Systems*, 60–72. Springer.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. 'Elan: A Professional Framework for Multimodality Research'. In *Proceedings of LREC*. Vol. 2006.

Eye-trackers and Multimodal Communication Studies

Kristiina Jokinen
Institute of Behavioural Sciences
University of Helsinki
kristiina.jokinen@helsinki.fi

Abstract

This article provides an overview of eye-tracking technology in multimodal communication studies. It presents a short review of the human visual perception system and the eye-tracking technology, and discusses two types of eye-gaze studies as examples of how eye-trackers can be used in interaction management: in turn-taking analysis and involvement in conversation.

1 Introduction

The basic function of the eye is to provide visual information from the environment to the perceiving agent. Eye-gaze indicates where the speaker's focus of attention is directed, and it is thus one of the important multimodal feedback signals in human communication. For instance, if the gaze is rapidly wondering around, the person is understood as surveying the environment and collecting information from various points of interest, whereas looking straight at the partner normally signals interest and presence in the interaction with the partner. Gazing also has culturally determined interpretations related to appropriate social behaviour: looking into the partner's eyes can signal the speaker's reliability and truthfulness, although staring at the partner in general can be intimidating for the partner being scrutinized. Looking down can be a sign of humbleness, whereas the gaze wondering around can be interpreted as the person being absent-minded or demonstrating lack of interest in what the partner is presenting which would be considered socially unacceptable behaviour.

Due to the signalling of one's focus of attention, gaze is a powerful indicator of one's cognitive processing. It gives feedback to the partner of the mental efforts and emotions of the speaker, and also indicates the speaker's attitudes towards the partner in a given situation (Cassel et al. 2001). Eye-gaze is also used to control the interaction, as well as to build trust and rapport. Early work by Argyle and Cook (1976) described the role of eye-gaze in turn-taking and introduced the notion of "mutual gaze" for the point in interaction when the partners gaze at each other for a short time to agree on the change of the speaker. Much work on describing the functions and use of eye-gaze in human interactions has been conducted, and the reader is directed to the work e.g. by Cassell et al.1999; Goodwin 1981; Kendon 1990; Streeck and Knapp 1992; Gullberg 1999, among others.

Gazing forms the basis for joint visual attention, and is important when learning social cooperative behaviour. For a child, learning to follow the care-taker's gaze and to understand where their focus of attention is directed to, are important steps in the child's social development and language learning: they enable the child to distinguish "self" from "other" as well as to learn references to salient entities in the shared visual field (Trevarthen, 1984).

Gaze also has a strong cueing effect. Gullberg and Holmqvist (1999) show how interlocutors usually focus their attention on the speaker's face and not on their hands even though hand gestures may be large and peripheral. However, if the speaker fixates their gaze on the hand first, also the partner's gaze follows the hand. The gaze signals to the partner that something has attracted the speaker's attention in the hand movement, and the gaze following is thus automatic in order to maintain joint attention.

Gazing in the partner's face and gaze following seem to be conditioned to social face-

to-face situations. In an experiment Gullberg and Holmqvist (2006) demonstrate how interlocutors who communicate via a small-screen video-conference do not exhibit similar kind of gaze behaviour as those who communicate via a large-screen video-conference or in the presence of a live partner. In particular, they do not look at the partner's face nor follow gaze as often as the participants in the other conditions.

Eye-gaze is also accompanied by other type of eye activity, such as blinking of the eyes or change in the size of the pupil. They are mostly involuntary physiological reactions, but can tacitly indicate the person's cognitive occupation or emotions. For instance, the blinking of the eyes can signal the person's emotional state, and the size of the pupil is related to cognitive load. Facial expressions and eye-muscle movements are also related to gaze quality, and in ordinary language, one often talks about "twinkling" of the eyes, caused by the contraction of eye-muscles when laughing, or describes lack of emotion by the phrase "cold eyes". Along these lines, Poggi (2001) talks about the "alphabets of the eyes", where the shape and form of the eyes and eye-brows function as iconic communicative signals related to the speaker's mental attitudes: squeezed eyes can indicate that the speaker finds the partner's presentation unbelievable or the overall situation difficult or strenuous, while wide eyes usually signal surprise or fear.

Previous work has been mainly descriptive and based on manual analysis of videos. In the past years, eye-trackers have been introduced in the field of human communication studies, and it has become possible to collect more accurate and objective data on the interlocutors' gaze behaviour. Eye-trackers have already been used in medical and clinical research for a long time, but due to developments in the technology, they have improved in robustness and reliability, and also become cheaper, so their use in interaction studies has become feasible and more common.

This article aims to give a short overview of the use of eye-tracking technology in interaction studies. It is not meant to be an exhaustive and systematic overview of the research conducted in the area, but to provide a review of the state-of-the-art in eye-tracking research for understanding interaction and human communication.

The article is structured as follows. In Section 2, the human visual perception system is briefly presented, and in Section 3 an overview of the eye-tracking technology is given. In Section 4, some specific issues related to eye-tracking in the research of human communicative behaviour are discussed, and in Section 5 two research directions of the field are presented: turn-taking analysis and involvement in conversation. Section 6 concludes the article with a discussion of challenges of eye-tracking research.

2 Visual perception system

2.1 Basic concepts

Visual perception includes constant eye activity with *saccades* and *fixations*. The saccades are rapid eye movements when the eyes move simultaneously to the same direction, and the fixations are stops when the gaze is maintained on a single location. The eyes can also follow small moving objects in a manner called *smooth pursuit*, where the fovea is kept steady on the moving object. Another type of eye movement is *vergence movements* when the eyes move to opposite direction so as to adjust the fovea of both eyes to a near object. Seeing occurs only during fixations, and the area of accurate vision (foveal area) is of the size of about 0.3°-2° visual angle. On average there are three fixations per second, their length varying from less than 100ms to about two seconds.

As demonstrated already in the early vision studies (Buswell, 1935), the length of fixation can vary greatly. It is assumed that durations are determined by information processing and cognitive processes that concern interpretation of visual information and recognition of particular objects (Groner and Groner 1989). In other words, seeing is not the same as visual perception: the latter is affected by processing limitations, attention selection, memory capacity etc. and thus always includes interpretation of the visual information. In fact, despite the visual perception being based on discrete and mostly inaccurate seeing, human experience of the surrounding world is continuous and vision accurate. This is because the brain processes visual information by categorizing perceptions of the environment into objects and areas of interest, in a manner which seems to include elements of "problem solving" (Pylyshyn, 1999).

2.2 Eye-gaze and focus of attention

Our experience of the world is based on the things we attend to. Human attention is attracted by various multimodal aspects and features of

the environment, usually related to contrast and change. For instance, unusual shape, unexpected configurations, moving things and surprise appearances catch attention, but also familiarity and conventional forms can be important, especially when learning new skills. It should be emphasised that the focus of attention is not the same as visual attention, and although eye-gaze is commonly used to indicate what the agent is attending to, this can be different from what they are overtly looking at. For instance, looking at the person talking does not necessarily mean that the listener's focus of attention is on the speech: the listener may be attending to something else while directing their overt visual attention on the speaker. Visual attention can also be changed if something suddenly catches attention (ringing of a bell, being called by name, feeling cold, etc.).

One method to estimate the agent's focus of attention is to calculate *saliency maps* (see e.g. Walther and Koch, 2006). Saliency maps use low-level image features such as colour, intensity and orientation to identify high-contrast edges of possible objects of interest, and the bottom-up feature predictions produce output that concerns contrasting areas, e.g. face and background areas. However, eye-tracking experiments show that eye fixations do not coincide with the salience maps as such, but rather with the areas *inside* the salient edges, i.e. humans focus their eyes on the objects which can be perceived by the salient contours. For instance, human faces, eyes in particular, always attract attention, as well as text in images. Visual saliency maps thus need to be augmented by semantic knowledge of the vision scene, while saliency estimation must take into account those objects and activities that are meaningful to the agent.

Information for cognitive processes is selected through visual attention. Human experience is based on attending to salient objects and events in the environment and combining expectations with the actual perceptions via selective attention. Selective attention is a mechanism that is used to serialise the perception of objects in a complex scene. For instance, changes in the world are perceived by attending salient objects and features of the environment. However, perception presupposes interpretation of the visual data, and includes processing limitations, so what is seen is not exactly what is perceived. Simons and Chabris (1999) demonstrate this via selective attention tests. Striking experimental data exists about people's "change blindness", i.e. failure to notice apparent changes in images that are identical but one feature of minor importance (Rensink et al. 1997).

2.3 Visual attention

The manner in which human visual system works is complicated, and matches expectations of salient objects with the actual perception of the world. Current theories of visual attention hypothesise that human vision operates via two visual mechanisms, the global and the local one. They serve different objectives in continuous visual tasks and consequently employ different distribution of saccades and fixations (Unema et al 2005; Pannasch et al. 2011).

The global processing system is in the service of the ambient attention mode, and aims at getting a cursory view of the main regions of interest in the visual scene, whereas the local processing system contributes to the focal mode of attention, and focuses on examining details of the interesting objects. Global processing appears early in the viewing, and is associated with short fixations and long saccades (larger than 5°), so as to scan a larger area with accurate vision, whereas local processing occurs later in the viewing and is characterised by long fixations and short saccades (smaller than 5°), so it is possible to get more information from a particular object of interest. Fixations can thus be classified on the basis of the preceding saccade amplitude: larger amplitudes belong to the ambient attention mode, and shorter ones to the focal attention mode.

2.4 Coherence theory of visual attention

Visual attention on various objects depends on the task that the attention is to serve. The higher-level plans and goals provide targets on which to focus one's attention. The famous work by Yarbus (1967) showed that the eyes move differently in picture inspection depending on the initial goal given to the subject, although there is variation in the individual eye-movements. For instance, compared with free examination of the picture, the task to estimate the age of the people appearing in the picture resulted in a fixation pattern where the subjects focused their attention on the people's faces rather on scanning the whole scene. Task-related cognitive processes seem to control visual exploration of the environment in a top-down fashion.

The main problem in visual cognition is to account for the relation between higher-level

decisions that concern the attention of recognized objects and the visual perception itself: how the objects can be attended before they are recognized. The coherence theory (Rensink, 2000) proposes a solution which models bottom-up attention to salient objects, and uses the notion of *proto-object* to represent possible objects of attention in a salient region of the visual scene. The proto-objects are volatile structures based on bottom-up visual information processing, and they function as constantly re-generated units of visual information. It can be assumed that they function as expectations of the important events and objects in the environment, and they can be selected as the focus of attention depending on how they match with the cognitive requirements.

In computer vision, low-level categorisation of object features is used to produce salience maps within which the proto-objects can be accessed and be validated. These salient regions are used to restrict spatial locations which are likely to contain proto-objects, while the proto-objects can be validated as the actual objects of the scene by selective attention. The spatial location itself functions as an index that links the low-level features into proto-objects across space and time.

So far the visual attention studies have mostly dealt with static visual environments where the eye movement patters have been outlined with respect to a picture on a screen. Mobile eye-tracking technology has brought forward possibilities to study eye movements when the subject is in action, e.g. walking, typing, making tea, playing piano, etc. Two different principles have been identified in the eye-body movement correlation: fixations can focus exactly on the object the agent is engaged with, or they provide information of an action just before the particular action. The studies show that gaze is about one second ahead of the action start, see Land (2006) who gives an overview of the use of eye-gaze in action studies. In other type of tasks, for instance when driving a car, it has been observed that experts anticipate the route about 2-3 seconds ahead, while novices keep their eyes on the road just in front of the car (Sodhi et al., 2002).

In conclusion, we can say that human visual system is a complex mechanism which includes both bottom-up and top-down processes which function in integration. The system provides a means to attend the surrounding world, and maintain coherent experience of it.

3 Eye-tracking technology

As already mentioned, eye-trackers have long been used in medical diagnostic and clinical work. However, technology has developed much from the mechanical eye-trackers used by Huey (1898) to the present-day infra-red light reflection devices with advanced video image processing techniques. Eye-trackers have become more robust and practical, and available for interaction researchers in computer science, social and communication studies. In this article we focus on a short review of the technology only, and refer to Räihä and Majaranta (2007) for a more comprehensive overview of the development of eye-trackers and their use in human-computer interaction research.

Eye tracking refers to measuring where the agent is looking, i.e. their point of gaze. The eye tracker device measures gaze points and eye movement in real time, and reports gaze fixations as scan paths (gaze plots, Figure 1), or heatmaps.

Figure 1Eye fixations and a scan path.

The operation of modern eye-trackers is based on infra-red light reflection from the corneas of the user's eyes. The reflection patterns are collected by image sensors and image processing algorithms are used to identify relevant features, with the help of which gaze point coordinates on the screen can be calculated. Sampling rate is usually 50-120 Hz, which determines the relative accuracy between two consecutive gaze points.

In order to compensate for head movements, two reference points on the eye are needed and the difference in the reflection patterns account for head movements. Usually the pupil centre and the corneal reflection point are used as the reference points. The gaze points used to be measured with respect to head, which requires that the head has to be kept still with the help of a head rest. Modern computer vision-based eye-trackers can take head movements into account, although they still require that the subjects do not

turn their head sideways or move head up-down or back- and forward beyond certain limits. For most table-top trackers an optimum distance from the screen is 50-90 cm, while tolerance to side-turns is less than 20 degrees. Mobile head-mounted trackers or eye-tracking glasses allow subjects to move their head freely as well as walk around. The optics is similar to the table-top trackers except that it is in a miniature form. Measured accuracy is about 0.5 degree, and will always stay in order of 1° visual angle since the exact focus point can be determined only within the foveal area of about 2° visual angle.

Calibration of the tracker with respect to the user's gaze patterns is done before the tracking starts, and sometimes repeated also during long tests so as to compensate possible slight changes. This kind of calibration consists of recording the user's gaze when they are looking at the fixed points on the screen.

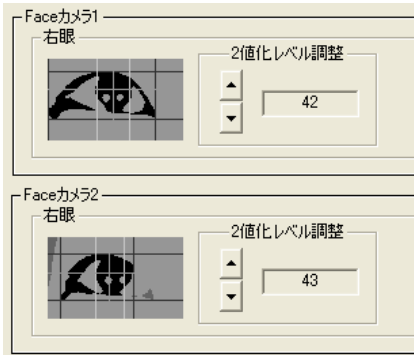

Figure 2 Calibrating the shape of the user's right eye.

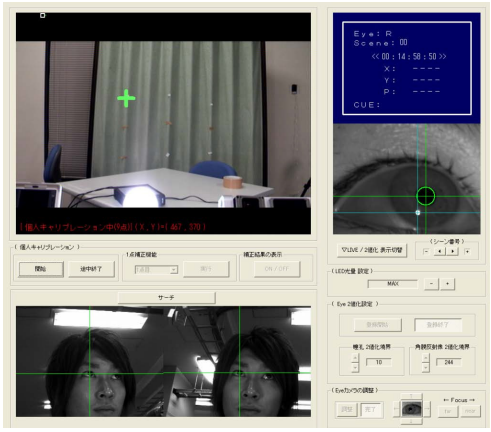

Figure 3 Control panel of an eye-tracker.

Visual information of the user is usually also included into calculations, e.g. facial features as well as eye shape and relative portion of white in the eye. Figure 2 shows the right eye of a user and calibration of the amount of light and dark areas in the eye shape.

Figure 3 shows a control panel of an eye-tracker, featuring the camera views of the visual scene (top left), the user's right and left eyes, and the reflection of the right eye.

4 Eye-tracking and interaction studies

In human communication and interaction research eye-tracking is a useful method as it adds objective information to descriptive observations. It supports analytical approach to estimate where the person is looking at, what they might have perceived, and what has drawn their attention. Moreover, it enables quantitative measures on gaze position, gazing time, and gaze plots, and thus support experimental studies and hypothesis testing on human cognitive processes and communication, e.g. studies on how the task affects gaze behaviour, or how gaze patterns indicate turn-taking. Eye-tracking experiments are also used for usability testing, cognitive load measurements and user evaluation of computer applications. Furthermore, gaze monitoring by an eye-tracker allows development of applications that make use of visual attention: human-computer interfaces for special user groups, computer-mediated communication, and controlling home appliances.

4.1 Metrics in eye-tracking studies

Common metrics used in eye-tracking studies deal with the number and length of fixations, gaze (cumulative duration of consecutive fixations on a particular spatial location), and scan paths (spatial arrangement of a sequence of fixations). Time to the first fixation on the target area of interest can also be useful. Fixations are defined as relatively stable eye positions with some threshold of spatial variation over a minimum duration (typically 100-200 ms). A set of several fixations on the area of interest together with short saccades between these fixations is referred to as "gaze". Jakob and Karn (2010) suggest that gazes are often more meaningful than counting the number of individual fixations. According to them, some authors have also used the term "dwell" in this meaning, although it has not yet become a common term.

Usual measurements include mean and overall number and duration of fixations, commonly measured with respect to particular areas of interest. The areas of interest are defined in advance by the researcher, and in human interaction studies they can include face, certain

(task-related) objects, background, etc., as well as temporal events like turn changes or gesturing.

The number of overall fixations is assumed to have negative correlation with search efficiency (more fixations tell about poor arrangement of the screen or visual scene) but the number should be normalised with respect to the task time, since more complicated tasks are longer and require more fixations. On the other hand, a large number of fixations on a particular object or an area of interest can also signal importance of the object; interpretation depends on the task. Overall fixation rate (the number of fixations in a time unit) is also used; it can signal about the person's emotional state, or about time pressure to learn about the important objects in the scene as quickly as possible.

Longer fixations are usually associated with problems on the particular object (unknown word in reading, complex object on a display), but can also indicate the importance of the object to the user. It has been pointed out that length and frequency may server different purposes: while duration of fixation may reflect difficulty of extraction information, frequency may reflect the importance of object or the area of interest.

4.2 Eye-tracker research

Visual attention studies with the help of eye-trackers started in the 1970's, focussing on experimental investigations on visual perception and cognitive processes, on tasks such as reading texts, searching information, or evaluating image quality. From early on, the typical use of eye-trackers has concerned medical diagnostics and clinical research, while in human-computer interaction and ergonomics, eye-trackers have been used, together with various other bio-physiological devices, to study human reaction, perception, and cognitive load, on complex practical tasks. By measuring the user's gaze patterns and how these differ depending on the user's experience as well as the task and overall layout of the environment, it is possible to get information about the human factors, i.e. about the user needs, skills, and processing constraints, which can help in the design and development of appropriate, efficient, and user-friendly applications. Research on human behaviour in complex tasks often use fairly sophisticated simulation environments, e.g. flight or car simulators, which enable observations in realistic but safe situations, about how various human factors affect the user's control and operation of an application at a particular point in time.

The "applied eye-tracking" deals with interface design and application development where eye-trackers are used to monitor the user's visual attention. This kind of information can then be used to infer the user's intention so as to adapt the application to the needs of the user's and thus serve the user better. For instance, eye-typing interfaces (Hyrskykari et al. 2005) allow the user to input text by focussing on particular letters on the screen, whereas gaze-aware systems aim to anticipate the user's mouse clicks by moving the mouse close to the point where the use's visual attention is located (MIDAS). The European network COGAIN maintains the research activities in this respect, while Räihä and Majaranta (2007) provide an overview of the issues related to gaze-based interfaces. Several workshops and conferences are also associated with the growing interest in eye-gaze studies. For instance ETRA and the series of GAZE-IN workshops (Gaze in Interaction) at ICMI provide annual meetings for studies on gaze and interaction.

5 Eye-gaze in human communication

In human communication and social studies, gaze has been extensively studied, although quantitative measures with eye-trackers are only recently being used in this context. In this paper we will not give an exhaustive and systematic overview of how eye-trackers have been used in human communication studies, but review two directions, where eye-trackers have been used to provide an objective basis for certain human behaviours: the coordination of turn taking and the effect of silent partners in multiparty communication situations.

5.1 Eye-gaze in turn-taking

Gaze is an effective means to coordinate turn-taking and to organize talk: by eye-gaze, the interlocutor can indicate which participant they are addressing their speech to or whether they have understood the speaker's utterance. Besides conversational feedback, eye-gaze is also used to coordinate and control turn-taking: looking at the conversational partner or looking away from the partner provides cues of the agent's willingness to continue interaction (Kendon 1967, Argyle

and Cook 1976, Nakano et al. 2007, Edlund et al. 2004, 2005, 2009).

Earlier studies on spoken interaction have identified several turn taking signals. Acoustic correlates related to distinctive intonation patters have been confirmed in many languages: low or low and falling intonation patterns are associated with turn-yielding and thus suitable turn-taking places, while mid- and high-level intonation patterns indicate turn-keeping and are inappropriate places for turn-taking (Koiso et al. 1998, Noguchi and Den 1998, Edlund et al. 2009). In the absence of boundary tones, also pauses play a role (Wennerström and Siegel 2003). Listeners are likely to wait longer before taking the turn, but the speaker is likely to continue speaking if the pause is longer than 0.3 seconds. After 0.5 seconds, the current speaker was likely to resume talking.

Gaze is a convenient way to convey meaning as it can occur simultaneously with speaking. Simultaneous gazing, or the mutual gaze by the speakers, is important when agreeing on the speaker change (Kendon 1967, Argyle and Cook 1976, Novick et al. 1996, Bavelas 2005, Jokinen et al. 2009, 2010). As described above, gazing at particular elements in the vision field can tell where the speaker's focus of attention is, and this is used in manage turn taking: the speaker who wants to yield the turn, signals this by directing their attention to a potential next speaker, while the partner who is willing to take the turn, focuses their attention onto the current speaker. If this happens simultaneously, the partners can thus synchronise their intentions, and turn taking is possible. Once the partners have visually shared and agreed on the speaker change, the next speaker will start their turn, and also break the mutual gaze by looking away. In fact, in casual conversations, the pressure on the next speaker to speak is so high that uttering nothing is considered extremely rude or it requires an explanation why the listener is not able to react as expected.

Coordination of turn taking in dialogues is often unproblematic since the two partners can fairly easily manage their intentions by gaze in fact-to-face situation. However, in multiparty conversations, gaze is not so reliable since the participants can focus their attention on other than the speaker, and also the speaker need not look at the partner who is willing to talk next. In these cases, head movement functions as an important signal: it is more visible than eye-gaze but still associated with visual attention and fairly reliable in group configurations where the participants need to turn their head to have a straight look at the partner.

In the series of studies by Jokinen et al. (2009, 2010a, 2010b, 2010c), the research centered on questions how eye-gaze affects turn-taking coordination in multiparty conversations, and if eye-gaze can help in predicting turn-taking possibilities. The work is based on the Doshisha Conversational Eye-gaze Data (Jokinen 2010b) which consists of 28 three-party conversations on free topics of interests, among participants who either know each other or are unfamiliar with each other. The corpus was collected using the NAC EMR-AT VOXER eye-tracker, and each conversation is about 10 minutes long, totalling almost 5 hours of data. Figure 1 shows a screen shot of the data.

The study found out that eye-gaze improves the prediction of turn-taking possibilities in spoken conversations, and used together with speech features, it is effectively used to distinguish between two different types of long pauses: those associated with turn-holds and those with turn-change. Long pauses and focussing of gaze on the partner indicate that the partner wants to give the turn to the partner, while gaze aversion during long pauses indicates turn-holds: because of hesitation or need to plan their utterance, the speaker does not focus on the partner and there is no possibility to yield the turn due to the lack of mutual gaze.

When studying the mean and standard deviation of gaze offset related to speech, Levitski et al., (2012) noticed that more gaze activity takes place in the beginning of the utterance than in the middle or at the end of the utterance, the times measured as one second before and after the start or the end of the utterance. The eyes are fixated significantly more often and longer in the beginning than at end of one's utterance, which corroborates with the notion of mutual gaze: the next speaker needs to make sure that the previous speaker indeed agrees to yield the turn and only then can break the mutual gaze, while the previous speaker only needs to scan if the partner is willing to accept the speaker change.

The analysis method advocated in the research is called Multi-level Hybrid Method, and it contains both top-down and bottom-up

techniques. The top-down approach refers to human observation and analysis of conversational phenomena, and uses annotation of the dialogues at the dialogue meaning level. The top-down approach reflects the observer's theoretical view-point and understanding of the phenomena in questions, and is to be validated by the inter-coder agreement calculation so as to reach more objective view or "gold standards".

The bottom-up approach refers to the analysis of the data at the signal level, and uses statistical and machine learning techniques to produce meaning correlations among the data. This can include common data mining techniques related to segmentation and clustering of the data, and can be used to produce meaningful relations. For instance, eye-trackers can be used to provide quantitative data about eye-gaze in various interactive situations which can be automatically analysed.

5.2 Conversational activity

Since turn taking coordination is a matter of the participants' engagement in the conversation, it is interesting to study how the interlocutors' engagement, as measured via their non-verbal activity, influences the other participants' gaze behaviour, and especially how the participants' focus of attention is changed in conversational situations when some of the partners are more active than the others?

In the study of Levitski et al. (2012), conversational activity refers to the interlocutor's general activeness in the interaction. It is defined as an individual speaker's intentional state characterised by energy and liveliness that produces expressive behaviour by speech, gaze, and body. Engagement, on the other hand, refers to the participant's presence in the interaction, and it is measured through their gaze activity. The definition coincides with the notion of entrainment, but is related to gaze.

The experiment used video data where three people discuss about their favourite films, with the subject being eye-tracked and one of the two discussants being naturally active in speaking and making many questions, while the other being naturally more quiet and passive. The Tobii X120 free standing eye tracking device was used in the experiments. Figure 4 is a screen shot of the experimental situation showing the eye-tracked person's eyes fixated at the person on the right.

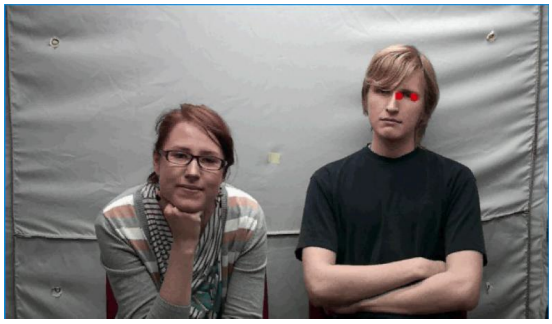

Figure 4 Experimental setup for the conversational activity studies: the active partner is on the left and the silent partner on the right.

In the study, the measurement is gaze activity rather than fixations. Gaze activity is defined as uninterrupted focussing on a particular target, so one token of gaze activity may contain many fixations. As is expected, more gazing is directed to the active partner than to the silent one, and also, the subjects had more gaze activity to both partners when speaking than when listening or backchannelling. When speaking, the subject directs gaze at the active partner, but when they are listening, gaze is divided between the two partners. This confirms the general observation that the participant's own gaze activity is related to speaking, i.e. to a more energetic (active) situation, and that speaking and showing active engagement also attracts the participants' focus of attention.

The experiment also suggests that the silent partner influences the subject's gaze behaviour, and indicates their awareness of the other partner. As expected, the subject's fixation targets and the silent partner's engagement (measured in number of overlapping segments) are correlated: there are more fixations on the silent partner if this is engaged, and if the silent partner is passive, there are more fixations on the other, active partner. However, if the silent partner is passive rather than engaged, the subject gazes at the partner's face less often, but twice as long.

Moreover, it also appears that there are more fixations on the active partner's face when the silent partner is engaged. This gives rise to the hypothesis that the silent partner's activity increases the subject's activity level, since the subject now needs to check the other partner's reaction, too: the increased engagement by the silent partner may cause a reaction in the other

partner as well. The subject is aware of both participants, so in order to keep up-to-date with the whole situation, the subject needs to quickly focus their attention to the other partner as well.

These observations and experimental results confirm the fact that turn-taking is a highly regulated event in the conversations, and that interactions involve social issues that need accurate gaze activity and rapid change of focus of attention so as to be able to manage smooth turn taking.

6 Visual Interaction Management

When looking at images of a person, people look at their faces, especially the area of eyes and mouth, if the faces are big. People are trying to see what social messages there are in the image, and gaze is mainly used for looking and retrieving information. However, in interactive situations, visual attention does not only function as a means to get information about the environment, but it is also a strong signal about communication. As discussed above, gaze can be used to direct the partner's attention to some important aspect of the environment (or distract them from something), and it is also effectively used to coordinate and control the interaction, i.e. there are social rules that regulate attention allocation (see also e.g. Skarratt et al. 2012).

Eye-gaze functions like other multimodal means, such as head nods, hand gestures, and body movement, in enabling the construction of shared understanding among the interlocutors. These means allow unobtrusive signalling of the speaker's conversational status simultaneously with their speaking, and are important in providing feedback about the basic enablements of the communication: whether the partner is willing to be in contact, if they are able to perceive and understand the partner's message, and consequently willing and capable to produce relevant continuation to the interaction. They can all signal the participants' engagement in the interaction.

It is also necessary that the interlocutors are familiar with the non-verbal means and have a similar set of interpretations so that they can be interpreted in the intended way. There are differences in the interpretation of a particular behaviour in different (cultural) contexts, and thus the interlocutors must learn the necessary and important gaze signals in order for the communication to be smooth and efficient. Through gaze studies one can also increase this kind of the awareness in the communication: although gaze patterns often are unconscious and unintentional, the speakers can learn to control them intentionally.

Acknowledgments

Thanks to Seiichi Yamamoto and his staff and students at Doshisha University for collaborating on the research on turn-taking, and to Jenni Radun and students at the University of Helsinki for research on interaction engagement.

References

Argyle, M. and Cook, M. 1976. *Gaze and mutual gaze.* Oxford, England: Cambridge U Press.

Bavelas, J. B. 2005. The two solitudes: Reconciling Social Psychology and Language and Social Interaction. In K. Fitch & R. Sanders (Eds.), *Handbook of Language and Social Interaction* (pp. 179-200). Mahwah, NJ: Erlbaum.

Buswell, G. T. 1935. *How people look at pictures*. University of Chicago Press, Chicago.

Cassell, J., Nakano, Y., Bickmore, T., Sidner, C. and Rich, C. 2001. "Non-Verbal Cues for Discourse Structure." *Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics,* pp. 106-115. July 17-19, Toulouse, France.

Cassell, J., Vilhjálmsson, H. and Bickmore, T. 2001 "BEAT: the Behavior Expression Animation Toolkit." *Proceedings of SIGGRAPH '01*, pp. 477-486. August 12-17, Los Angeles, CA.

Cassell, J. and Ryokai, K. 2001. "Making Space for Voice: Technologies to Support Children's Fantasy and Storytelling." *Personal Technologies* 5(3): 203-224.

Cassell, J., Bickmore, T., Vilhjalmsson, H. and Yan, H. 2001. "More Than Just a Pretty Face: Conversational Protocols and the Affordances of Embodiment." *Knowledge-Based Systems* 14: 55-64.

Cassell, J. and Bickmore, T. 2001 "A Relational Agent: A Model and Implementation of Building User Trust." *Proceedings of the CHI'01 Conference*, pp. 396-403. March 31-April 5, Seattle, Washington.

Cassell, J., D. McNeill, and K. E. McCullough. 1999. Speech-gesture mismatches: evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics and Cognition* 7(1):1–34.

Cogain Network for Gaze and interaction studies: http://www.cogain.org/wiki/Bibliography_Gaze_Interaction

Duchowski, A.T. 2003. *Eye-tracking Methodology: Theory and Practice*. Springer

Edlund, J., Skantze, G. and Carlson, R. 2004. Higgins - a spoken dialogue system for investigating error handling techniques- In Proceedings of ICSLP , 2004

Edlund, J., House, D. and Skantze, G. 2005. The effects of prosodic features on the interpretation of clarification ellipses- In *Proceedings of Interspeech* 2005.

Edlund, J., Heldner, M. and Hirschberg, J. 2009a. Pause and gap length in face-to-face interaction. In *Proceedings. of Interspeech 2009*, Brighton.

Edlund, J., Heldner, M. and Pelcé, A. 2009b. Prosodic features of very short utterances in dialogue. In *Proceedings of the Nordic Prosody* 2008, pp. 57-68. Frankfurt am Main.

Goodwin, C. 1981. *Conversational Organization: Interaction Between Speakers and Hearers.* New York, NY: Academic Press.

Groner, R. and Groner, M. T. 1989. Attention and eye movement control: An overview. *European Archives of Psychiatry and Clinical Neuroscience*, 239, 9–16.

Gullberg, M. and Holmqvist, K. 1999. Keeping an Eye on Gestures: Visual Perception of Gestures in Face-to-Face Communication. *Pragmatics and Cognition* 7 (1):35-63.

Gullberg, M. and Holmqvist, K. 2006. What speakers do and what addressees look at: Visual attention to gestures in human interaction live and on video. *Pragmatics & Cognition,* 14(1), 53-82

Huey, E. B. 1898. Preliminary experiments in the physiology and psychology of reading. *American Journal of Psychology, 9*, 575-586.

Hyrskykari, A., Majaranta, P. and Räihä, K.-J. 2005. From gaze control to attentive interfaces. *Proceedings of HCII 2005*, Las Vegas, NV.

Jakob, R.J.K. and Karn, K.S. 2010. Commentary on Section 4. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises.

Jokinen, K., Nishida, M. and Yamamoto, S. 2009. Eye-gaze Experiments for Conversation Monitoring. *The 3rd International Universal Communication Symposium*, Tokyo, Japan.

Jokinen, K. and M. McTear 2009. *Spoken Dialogue Systems*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool.

Jokinen, K. and F. Cheng 2010. *New Trends in Speech-based Interactive Systems*. Springer Publishers.

Jokinen, K. and J. Allwood 2010. Hesitation in Intercultural Communication: Some Observations and Analyses on Interpreting Shoulder Shrugging. In: T. Ishida (Ed.): *Culture and Computing*, LNCS 6259, pp. 55--70. Springer, Heidelberg.

Jokinen, K., K. Harada, M. Nishida and S. Yamamoto 2010a. Turn-alignment using eye-gaze and speech in conversational interaction. *Proceedings of Interspeech 2010*. Makuhari, Japan.

Jokinen, K., M. Nishida and S. Yamamoto 2010b. Collecting and Annotating Conversational Eye-Gaze Data. *Proceedings of Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality (MMC 2010)*, Language Resources and Evaluation Conference (LREC-2010). Valetta, Malta

Jokinen, K., Nishida, M. and Yamamoto, S. 2010c. On Eye-gaze and Turn-taking. Proceedings of the Workshop on *Eye-gaze in Intelligent Human-Machine Interaction. International Conference on Intelligent User Interfaces*.

Kendon, A. 1967. Some functions of gaze direction in social interaction. *Acta Psychologica*,*26*, 22–63.

Kendon, A, 1990. Signs in the cloister and elsewhere. *Semiotica. 79, 307-29.*

Koiso, H., Horiuchi, Y., Tutiya, S. Ichikawa, A. and Den, Y. 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs. *Language and Speech,* 41(3-4):295–321.

Land, M. F. 2006. Eye movements and the control of actions in everyday life. *Progress in retinal and eye research*, *25*(3), 296–324.

Land, M. F. 2009. Vision, eye movements, and natural behavior. *Visual neuroscience*, *26*(1), 51–62.

Levitski, A., Radun, J. and Jokinen, K. 2012. Visual interaction and conversational activity. *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction: Eye Gaze and Multimodality.* Santa Monica, USA

Majaranta, P. and Räihä, K.-J. 2007. Text entry by gaze: Utilizing eye-tracking. In MacKenzie, I. S., and Tanaka-Ishii, K. (Eds.), *Text entry systems: Mobility, accessibility, universality,* pp. 175-187. Morgan Kaufmann

Noguchi, H. and Den, Y. 1998. Prosody-Based Detection of the Context of Backchannel Responses. In Fifth International Conference on Spoken Language Processing.

Nakano, Y. and Nishida, T. 2007. Attentional behaviours as nonverbal communicative signals in situated interactions with conversational agents. In Nishida, T. (Ed.), *Engineering approaches to conversational informatics,* pp. 85-102. John Wiley & Sons

Novick, D., Walton, L. and Ward, K. 1996. Contribution graphs in multiparty conversations, *Proceedings of the International Symposium on Spoken Dialogue* (ISSD-96), Philadelphia, PA, October, 1996, 53-56.

Pannasch, S., Schulz, J. and Velichkovsky, B. M. 2011. On the control of visual fixation durations in free viewing of complex images. *Attention, Perception, & Psychophysics*, Psychnomic Society, Inc. DOI 10.3758/s13414-011-0090-1

Poggi I. 2001. The lexicon and the Alphabet of Gesture, Gaze, and Touch. *Proceedings of the Third International Workshop on Intelligent Virtual Agents (IVA),* p. 235-236. http://dx.doi.org/10.1007/3-540-44812-8_20

Pylyshyn, Z. 1999. Is vision continuous with cognition? The case for cognitive impenetrability of visual perception *Behavioral and Brain Sciences* 22:341–423. Cambridge University Press.

Rensink, R.A., O'Regan, J., Kevin and Clark, J. 1997. To see or not to see: the need for attention to perceive changes in scenes. *Psychological Science* 8 (5): 368–373.

Rensink, R. A. 2000. The dynamic representation of scenes. *Visual Cognition*, 7(1/2/3), 17–42.

Simons and Chabris 1999. In youtube: http://www.youtube.com/watch?v=vJG698U2Mvo

Skarratt, P.A., Cole, G.G. and Kuhn, G. 2012. Visual cognition during real social interaction. *Frontiers in human neuroscience*, *6*, 196.

Sondhi, A., O'Shea, J. and Williams, T. 2002 *Arrest Referral: emerging findings from the national monitoring and evaluation programme.* DPAS paper 18. London: Home Office.

Streek, J. and Knapp, M. L. 1992. The interaction of visual and verbal features in human communication. In F. Poyatos (Ed.), *Advances in Nonverbal Communication; sociocultural, Clinical, Esthetic and Literary Perspectives. pp. 3-23.* Amasterdam and Philadelphia: John Benjamins.

Trewarthen, C. 1984. "Emotions in Infancy: Regulators of Contact and Relationships with Persons." Pp. sivunumero, ei löydy!! in *Approaches to Emotion,* edited by K. R. Sherer and P. Ekman. Hillsdale, NJ: Lawrence Erlbaum.

Unema, P. J. A., Pannasch, S., Joos, M. and Velichkovsky, B. M. 2005. Time course of information processing during scene perception: The relationship between saccade amplitude and fixation duration. *Visual Cognition*, 12, 473–494.

Walther, D. and Koch, C. 2006. Modeling attention to salient proto-objects. *Neural Networks* **19**, 1395-1407.

Wennerstrom, A. and Siegel, A. F. 2003. Keeping the Floor in Multiparty Conversations: Intonation, Syntax, and Pause. *Discourse Processes* 36, 77-107.

Yarbus, A. 1967. *Eye Movements and Vision*, Plenum Press, New York.

Yonezawa, T., Yamazoe, H., Utsumi, A. and Abe, S. 2007. Gaze-communicative behavior of stuffed-toy robot with joint attention and eye contact based on ambient gaze-tracking. *Proceedings of the 9th International Conference on Multimodal Interfaces* (ICMI'07), pp. 140-145. New York, NY: ACM.

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

Magdalena Lis
Centre for Language Technology
University of Copenhagen
Njalsgade 140
2300 Copenhagen
magdalena@hum.ku.dk

Costanza Navarretta
Centre for Language Technology
University of Copenhagen
Njalsgade 140
2300 Copenhagen
costanza@hum.ku.dk

Abstract

This paper deals with the relation between speech and form of co-occurring iconic hand gestures. It focuses on multimodal expression of eventualities. We investigate to what extent it is possible to automatically classify gestural features from the categorization of verbs in a wordnet. We do so by applying supervised machine learning to an annotated multimodal corpus. The annotations describe form features of gestures. They also contain information about the type of eventuality, verb Aktionsart and Aspect, which were extracted from plWordNet 2.0. Our results confirm the hypothesis that the Eventuality Type and Aktionsart are related to the form of gestures. They also indicate that it is possible to some extent to classify certain form characteristics of gesture from the linguistic categorization of their lexical affiliates. We also identify the gestural form features which are most strongly correlated to the Viewpoint adopted in gesture.

Keywords: multimodal eventuality expression, iconic co-speech gesture, wordnet, machine learning

1 Introduction

In face-to-face interaction humans communicate by means of speech as well as co-verbal gestures, i.e. spontaneous and meaningful hand movements semantically integrated with concurrent spoken utterances (Kendon, 2004; McNeill, 1992). Gestures which depict entities are called iconic gestures. Such gestures are co-expressive with speech, but not redundant. According to inter alia McNeill (1992; 2005) and Kendon (2004), they form an integral part of a spoken utterance.

Iconic gestures are especially well-suited to express spatio-motoric information (Alibali et al., 2001; Krauss et al., 2000; Rauscher et al., 1996) and, thus, often accompany verbal expressions of eventualities, in particular motion eventualities. Eventuality is an umbrella term for entities like events, actions, states, processes, etc. (Ramchard, 2005).[1] On the level of language, such entities are mostly denoted by verbs. Gesturally, they are depicted by means of iconicity relation (McNeill, 1992; McNeill, 2005; Peirce, 1931). This relation does not, however, on its own fully explain the form that a gesture takes - a referent can be depicted in gestures in multiple ways, for instance from different perspectives - that of the observer or that of the agent. How a speaker chooses to represent a referent gesturally determines which physical form a gesture takes. Knowledge about the factors influencing this choice, is still sparse (Kopp et al., 2008). It is, however, crucial not only for our understanding of human communication but also for theoretical models of gesture production and its interaction with speech. Such models can in turn inform generation of natural communicative behaviors in Embodied Conversational Agents (Kopp et al., 2008).

The present paper contributes to this understanding. It addresses a particular aspect of gesture production and its relationship to speech with focus on multimodal expression of eventualities. Various factors have been suggested to influence eventuality gestures, including referent characteristics (Parrill, 2010; Poggi, 2008), verb Aspect (Duncan, 2002) and Aktionsart (Becker et al., 2011). We present a pilot study investigating the extent to which hand gestures can be automatically

[1] In gesture studies the terms 'action' or 'event' are habitually used in this sense. We adopted the term 'eventuality' to accommodate the terminology for Aktionsart categories reported in Subsection 3.2.2, where 'action' and 'event' are subcategories of what can be termed 'eventualities.'

classified from the information about these factors. We extract this information from the categorization of verbs in a lexical-semantic database called wordnet. The theoretical background and methodological framework are discussed in (Lis, 2012a; Lis, 2012b; Lis, submitted).

In the present paper, differing from preceding studies on the multimodal expression of eventualities, we test the hypotheses by applying supervised learning on the data. Our aim in employing this method is to test the annotation scheme and potential application of the annotations in automatic systems and to study the relationship between speech and gesture not only for relations between single variables but also groups of attributes. In this, we follow the approach adopted by a number of researchers. For example, Jokinen and colleagues (2008) have used classification experiments to test the adequacy of the annotation categories for the studied phenomenon. Louwerse and colleagues (2006a; 2006b) have applied machine learning algorithms on annotated English map-task dialogues to study the relation between facial expressions, gaze and speech. A number of papers (Fujie et al., 2004; Morency et al., 2009; Morency et al., 2005; Morency et al., 2007; Navarretta and Paggio, 2010) describe classification experiments testing the correlation between speech, prosody and head movements in annotated multimodal corpora. Machine learning algorithms have also been applied to annotations of hand gestures and the co-occurring referring expressions in order to identify gestural features relevant for coreference resolution (Eisenstein and Davis, 2006; Navarretta, 2011).

Moreover, in the present work, we extend the annotations reported in (Lis, 2012b) with two more form attributes (Movement and Direction). These attributes are chosen because they belong to fundamental parameters of gesture form description (Bressem, 2013) and they are associated with motion, so are expected to be of importance considering we study eventualities, especially motion ones.

The paper is organized as follows. In section 2, we shortly present the background for our study, and in section 3 we describe the multimodal corpus and the annotations used in our analyses. In section 4, we present the machine learning experiments, and in section 5 we discuss our results and their implications, and we propose directions for future research.

2 Background

The form of co-verbal, iconic gestures is influenced by, among others, the semantics of the co-occurring speech and by the visually perceivable characteristics of the entity referred to (Kita and Özyürek, 2003; McNeill, 1992). Poggi (2008) has suggested that not only the observable properties of the referent should be taken into consideration but also "the type of semantic entity it constitutes." She has distinguished four such types (Animates, Artifacts, Natural Objects and Eventualities) and proposed that their gestural representation will differ.

Eventualities themselves can still be represented in gesture in various ways, for example from different Viewpoints (McNeill, 1992; McNeill, 2005). In Character Viewpoint gestures (C-vpt), an eventuality is shown from the perspective of the agent, gesturer mimes agent's behavior; in Observer Viewpoint (O-vpt), the narrator sees the eventuality as an observer and in Dual Viewpoint (D-vpt), the gesturer merges the two perspectives. Parrill (2010) has suggested that the choice of Viewpoint is influenced by the eventuality structure. She has proposed that eventualities which have trajectory as the more salient element - elicit O-vpt gesture, while eventualities in which the use of character's hands in accomplishing a task is more prominent - tend to evoke C-vpt gestures.

Other factors suggested to influence eventuality gestures include verb Aspect and Aktionsart. Aspect marks "different ways of viewing the internal temporal constituency of a situation" (Comrie, 1976). The most common distinction is between perfective and imperfective aspect: the former draws focus to the completeness and resultativness of an eventuality, whereas with the latter the eventuality is viewed as ongoing. Duncan (2002) has analyzed the relationship between Aspect of verbs and Handedness in gestures in English and Chinese data. Handedness regards which hand performs the movement and, in case of bi-handed gestures, whether the hands mirror each other. Duncan has found that symmetric bi-handed gestures more often accompany perfective verbs than imperfective ones; the latter mostly co-occur with two handed non-symmetric gestures. Parrill and colleagues (2013) have investigated the rela-

tionship between verbal Aspect and gesture Iteration (repetition of a movement pattern within a gesture). They have found that descriptions in progressive Aspect are more often accompanied by iterated gestures. This is, however, only the case if eventualities are presented to the speakers in that Aspect in the stimuli.

Aktionsart is a notion similar to, but discernible from, Aspect.[2] It concerns Vendler's (1967) distinction between States, Activities, Accomplishments and Achievements, according to differences between the static and dynamic, telic and atelic, durative and punctual. Becker and colleagues (2011) have conducted a qualitative study on Aktionsart and temporal coordination between speech and gesture. They have suggested that gestures affiliated with Achievement and Accomplishment verbs are completed, or repeated, on the goal of the verb, whereas in case of gestures accompanying Activity verbs, the stroke coincides with the verb itself.

Lis (2012a) has introduced a framework in which the relationship between these factors and gestural expressions of eventualities is investigated using wordnet databases, i.e. electronic linguistic taxonomies. She has employed wordnet to, among others, formalize Poggi (2008) and Parrill's (2010) insights. Based on plWordNet 1.5 classification, she has distinguished different types of eventualities and showed their correlation with gestural representation (Lis, 2012b). The present study further builds up on that work, using updated (plWN 2.0), revised (Lis, submitted) and extended annotations and machine learning experiments.

3 The data

3.1 The corpus

Our study was conducted on the refined annotations (Lis, submitted) from the corpus described in (Lis, 2012a; Lis, 2012b), which has in turn been an enriched version of the PCNC corpus created by the DiaGest research group (Karpiński et al., 2008). Data collection followed the well-established methodology of McNeill (1992; 2005): the corpus consists of audio-video recordings of 5 male and 5 female adult native Polish speakers who re-tell a Canary Row cartoon to an addressee. The stimulus contains numerous eventualities and has proved to elicit rich multimodal output. The monologues were recorded in a studio as shown in Figure 1 and the whole corpus consists of approximately one hour of recordings.

[2]For a discussion on the differences between Aspect and Aktionsart and between the Germanic and Slavic traditions of viewing these two concept cf. (Młynarczyk, 2004).

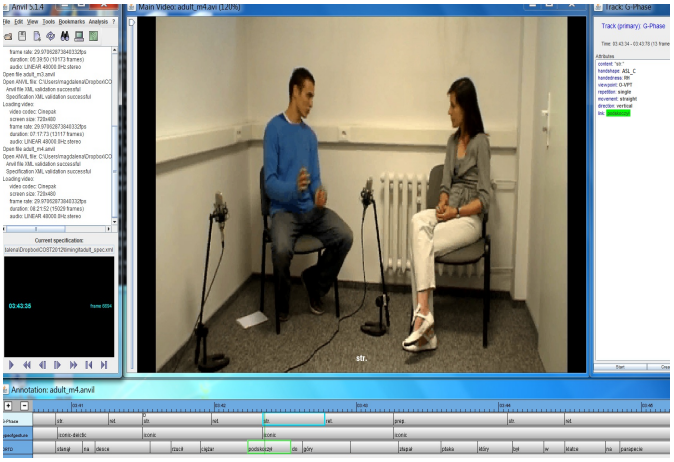

Figure 1: A snapshot from the ANVIL tool

3.2 The annotation

Speech has been transcribed with word time stamps by the DiaGest group, who has also identified communicative hand gestures and annotated their phases, phrases and semiotic types in ELAN (Wittenburg et al., 2006). Lis (2012a; 2012b) exported the annotations to the ANVIL tool (Kipp, 2004) and enriched it with coding of verbs and Viewpoint, Handedness, Handshape and Iteration of gestures. The annotations in the corpus were refined and, for the purpose of the present study, further extended with two more gesture form attributes (Direction and Movement) (Lis, submitted).

3.2.1 The annotation of gestures

Iconic hand gestures were identified based on DiaGest's annotation of semiotic types. Gestures depicting eventualities were manually annotated using six pre-defined features, as reported in detail in (Lis, submitted). Table 1 shows the attributes and values for gestures annotation used in this study. Viewpoint describes the perspective adopted by the speaker and was encoded using the values proposed by McNeill: C-, O- and D-vpt (1992). The attribute Handedness indicates whether one (*Right_Hand*, *Left_Hand*) or two hands are gesturing and whether they are symmetric or not (*Symmetric_Hands* versus *Non-symmetric_Hands*). Handshape refers to configu-

Table 1: Annotations of gestures

Attribute	Value
Viewpoint	Observer_Viewpoint, Character_Viewpoint, Dual_Viewpoint,
Handshape	ASL_C, ASL_G, ASL_5, ASL_O, ASL_S, Complex Other
Handedness	Right_Hand, Left_Hand, Symmetric_Hands, Non-symmetric_ Hands
Iteration	Single, Repeated, Hold
Movement	Straight, Arc, Circle, Complex, None
Direction	Vertical, Horizontal, Multidirectional, None

ration of palm and fingers of the gesturing hand(s); the values are taken from American Sign Language Handshape inventory (Tennant and Brown, 2010): *ASL_C, ASL_G, ASL_5, ASL_O, ASL_S* and supplemented with the value for hand shapes changing throughout the stroke (*Complex*) or not falling under any of the mentioned categories (*Handshape_Other*). Iteration indicates whether a particular movement pattern within a stroke occurs once (*Single*) or multiple times (*Repeated*), or whether the stroke consists of a static *Hold*. Movement regards shape of the motion, while Direction - the plane on which the motion is performed.

3.2.2 The annotation of verbs

Verbs were identified in the word stamp speech transcript. Information about verbs was extracted from the Polish WordNet, plWordNet 2.0, following the procedure explained in (Lis, 2012a; Lis, 2012b). In a wordnet, the lexical units are classified into sets of synonyms, called synsets, which are linked to each other via a number of conceptual-semantic and lexical relations (Fellbaum, 1998). The most frequently encoded one is hyponymy, also called IS_A or TYPE_OF relation, that connects a sub-class to its superclass, the hyperonym. Non-lexical synsets in the upper-level hierarchies of hyponymy encodings in plWordNet contain information on verb Aspect, Aktionsart and domain (Maziarz, 2012).

A domain denotes a segment of reality and all lexical units belonging to a particular domain share a common semantic property (Brinton, 2000). Lis (2012a; 2012b) has used wordnet domains to categorize referents of multimodal expressions according to their type. The attribute Eventuality Type was assigned based the domain of the verb used in speech to denote the eventuality. The choice of the domains in focus has been partially inspired by Parrill's distinction between eventualities with a more prominent trajectory versus eventualities with a more prominent handling element (Parrill, 2010). Based on this, Lis (2012a; 2012b) has distinguished two Eventuality Types:[3] Translocation and Body_Motion. The former refers to eventualities with traversal of a path of a moving object or focus on spatial arrangement and the latter refers to a movement of agent's body (part) not entailing displacement of the agent as a whole (cf: (Levin, 1993)). Lis has subsumed plWordNet domains to fit this distinction. The domains relevant to our study are (with examples of verbs from the corpus given in parentheses):

TRANSLOCATION

{location or spatial relations}[4](*spadać* 'to fall,' *zderzać się* 'to collide');

{change of location or spatial relations change}(*biegać* 'to run,' *skakać* 'to jump').

BODY_MOTION

{causing change of location or causing spatial relations change}(*rzucać* 'to throw,' *otwierać* 'to open');

{physical contact}(*bić* 'to beat', *łapać* 'to catch');

{possession}(*dawać* 'to give,' *brać* 'to take');

{producing}(*budować* 'to build,' *gotować* 'to cook').

Verbs from the synsets {location or spatial relations} and its alterational counterpart were subsumed under the type Translocation. More examples of the verbs from the corpus include: *wspinać się* 'to climb,'*chodzić* 'to walk,' *wypadać* 'to fall out'. Synsets {causing change of location or causing spatial relations change} and {physical contact}, as well as {possession} and {producing} were grouped under the type Body_Motion. Further verb examples are: *przynosić* 'to bring,' *trzymać* 'to keep,' *walić* 'to bang,' *dawać* 'to give,' *szyć* 'to sew.' Verbs from the remaining domains were collected under the umbrella term 'Eventuality_Other.' These verbs constituted less than 10% of all verb-gesture tokens found in the data. Examples include: {social relationships} *grać* 'to play,' {mental or emotional state} *oglądać* 'to watch.' For the purpose of the analyses in the present paper, they were combined with

[3]Note that these categories are orthogonal to Poggi's (2008) ontological types.

[4]In wordnets {} indicates a synset.

Table 2: Annotations of verbs

Attribute	Value
Eventuality Type	Translocation, Body_Motion, Other
Aspect	Perfective,Imperfective
Aktionsart	State, Act, Activity, Accident, Event, Action, Process

Table 3: Classification of Handshape

Hanshape	Precision	Recall	F-score
baseline	0.08	0.28	0.12
Aspect	0.08	0.28	0.12
Aktionsart	0.22	0.28	0.21
Type	**0.17**	**0.32**	**0.22**
all	0.19	0.27	0.21

the Body_Motion category.[5] The domains were semi-automatically assigned to the verbs in our data. Verb polysemy was resolved with a refined version (Lis, submitted) of the heuristics proposed in (Lis, 2012b).

Apart from the domains, the encoding of hyponymy-hyperonymy relations of verbs in plWordNet provides also information about Aktionsart and Aspect. The attribute Aspect has two possible values: Perfective and Imperfective. For Aktionsart, seven categories are distinguished: States, Acts, Activities, Accidents, Events, Actions and Processes. They are Laskowski's (1998) adaptation of Vendler's (1967) Aktionsart classification to the features typical for Polish language.[6] Table 2 shows the attributes and values for verbs annotation used in our study.

3.2.3 The annotation process

Gestures and verbs were coded on separate tracks and connected by means of MultiLink option in ANVIL. Gestures were linked to the semantically affiliated verb. The verbs and gestures were closely related temporally: 80% of the verb onsets fell within stroke phase or slightly preceded it (Lis, submitted). Figure 1 shows a screen-shot of the annotations in the tool. 269 relevant verb-gesture pairs were found in the data. Intercoder agreement was calculated for the majority of the gesture annotation attributes and ranged from 0.67 to 0.96 (Lis, submitted) in terms of κ score (Cohen, 1960), i.e. from substantial to almost perfect agreement (Rietveld and van Hout, 1993).

4 The classification experiments

In the machine learning experiments we wanted to test to which extent we can predict the form of hand gestures from the characteristics of eventualities and verbs, as reflected in plWordNet's categorization. The relevant data were extracted from gesture and orthography tracks in ANVIL, and combined using the Multilink annotation. Classification experiments were performed in WEKA (Witten and Frank, 2005) using ten-fold cross-validation to train and test the classifiers. As baseline in the evaluation, the results obtained by the ZeroR classifier were used. ZeroR always chooses the most frequently occurring nominal value. An implementation of a support vector classifier (WEKA's SMO) was applied in all other cases; various algorithms were tested, with SMO giving the best results. The results of the experiments are provided in terms of Precision, Recall and F-score (Witten and Frank, 2005).

4.1 Classifying the gesture form features from linguistic information

In these experiments we wanted to test whether it is possible to predict the form of the gesture from the type of the eventuality referred to and information about Aspect and Aktionsart. The first group of experiments regards the Handshape attribute with seven possible values. In Table 3, the results of these experiments are shown. They indicate that Aspect information does not at all affect the classification of Handshape, and Eventuality Type and Aktionsart only slightly contribute to the classification (the best result is obtained using Eventuality Type annotation, F-score improvement of 0.1 with respect to the baseline, but is not significant).[7] Not surprisingly, the confusion matrix from this experiment shows that the categories which are assigned more correctly are those that occur more often in the data (*ASL_5* and *ASL_S*).

In the following experiment, we wanted to test whether Aktionsart, Aspect and Eventuality Type are related to the employment of hands in the gestures. Thus, Handedness was predicted using the

[5] The resulting frequency distribution of Type in the verb-gesture pairs: Translocation(150) and Body_Motion+Other(119).

[6] Laskowski's (1998) categories of Vendler's (1967) Aktionsart are called Classes. For the sake of simplicity, we use the term Aktionsart instead of Class to refer to them.

[7] We indicate significant results with $*$. Significance was calculated with one-tailed t-test and $p<0.05$.

Table 4: Classification of Handedness

Handedness	Precision	Recall	F-score
baseline	0.2	0.44	0.27
Aspect	0.2	0.44	0.27
Aktionsart	0.33	0.45	0.37
Type	**0.36**	**0.48**	**0.41**
all	0.35	0.47	0.40

Table 5: Classification of Iteration

Iteration	Precision	Recall	F-score
baseline	0.55	0.74	0.63
Aspect	0.55	0.74	0.63
Aktionsart	0.55	0.74	0.63
Type	0.55	0.74	0.63
all	0.55	0.74	0.63

verb related annotations. The results of these experiments are in Table 4. Also in this case, Aspect does not contribute to the prediction of gesture form. However, the results show that information about the Eventuality Type to some extent improves classification with respect to the baseline (F-score improvement: 0.14∗). The most correctly identified gestures were performed with *Right_Hand* and *Symmetrical_Hands*, which are the most frequently occurring Handedness values in the data.

In the third group of experiments, we wanted to investigate whether the linguistic categorization of verbs improves the prediction of the gesture Iteration. The results of these classification experiments are in Table 5. They indicate that no single feature contributes to the classification of hand repetition: in all cases the most frequently occurring value, *Single*, is chosen as in the baseline.

In the fourth group of experiments we analyzed whether the linguistic categorization of verbs enhances the prediction of Movement. We present the results of these classification experiments in Table 6. They show that none of the investigated verbal attributes has a relation to the Movement in gesture.

In the fifth group of experiments the relation between the linguistic categorization of verbs and the direction of the hand movement was determined. The results of these classification experiments are given in Table 7. They indicate that only Aktionsart contributes to the prediction of Direction (the improvement with respect to the baseline: 0.16∗).

Table 6: Classification of Movement

Movement	Precision	Recall	F-score
baseline	0.37	0.61	0.46
Aspect	0.37	0.61	0.46
Aktionsart	0.37	0.61	0.46
Type	0.37	0.61	0.46
all	0.37	0.61	0.46

Table 7: Classification of Direction

Direction	Precision	Recall	F-score
baseline	0.26	0.50	0.34
Aspect	0.26	0.50	0.34
Aktionsart	**0.47**	**0.55**	**0.50**
Type	0.26	0.50	0.34
all	0.47	0.55	0.50

Table 8: Predicting the Viewpoint type from linguistic information

Viewpoint	Precision	Recall	F-score
baseline	0.29	0.54	0.38
Aspect	0.29	0.54	0.38
Aktionsart	0.53	0.59	0.53
Type	**0.71**	**0.78**	**0.74**
all	**0.71**	**0.78**	**0.74**

4.2 Classifying the Viewpoint

In the following experiments we investigated to what extent it is possible to predict the Viewpoint in gesture from a) the linguistic categorization of the verb and b) from the gesture form.

In the first experiment, we tried to automatically identify the Viewpoint in the gesture from the Eventuality Type annotation. We also investigated to which extent the verb Aspect and Aktionsart contribute to the classification. The results of these experiments are in Table 8. The results confirm that there is a strong correlation between Viewpoint and Eventuality Type (F-score improvement with respect to the baseline: 0.36∗). We also found a correlation between Viewpoint and Aktionsart.

In Figure 2 the confusion matrix for the best classification results are given. Not surprisingly, the classifier did not perform well on the very infrequent category, i.e. D-vpt.

```
  a   b   c   <-- classified as
 89   0  12 |   a = C-VPT
  5   0  18 |   b = D-VPT
 25   0 120 |   c = O-VPT
```

Figure 2: Confusion matrix for predicting Viewpoint from linguistic information

In the last group of experiments we applied the SMO classifier to the data to predict Viewpoint

Table 9: Predicting the Viewpoint type from form features

Viewpoint	Precision	Recall	F-score
baseline	0.29	0.54	0.38
Handshape	**0.64**	**0.7**	**0.67**
Handedness	**0.58**	**0.64**	**0.60**
Iteration	0.67	0.57	0.44
Movement	0.55	0.55	0.43
Direction	0.67	0.57	0.44
all	**0.68**	**0.72**	**0.69**

from Handshape, Handedness, Iteration, Movement and Direction. Table 9 summarizes the results of these experiments. They demonstrate a strong correlation between the form of a gesture and the gesturer's Viewpoint: F-score improvement with respect to the baseline is $0.31*$ when all form related features are used, and all features contribute to the classifications. Handshape and Handedness are the features most strongly correlated to Viewpoint. In Figure 3 the confusion matrix for the best classification results is given.

```
   a   b   c   <-- classified as
84   0  17 |   a = C-VPT
20   0   3 |   b = D-VPT
36   0 109 |   c = O-VPT
```

Figure 3: Confusion matrix for predicting Viewpoint from form features

5 Discussion and future work

The results of our first group of experiments indicate that it is to some extent possible to automatically predict certain form characteristics of hand gestures from the linguistic categorization of their lexical affiliates. We found that the Eventuality Type extracted from wordnet categorization of verbs improves classification of Viewpoint in the co-occurring gesture. Our results are in line with Lis' (2012b) claim that the type of referent influences gestural representation. This claim has in turn been inspired by Poggi (2008) and Parrill's (2010) hypotheses.

Lis (submitted) interprets the finding in terms of Gricean Maxims (Grice, 1976), which among others state that speakers tend to convey as much relevant information in as economic way as possible. Body Motion refers to a movement of agent's body (part) not entailing displacement of the agent as a whole, which can be easily mimed with hand gestures from an internal perspective. The trajectory or spatial arrangement of Translocation eventualities, on the other hand, is less readily reenacted without the risk of hindering communicative flow between interlocutors. It can, however, be easily depicted from an external perspective with gestures drawing paths. Moreover, we have identified the form features of gestures which are most tightly related to the Viewpoint, that is Handshape and Handedness. In line with the previous interpretation, Lis (submitted) suggests that C-vpt gestures often depict interaction with an object and the hand shapes reflect grasping and holding. O-vpt gestures, on the contrary, focus on shapes and spatial extents and utilize, thus, hand shapes convenient for depicting lines, i.e. a hand with extended finger(s). It needs to be, however, further examined in how far the distribution of Handshape and Handedness in our data is motivated by the specifics of the stimuli.

Our findings also show that the type of eventuality improves prediction of Handedness. However, Eventuality Type provides a more substantial improvement in the prediction of Viewpoint, i.e. aspect of gestural representation rather than of purely physical form of gesture. This suggests that considering such representational format as an intermediate step in modeling gesture production may be appropriate. Having found that referent properties are only partially predictive of the form of iconic gesture, Kopp and colleagues (2008) consider direct meaning-form mapping to have a weak empirical support. They have instead suggested a two-step micro-planning procedure where the relationship between referent properties and gesture physical form is mediated by representational format. The present experiments do not provide an answer as to whether the two-step approach could lead to modeling aspects of eventuality gesture production. More analyses are needed, and they should be addressed in future work.

While our results indicate that Eventuality Type is the strongest predictor of gesture form, we have also found that Handedness and Viewpoint are related to Aktionsart, whereas none of the considered form features showed correlation with verb Aspect. An explanation might be that both the Eventuality Type and Aktionsart regard more inherent characteristics of eventuality, while Aspect regards the speaker's external perspective on the eventuality. It also needs to be noted that not all Aktionsart categories are equally represented in

our data.[8] The three most frequent Aktionsart categories share the feature 'intentionality,' but belong to different groups in Vendler's classification (Maziarz et al., 2011). It should be investigated in how far different Aktionsart types in our data are represented for different Eventuality Types, as that may provide a further explanation of the obtained results.

Aspect does not improve the classification for any feature. The observation that Aspect is related to Handedness (Duncan, 2002) and Iteration (Parrill et al., 2013) is, thus, not reflected in this corpus. It needs to be remembered that the relationship between Aspect and Iteration was found by Parrill and colleagues (2013) only when the eventualities were presented to speakers in the appropriate Aspect in the stimuli. Our results suggest it may not be generalizable to an overall correlation between Aspect and gesture Iteration. Moreover, Aspect is expressed very differently in the three languages under consideration (Polish - the present study, English (Parrill et al., 2013), and English and Chinese (Duncan, 2002)). Cross-linguistic differences have been found to be reflected in gesturing (Kita and Özyürek, 2003). Whether such differences in encoding of Aspect impact gestures should be, thus, investigated further.

The results of the experiments also indicate that gestural Iteration and Movement are not at all related to the linguistic characteristics of the co-occurring verb and that the only feature improving classification of gesture direction is Aktionsart. For Iteration, however, our data are biased in that single gestures are predominant, which may have affected the results. Regarding Movement and Direction, we suggest that they may be primarily dependent on visual properties of the referent, rather than the investigated factors. For example, Kita and Özyürek (2003) have found that the direction of gesture in elicited narrations reflects the direction in which an eventuality has been presented in the stimuli. The only improvement identified in our experiments in the classification of Direction (due to Aktionsart) requires further investigation.

Our results suggest the viability of the framework adopted in the paper, i.e. application of wordnet for investigation of speech-gesture ensembles. Wordnet classification of lexical items can be used to shed some light on speech-related gestural behavior. Using wordnet as an external source of annotation increases coding reliability and due to the wordnet machine-readable format, it enables automatic assignment of values. Wordnets exist for numerous languages and the approach may, thus, be applied cross-linguistically and help to uncover universal versus language-specific structures in gesture production. The findings support the viability of a number of categories in the annotation scheme used - they corroborate that the type of referent is a category relevant to studying gestural characteristics and they validate the importance of introducing distinctions among eventualities for multimodal phenomena. The experiments also identify another attribute, i.e. Aktionsart, as relevant in the framework.

It has to, however, be noted that our study is only preliminary, because the results of our machine learning experiments are biased by the fact that for some attributes certain values occur much more frequently than others in the data. Future work should address normalization as a possible solution. Moreover, our findings are based on narrational data, and need to be tested on different types of interaction. Most importantly, the dataset we used is small for machine learning purposes. Due to time load of multimodal annotation, small datasets are a well-known challenge in gesture research. Our results await, thus, validation on a larger sample. Also, cross-linguistic studies on comparative corpora should be performed.

In the present work only one type of bodily behaviors, i.e. hand gestures, was taken into account, but people use all their body when communicating. Thus, we plan to extend our investigation to gestures of other articulators, such as head movements and posture changes. In the present work only gestures referring to eventualities were considered. Lis (submitted) has recently started extending the wordnet-based framework and investigation to animate and inanimate objects.

[8] The frequency distribution of Aktionsart in the verb-gesture pairs: Activities(115), Acts(56), Actions(58), Events(23), Accidents(15), States(2), Processes(0), and of Aspect: Imperfective(179) and Perfective(126).

References

Polish WordNet. Wrocław University of Technology. http://plwordnet.pwr.wroc.pl/wordnet/.

Tennant, R. and M. Brown *The American Sign Language Handshape Dictionary.* Washington, DC: Gallaudet University Press (2010).

Alibali, M. W., Heath, D. C., and Meyers, H. J. Effects of visibility between speakers and listeners on gesture production: Some gestures are meant to be seen. *Journal of Memory and Language*, 44:159–188 (2001).

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 *Gesture and Speech in Interaction '11*, (2011).

Bressem, J. A linguistic perspective on the notation of form features in gestures. *Body – Language – Communication. Handbooks of Linguistics and Communication Science.* Berlin, New York: Mouton de Gruyter (2013).

Brinton, L. *The structure of modern English: A linguistic introduction.* John Benjamins Publishing Company (2000).

Comrie, B. *Aspect.* Cambridge: Cambridge University Press (1976).

Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46 (1960).

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

Eisenstein, J.and Davis, R. Gesture features for coreference resolution. In Renals, S., Bengio, S., and Fiscus, J., editors, *MLMI 06*, pages 154–155 (2006).

Eisenstein, J.and Davis, R. Gesture improve coreference resolution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 37–40, New York (2006).

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

Fujie, S., Ejiri, Y., Nakajima, K., Matsusaka, Y., and Kobayashi, T. A conversation robot using head gesture recognition as para-linguistic information. In *Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication*, 159–154 (2004).

Grice, H. *Logic and Conversation. Syntax and Semantics*, 3:41–58. Academic Press, New York (1976).

Jokinen, K. , Navarretta, C., and Paggio, P. Distinguishing the communicative function of gesture. *Proceedings of MLMI* (2008).

Karpiński, M., Jarmołowicz-Nowikow, E., Malisz, Z., Szczyszek, M., Juszczyk, J. Rejestracja, transkrypcja i tagowanie mowy oraz gestów w narracji dzieci i dorosłych. *Investigationes Linguisticae*, 17 (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).

Kita, S. and A. Özyürek. 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 (2003).

Kopp, S., Bergmann, K., and Ipke, W. Multimodal communication from multimodal thinking – towards an integrated model of speech and gesture production. *Semantic Computing*, 2(1):115–136 (2008).

Krauss, R. M., Chen, Y., and Gottesman, R. F. Lexical gestures and lexical access. a process model. In McNeill, D., editor, *Language and Gesture*, pages 261–283. Cambridge University Press, New York (2000).

Laskowski, L. *Kategorie morfologiczne języka polskiego — charakterystyka funkcjonalna.* PWN, Warszawa (1998).

Levin, B. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago (1993).

Lis, M. Annotation scheme for multimodal communication: Employing plWordNet 1.5. In *Proceedings of the Formal and Computational Approaches to Multimodal Communication Workshop. 24th European Summer School in Logic, Language and Information (ESSLLI'12)* (2012).

Lis, M. Influencing gestural representation of eventualities: insights from ontology. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction (ICMI'12)*, 281–288, (2012).

Lis, M. Multimodal representation of entities: A corpus-based investigation of co-speech hand gesture. PhD dissertation, University of Copenhagen (submitted).

Louwerse, M., Jeuniaux, P., Hoque, M., Wu, J., and Lewis, G. Multimodal communication in computer-mediated map task scenarios. In Sun, R. and Miyake, N., editors, *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, Mahwah, NJ. Erlbaum (2006).

Louwerse, M. M., Benesh, N., Hoque, M., Jeuniaux, P., Lewis, G., Wu, J., and Zirnstein, M. Multimodal communication in face-to-face conversations. In Sun, R. and Miyake, N., editors, *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, Mahwah, NJ. Erlbaum (2006).

Maziarz, M. Non-lexical verb synsets in upper-hierarchy levels of polish wordnet 2.0. Technical report, Wrocław University of Technology (2012).

Maziarz, M., Piasecki, M., Szpakowicz, S., Rabiega-Wiśniewska, J. and B. Hojka. Semantic relations between verbs in polish wordnet 2.0. *Cognitive Studies*, (11):183–200 (2011).

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

McNeill, D. *Gesture and Thought.* University of Chicago Press, Chicago (2005).

Melinger, A. and Levelt, W. Gesture and the communicative intention of the speaker. *Gesture*, 4(2):119–141 (2005).

Młynarczyk, A. Aspectual pairing in Polish. PhD dissertation, University of Utrecht (2004).

Morency, L.-P., de Kok, I., and Gratch, J. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, 20:70–84 (2009).

Morency, L.-P., Sidner, C., Lee, C., and Darrell, T. Contextual recognition of head gestures. In *Proceedings of the International Conference on Multimodal Interfaces* (2005).

Morency, L.-P., Sidner, C., Lee, C., and Darrell, T. Head gestures for perceptual interfaces: The role of context in improving recognition. *Artificial Intelligence*, 171(8–9):568–585 (2007).

Navarretta, C. Anaphora and gestures in multimodal communication. In *Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011)*, pages 171–181, Faro, Portugal (2011).

Navarretta, C. and Paggio, P. Classification of feedback expressions in multimodal data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10)*, pages 318–324, Uppsala, Sweden (2010).

Parrill, F. Viewpoint in speech–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).

Peirce, C. S. *Collected Papers of Charles Sanders Peirce (1931-58)*. Hartshorne, P. Weiss and A. Burks, Cambridge, MA: Harvard University Press (1931).

Poggi, I. Iconicity in different types of gestures. *Gesture*, 8(1):45–61 (2008).

Ramchard, G. Post-davidsionianism. *Theoretical Linguistics*, 31(3):359–373 (2005).

Rauscher, F. H., Krauss, R. M., and Chen, Y. Gesture, Speech, and lexical access: The Role of Lexical Movements in Speech Production. *Psychological Science*, 7(4):226–231 (1996).

Rietveld, T. and Hout, R. v. *Statistical Techniques for the Study of Language and Language Behavior*. Mouton De Gruyter, Berlin (1993).

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, 2 edition (2005).

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. Elan: a professional framework for multimodality research. In *LREC'06, Fifth International Conference on Language Resources and Evaluation* (2006).

Verbal and gestural representation of the space-time relation in multimodal communication[1]

Svetlana Mishlanova
Perm State University
mishlanovas@mail.ru
Anna Khokhlova
Perm State University
aekhokhlova@mail.ru
Ekaterina Morozova
Perm State University
kattie_mistaken@mail.ru

Abstract

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

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.

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.

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 greatest number of gestures was revealed in past events.

The greatest number of gestures in all episodes was made by 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 European languages.

Keywords: multimodal communication, gestural unit, oral narrative

Introduction

In Russia studying of gesticulation is still regarded as the area which is more likely to be interesting for psychologists, than for linguists. Thus, when studying oral speech the gestures that accompany it are simply brought out of consideration and practically not perceived as one of the sources of knowledge about what linguistic processes occur in this aspect of the oral statement. In a number of the western universities studying of basic means of gestural behavior is included into an obligatory course for linguistics and philology students, regular summer schools on gesticulation take place, as a result the linguists are accustomed to use gestural data at the very beginning of professional career. In Russia gesticulation studies are still considered as an exotic by-product of studying of oral speech.

With the development of new informational technologies the possibility of research in multimodal communication has increased. Digital video filming allows to carry out comprehensive segmentation of oral narrative and, thus, gives an opportunity of drawing up the multimedia corpus, allowing to investigate nonverbal components of oral speech.

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

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.

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.

Relevant researches

We were inspired by the research of the psycholinguist Daniel Casasanto that deals with studying of gestures, expressing space-time metaphor. In "Hands of Time" he writes that when one speaks about time, he usually uses the words expressing some position in space, and it is thus logical to assume that gestures used by the speaker indicates the way the speaker imagines time is settled down in space. According to the results of his research, English-speakers have an accurate model of a horizontal time axis on which the past is located on the left, and the future – on the right from the speaker. However, this model represents only spontaneous gesticulation. When examinees gesticulated intentionally, this axis took a vertical (sagittal) position, i.e. the past is behind and the future is ahead. The similar model of the time axis is built by Alan Cienki and Cornelia Müller in the number of their research works.

In our research we considered only spontaneous gesticulation, for this research deals with the representation of space-time relation is *oral* narrative.

Method

1 Materials

As the material for this research we used the video recordings of the interviews of the Perm state national research university students. In this particular research 14 students' interviews were analyzed.
The task for informants was to tell about anything that makes them happy. As the task given to the students presupposes some bias in the direction of talk about past, it is clear that the whole amount of events they mention refers to the past time. However, there is a sufficient number of students that associate happiness with present and future events.

2 Participants

Participants were 8 female and 6 male students. Their age varies from 17 to 19. At the time of the experiment they all were first year students of the Perm state university.

3 Procedure

Step 1.
First we divided the narrative into two modalities to consider the verbal part separately from the gestural one.
To make it more convenient the text of each of the14 interviews was transcribed. We divided each interview into three groups of events relating to the past, the present and the future.
Step 2.
After that we divided the events of the past, the present and the future into thematic groups according to the following criteria:
• general semantics of the sentences,
• words and expressions having temporal semantics, marking this or that time, for example adverbs (*long ago, soon, now, earlier, later, then*)
• the direct nomination of this or that time.
Thus, we received 65 events:
• 37 events relating to the past,
• 22 events relating to the present,
• 6 events relating to future time.
Step 3.
The next step is the separation and calculation of gestural units in each episode and their division into three groups: right-handed, two-handed and left-handed gestures.
In this particular research we did not take into consideration the type of gesture, however this aspect is definitely going to be regarded in the following researches.
Step 4.
The final step is the comparison of verbal and gestural representation of the space-time relation.

4 Data analysis

We analysed both verbal and gestural codings. During the data analysis we found out that 57% of events (37) belong to the past, 34% (22) - to the present, and 9% of events (6) belong to future time.
It made it possible to draw a conclusion that for the students aged 17-19 years the happiness associates with the past in most cases.
As each of interrogated students told about his or her happy moments from the real life experience, we also divided the events according to the referent of happiness further to compare, what particular events the students associate with the past, the present, and the future.
As a result we allocated 9 types of referents:
a) Close people – 28% (18);
b) Feelings and emotions – 26% (17);
c) Study and work – 14% (10);
d) The purposes – 12% (8);
e) The nature – 5% (4);
f) Surprises – 5% (4);
g) Travel – 4% (2);
h) Creativity – 3% (1);
i) Pets – 3% (1).

In the past events the referents "Study" (7 - 70%), "Feelings" (15 - 89%), "Surprise" (4 - 100%), "Nature" (4 - 92%) prevail; in present events - the referent "Relatives and friends" (9 - 50%); the referents "Pets" and "Creativity" are most frequent in future events.
According to the obtained data the interrogated girls gesticulate much more actively, than the interrogated young men. Out of 122 recorded gestures 107 were made by girls, and only 15 – by the interviewed young men. Thus, out of total amount of gestural units 88% (107) gesticulation was received from female and only 12% (15) from male informants.
All in all we recorded 122 well expressed gestural units. Out of them 80 gestures were referred to "past events", 35 gestures - to "present events", in "future events" 7 gestures are recorded.
Thus, 80 (65%) gestures - past events, 35 (29%) - present events, 7 (6%) of gesticulation is presented in future events.
Out of 80 gestures relating to events of the past, there are 19 (24%) left-handed gestures, 11 (14%) right-handed gestures, 50 (62%) two-handed gestures.
Out of 35 gestures relating to events of the present, there are 3 (9%) left-handed gestures, 4 (11%) right-handed gestures, 28 (80%) two-handed gestures.
Out of 7 gestures relating to events of the future, there is one (14%) left-handed gesture, 4 (57%) right-handed gestures, 2 (29%) two-handed gestures.
It turns out that the most active gesticulation is shown in past events.

Thus, it is possible to draw a conclusion that students gesticulated most actively speaking about the past and vice versa, speaking about the future students gesticulated least actively.

1. Informants gesticulated with two hands more often when speaking about the present events (50 - 80%) than about the events of the past (63%) and the events of the future (28%) ($\chi 2 = 20.45$; $df = 2$; $p = 0.001$).
2. Informants gesticulated with their left hand more often when speaking about the events of the past (24%) than about the events of the future (14%) and the present (9%) ($\chi 2 = 7.32$; $df = 2$; $p = 0.05$).
3. Informants gesticulated with their right hand more often, whereas speaking about the events of the future (58%) than about the events of the past (28%) and the events of the present (14%) ($\chi 2 = 18.45$; $df = 2$; $p = 0.001$).

Results

1. According to the obtained data the frequency of gesticulation depends on gender accessory of an informant.
2. The biggest number of gestures (37gestures - 65%) was revealed in past events.
3. The biggest number of gestures in all episodes (80 gestures - 65,5%) was made with two hands.
4. 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.
5. 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.

References

1. Casasanto, Daniel. 2009b. When is a linguistic metaphor a conceptual metaphor? In V. Evans & S. Pourcel (eds.), New Directions in Cognitive Linguistics, 127–145. Amsterdam: John Benjamins.
2. Cienki, Alan. 1998. Metaphoric gestures and some of their relations to verbal metaphorical expressions. In J. Koenig (ed.), Discourse and Cognition: Bridging the Gap, 189–204. Stanford: CSLI Productions.
3. McNeill D. (2005). Gesture. The Cambridge encyclopedia of the language sciences. Hogan, P. C. (ed.). Cambridge: Cambridge University Press.: 344–346
4. Müller C. A Dynamic View of Metaphor, Gesture and Thought. // Gesture and the Dynamic Dimension of Language. Essays in honor of David McNeill. Eds. S. Duncan, J. Cussell, E.Levy. Pp. 109-116. Amsterdam/Philadelphia, 2007.
5. Casasanto, D. & Jasmin, K. (2012). The Hands of Time: Temporal gestures in English speakers. Cognitive Linguistics, 23(4), 643 – 674. URL: http://www.casasanto.com/papers/Casasanto&Jasmin_HandsOfTime.pdf
6. Casasanto, D. (2010). Space for Thinking. *In Language, Cognition, and Space: State of the art and new directions*. V. Evans & P. Chilton (Eds.), 453-478, London: Equinox Publishing. URL: http://www.casasanto.com/papers/Casasanto_SpaceForThinking.pdf

[1] The research is funded by the Russian humanitarian foundation (projects № 14-13-59007, №14-16-59007)

Overlaps in Maltese Conversational and Task-Oriented Dialogues

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

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

Abstract

This paper deals with overlaps in spoken Maltese. Overlaps are studied in two different corpora recorded in different communicative situations. One is a multimodal corpus involving first acquaintance conversations; the other consists of Map Task dialogues. The results show that the number of overlaps is larger in the free conversations, where it varies depending on specific aspects of the interaction. They also show that overlaps in the MapTask dialogues tend to be longer, serving the function of establishing common understanding to achieve optimal task completion.

Keywords: overlaps, MapTask dialogues, face-to-face conversations, Maltese

1 Background

We know that overlap, the phenomenon by which two or more speakers talk over one another, plays a significant role in spontaneous interaction (Schegloff, 2000). We also know that the amount and function of overlap varies depending on the type of communicative situation (Cetin and Shriberg, 2006; Adda-Decker M. et al., 2008; Campbell et al., 2010).

Several factors seem to correlate with the occurrence of overlap. One is the existence of predefined roles: for instance Cetin and Shriberg (op. cit.) observe that in chaired meetings, in which the general interaction is controlled by the chair, there is little overlap. Conversely, the more spontaneous and free the conversation, the more overlap can be expected. Moreover, Campbell et al. (op. cit.) claim that familiarity is also an important factor, such that the more familiar people are with each other, the more overlap they produce when they talk.

This paper examines overlap in two different corpora of spoken Maltese: the MAMCO multimodal corpus of first acquaintance conversations, and the Maltese Map Task dialogues. The two corpora differ substantially in ways that are expected to be directly related to the occurrence of overlap. Thus the paper aims to verify previous claims about the relation between overlap and communicative situation. It also provides an analysis of overlaps in a type of situation, first acquaintance dialogues, which has not been studied earlier in this respect[1].

The aims of the study are to see (i) how frequent overlaps are in the two corpora; (ii) what types of overlap occur; (iii) how overlaps are distributed between the speakers; (iv) whether the occurrence of overlap varies as the interaction proceeds. In general, we are interested in investigating whether there are systematic differences in the two corpora due to different features such as the presence or absence of pre-defined roles, and the nature of the conversation.

2 Overlaps: definition and types

An overlap is a stretch of time of variable duration where two or more conversation

[1] We report on a pilot study in Vella and Paggio (2013).

participants speak over one another, and which may or may not result in a change of speaker. In what follows, overlap is always between two speakers, since all the interactions examined are dyadic.

Different types of overlap may also be distinguished based on different functional categories. In our corpora the following three general types can be noted:

1. *Feedbac*k-related overlap (*ACKNOWLEDGE* move in Carletta et al., 1997): there is no competition for the floor and change of speaker is possible but not necessary. This can be lexical (e.g. *orrajt*/*owkey* 'all right, okay', *sewwa*/*tajjeb* 'good') or quasi-lexical (e.g. *mhm*/*eħe*).
2. *Question*-related overlap, especially in answers involving a yes or a no (*REPLY-YN* in Carletta et al., 1997): the current speaker relinquishes the floor and a change of speaker is expected. (Overlap is less likely, though not impossible with *wh*-questions – *REPLY-W* in Carletta et al., 1997).
3. *Competitive* overlap: the two speakers are competing for the floor. In some cases, this competition seems to result from an attempt at establishing common ground (mutual understanding, a common topic, etc.). The current speaker can retain or relinquish the floor.

In section 5 we will give examples of the various types, and discuss how they relate to the communicative situation specific to the two corpora investigated.

3 The corpora

The two corpora used in this study are the multimodal corpus of Maltese MAMCO and the Maltese Map Task dialogues. In Vella and Paggio (2013), which this paper builds upon, only one example from each corpus was considered. This study, by contrast, considers both corpora in their entirety.

3.1 The multimodal conversational corpus

The multimodal corpus of Maltese MAMCO consists of twelve video-recorded first acquaintance conversations between pairs of Maltese speakers.

Twelve speakers participated (6 females and 6 males). Each speaker took part in two different conversations, one involving another female and another involving a male interlocutor. An important prerequisite was that the two participants had not met before: they were instructed to try to get acquainted during the conversation. They could, however, freely decide what to talk about. Recording was stopped after about 5 minutes. All conversations were recorded in a studio using three different cameras, as shown in Figure 1. The general set-up was very similar to the one used in the Nordic NOMCO corpus (Paggio et al., 2010) so that it will be possible in future to use the corpora for inter-cultural comparisons.

Figure **1**: Screenshots from the MAMCO corpus.

3.2 The Map Task dialogues

The eight Maltese Map Task dialogues form part of the MalToBI corpus (Vella and Farrugia, 2006), which was designed to be representative of spoken Standard Maltese, participants being carefully selected with a view to balance in terms of age, sex and educational background. The Maltese Map Task design is similar to that used for 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 – which is absent from the Follower Map – to the participant in the Follower role, who has to draw the route following the Leader's information. The locations on the Maps are not identical, so that negotiation is sometimes required. The Maltese Map Task dialogues involve 16 speakers (8 females and 8 males): half of the speakers of each gender fulfil the Leader role and the other half the Follower role.

Contrary to other similar collections, in the Maltese Map Task corpus all participants could see each other. As a result, the Maltese Map Task data are directly comparable to the MAMCO data in that non-verbal as well as verbal means of communication were available to speakers for use (only audio recordings of the Maltese Map Task data are available, however).

3.3 Initial comparison of the two corpora

Similarities and differences between the two corpora are summarised in Table 1, reproduced here after Vella and Paggio (2013).

MAMCO	Map Task
Dialogues	Dialogues
Subjects standing at comfortable speaking distance	Subjects sitting facing each other with two tables between them
Lapel microphones	Unidirectional microphones
Cameras	No cameras
Can see each other (entire body)	Can see each other (face and torso)
Talk freely	Have to solve a task
No predetermined role	Different roles
Do not know each other	Familiarity not an issue

Table 1: Similarities and differences between the corpora

The last three rows in the table refer to the most interesting features from the point of view of this study. In MAMCO, there are no pre-defined topics and no task (we don't consider the sole instruction to get to know each other as a real, well-defined task), and participants have no predetermined roles in the dialogue. In the Map Task dialogues, on the contrary, participants have to complete a task and have been assigned specific roles for how to achieve this goal. As for familiarity, there is also a difference in that the MAMCO participants' starting point is that they do not know each other. Participants in the Map Task dialogues do know each other, however they do not talk about personal matters. Therefore, in a sense familiarity is not really an issue in those interactions.

A simple way to compare the two corpora is to look at how much participants speak, and how speaking time is distributed between the two speakers. In MAMCO, the average speaking time per participant is 248.56s. There is no clear pattern as to which participant speaks the most: sometimes Speaker 1 does, sometimes Speaker 2. The difference in speaking time between the two speakers is shown in Figure 2 in terms of seconds and time percentage. Bars above zero indicate predominance by Speaker 1 and those below by Speaker 2.

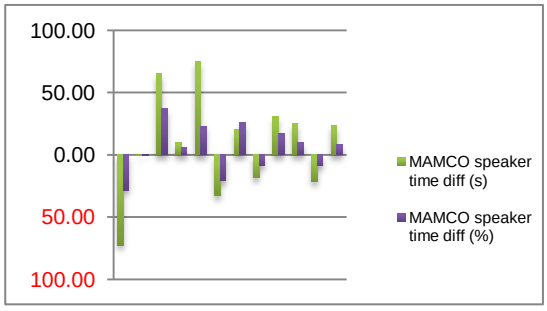

Figure 2: Distribution of speaking time in MAMCO: Bars above zero indicate predominance by Speaker 1 and bars below by Speaker 2.

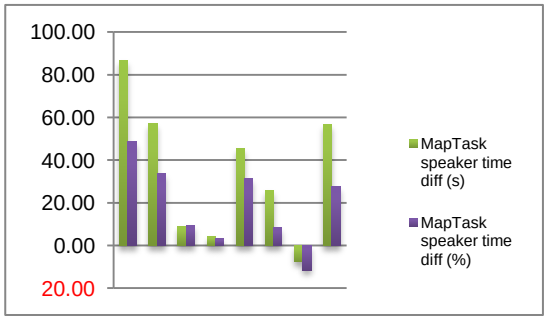

Figure 3: Distribution of speaking time in the Map Task dialogues. Bars above zero indicate predominance by Speaker 1 and bars below by Speaker 2.

The picture is quite different for the Map Task dialogues (Figure 3). The average speaking time, 257.83s, is comparable, but in this case one of the speakers nearly always has the most speaking time. Not surprisingly, this speaker is Speaker 1, always the Leader in these data. In the one exception in which the Follower – Speaker 2, bars below zero – speaks more than the Leader, this is due to this speaker often replicating part of the instruction given before adding an own comment. This noticeable difference in the way the two participants share speaking time in the two corpora is one of the consequences of the different type of communicative situation. Based on this difference, and on the claims made in the sources quoted above about how pre-defined roles and degree of familiarity impact the occurrence of overlap, we would expect the following facts concerning overlap to hold:

- a greater degree of overlap in the MAMCO conversations because both participants have to negotiate the floor;
- fewer overlaps resulting in change of speaker in the Map Task dialogues, since we expect the Follower to overlap in order to confirm, and the Leader to keep the floor;
- an increase in overlapping as the dialogue proceeds, as speakers get more comfortable with the situation and also more familiar with each other.

4 Quantitative analysis of overlap

4.1 Degree of overlap

The first dimension along which we want to compare the two corpora is the degree of overlap. We looked at this in several ways by measuring (i) the number of overlaps, (ii) the proportion of overlap time over total conversation time, and (iii) the length of the overlaps. These sets of measures are shown in Figures 4-6. For each measurement, the box on the left represents MAMCO, and the one on the right the Map Task dialogues.

Figure 4 shows that there is a significant difference in the average no. of overlaps (Two Sample t-test: t = 3.6413, df = 14.84, p-value = 0.002451), and that the difference is in the expected direction, with MAMCO showing more overlap as well as more variation in degree of overlap in the various conversations. The picture for the Map Task dialogues is much more uniform with the exception of a single outlier.

The difference in the proportion of overlap time between the two corpora, shown in Figure 5, is also significant (Two Sample t-test: t = 3.3975, df = 14.393, p-value = 0.004187). The explanation is that on average the length of the overlaps in the Map Task dialogues is higher, although not in a statistically significant way (Figure 6).

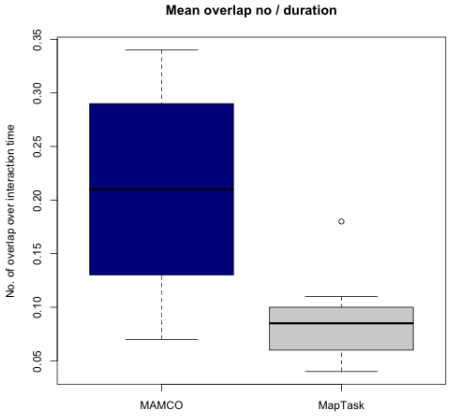

Figure 4: Overlap number over duration

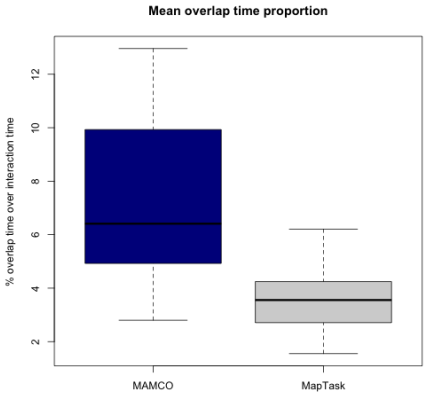

Figure 5: Proportion of overlap time over duration

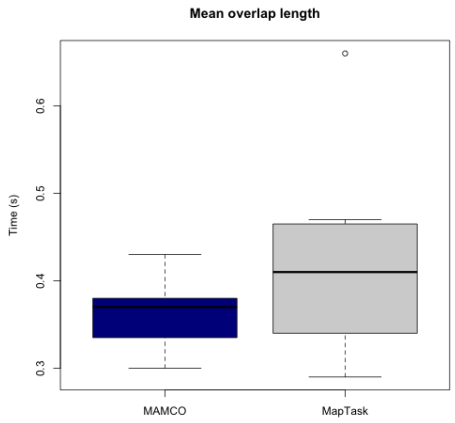

Figure 6: Overlap length in the corpora

The length of an overlap can be seen to relate to the functional types listed earlier. We hypothesise that the so-called *competitive* type of overlap, in which speakers compete for the floor, sometimes in an attempt at homing in on a topic of common interest, tends to be longer. As will be shown in section 5, examples of this type of overlap occur in the Map Task dialogues in places where there is breakdown of communication, or a misunderstanding of an instruction on the part of the Follower. In MAMCO, on the contrary, there are no inherent reasons for speakers needing to interrupt each other to clarify misunderstandings.

4.2 Overlap and change of speakers

To verify our second prediction, we measured the proportion of overlaps resulting in a change of speaker (Figure 7). As expected, the proportion of overlaps resulting in speaker change is (slightly) larger in the MAMCO corpus.

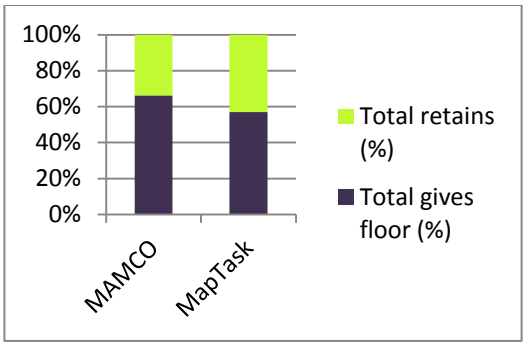

Figure 7: Overlap and change of speakers

Contrary to our expectations, however, in both corpora both speakers take the turn equally often when there is a change of speaker.

4.3 Overlap and familiarity

Finally, we wanted to verify whether increase in familiarity is proportional to amount of overlap. We tested this by looking at whether overlap increases as the dialogue progresses. We chose 60 seconds as a threshold, corresponding more or less to one third of the interaction, hypothesising that the participants would by then have broken the ice. Interestingly, there is no effect in the Map Task dialogues, whereas we see in fact a decrease in MAMCO (Figure 8).

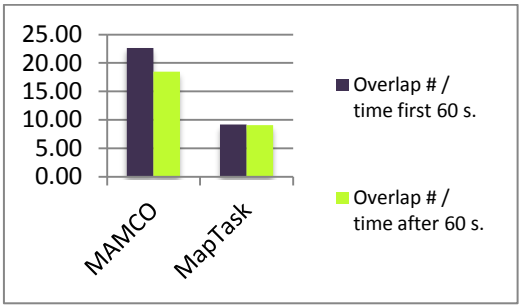

Figure 8: Overlap progression during the dialogues

It is debatable whether the effect we see is a counterexample to the familiarity effect observed by Campbell et al. Arguably, speakers in the MAMCO dialogues are not familiar with each other after 60 seconds of interaction. The decrease in overlap, therefore, has probably nothing to do with familiarity, but is due to the speakers adjusting their turn taking mechanism to each other after having broken the ice, introduced each other and familiarised themselves with the situation, although other factors cannot be excluded at this stage. In this sense it is significant that this does not happen in the Map Task dialogues, where what is important is that the task assigned be completed. In these dialogues therefore, whatever adjustment creates this effect may be overridden by the need to move the interaction forward, a goal which overlaps may in part help achieve (see also section 5.1).

5 Overlap functions and examples

Examples of the different functional types of overlap identified in section 2 above are presented below.

5.1 Feedback-related overlap in the context of feedback

It is worth noting at this point that negotiating turn-taking in dialogues is a logical necessity. In addition to this, providing one's interlocutor with feedback is also an important element if interaction is to succeed, and feedback and turn-taking are often related. Examples of turn-taking involving smooth changes between the two speakers one of whom is providing the other with feedback, are therefore, not unexpectedly, frequent in the data analysed. One such example is the following from one of the Map Tasks:

SP1: *Mela* (0.2) *nitilqu mill-Bajja ta' Ray* (0.6)
SP2: *Sewwa.* (0.7)
SP1: *għan-naħa tat-Tramuntana* (0.1)
mill-ewwel (0.1)
SP2: *Mhm.* (0.7)
SP1: *fejn fiha* (0.6) *tgħaddi bejn* (0.2)
Triq Mannarino

SP1: So (0.2) we leave from Ray's Bay (0.6)
SP2: Right. (0.7)
SP1: towards the North (0.1)
from the beginning (0.1)
SP2: Mhm. (0.7)
SP1: where in it (0.6) you pass through (0.2)
Mannarino Street

The numbers in parentheses in these examples indicate the duration of both inter- and intra-speaker pauses. Exchange of information in this example is generally evenly paced with both inter-, and on occasion, also intra-speaker pauses with a duration of 0.6-0.7s. Examples of feedback items in the above include the lexical items *sewwa* 'right' and the quasi-lexical *mhm*. Other frequent lexical or quasi-lexical feedback elements include *iva* 'yes' (also *ija*' or *iwa*), as well as *le* 'no', *orrajt* 'alright', *owkey* 'okay' and *tajjeb* 'good'.

Feedback of the sort illustrated above, however, can frequently be seen to involve overlap in both corpora. Sections of overlap in the examples provided below are enclosed within square brackets and the overlapping elements in the original indicated in bold.

A first example from the MAMCO corpus is given below:

SP1: *għandi z-zijiet hemmhekk.*
SP2: ***[Mhm.***
SP1: ***In-nan]*** *na: (0.2) minn Bormla*

SP1: I have aunts there
SP2: [Mhm.
SP1: My grand] ma's (0.2) from Bormla

SP2's *Mhm* in this example overlaps with part of SP1's continuing narrative on where different relatives come from without: there is however no competition for the floor. A second example, this time from the Map Task corpus is the following:

SP2: *jew Dar Millennia*
SP1: Dar Millennia **[sewwa**
SP2: **jew]** Vjal il-Mara (0.3).

SP2: either Millenia House
SP1: Millenia House [right
SP2: or] Women's Alley (0.3).

Again in this example, although it may appear, at a first glance, that SP1 is attempting to take the floor, this is in fact not the case since his contribution consists simply of a reaffirmation of the information he's been given by SP2 (*Dar Millenia*), followed by the lexical backchannel *sewwa* 'right'. It is in the context of this reassurance that transfer of information has been successful that SP2 comes in with her overlapping additional bit of information *jew Vjal il-Mara*.

To conclude on overlapping in the context of feedback, this type of overlapping in interaction often involves one speaker reassuring the other that transfer of information has been successful, which in turn, serves as a way to move the interaction forward. In most cases it does not

involve a change of speaker, but even where a change of speaker is involved, the overlap is co-operative rather than competitive.

5.2 Question-related overlap

In both corpora analysed, overlap also occurs when questions are answered. This is the case in the context of answers to both yes-no and wh-questions. An example from the MAMCO corpus is the following:

SP2: *Imma s-sitt waħda teżi* (0.2)*, hux ve* ***[ru?***
SP1: ***Ija.]***

SP2: But the 6th one's a thesis (0.2), isn't that [so?
SP1: Yes.]

An example from the Map Task corpus is:

SP2: *Minn Triq Mannari* ***[no***?
SP1: ***Ija.]***

SP2: From Mannari[no Street?
SP1: Yes.]

In these and similar examples, including examples involving wh-questions, although there is a change of speaker, there is no competition since, by virtue of the fact of asking a question, the current speaker is relinquishing the floor. Overlap in this context suggests engagement rather than competition, and once again serves the purpose of propelling the interaction forward. A characteristic of this type of overlap, which occurs in both corpora, is that it is short in virtue of the fact that the speaker who asks the question is relinquishing the floor on their own accord.

5.3 Competitive overlap

The third type of overlap identified in section 2 is competitive overlap. This can result in a change of speaker but does not always do so.

An example from the MAMCO corpus in which overlap leads to a change of speaker is given below:

SP1: *Mela mill-università for* ***[si***
ġieli rajt wiċċek
SP2: ***Imma: ee]*** (0.33) *għandi z-zijiet hemmhekk.*
[In-nanna:
SP1: ***Mhm.]***

SP1: So it's from University may **[**be
that I know your face
SP2: But ee] (0.33) I have aunts from there.
għandi z-zijiet hemmhekk.
[My grandmother
SP1: Mhm.]

In the above, SP2's overlap with SP1 results in SP2 succeeding in taking the floor.

Although examples similar to the above, in which competitive overlap leads to a change of speaker, can also be found in the Map Task data, the purpose of such examples in the Map Task dialogues seems different, in that speakers do not compete for the floor to contribute to the conversation with their personal stories or opinions, but to ensure that the task is completed successfully. Let us examine the following example from the Map Task corpus:

SP1: ***[Trid issib*** (0.1)
SP2: ***hemm naqra bogħod]***
[biex ngħaddi
SP1: ***Eħe.]*** (0.5)
SP2: *minnha.*
SP1: *Eħe.*

SP1: [You need to find (0.1)
SP2: it's a bit far **]**
[to go through
SP1: Eħe] (=Yes). (0.5)
SP2: from it
SP1: Eħe (=Yes).

The above contains two instances of overlap. The first of these is competitive and results in SP1, who was in the process of giving an instruction (*Trid issib)*, relinquishing the floor to SP2. Having lost the floor however, SP1 recalibrates, as it were. She proceeds immediately to acknowledge that yes (*Eħe*), the location they need to move to is rather far away, overlapping with SP2 again when she does this, but making no further attempt, at least at this point, to regain the floor.

A further example will serve to illustrate the complexity involved:

SP1: *Itla' l fuq.* (0.73)
SP2: *Mela.* (0.17)
SP1: ***[Fid-direzzjoni***
SP2: ***Tini sekonda]*** *ċans ta ħa nsib l-bajja* (2.47)
Iwa (0.21) *sibna l-bajja* (0.75)
Trid ***[titla 'l fuq***
SP1: ***Titla 'l fuq]*** (0.38)
*fid-****[direzzjoni***
SP2: ***Sewwa.]***
SP1: *ta' Triq Mannarino.*

SP1: Move upwards. (0.73)
SP2: So (0.17)
SP1: [In the direction of
SP2: Give me a second] to find the bay (2.47)
Yes (0.21) we've found the bay (0.75)
You need [to go up
SP1: You go up] (0.38)
in the [direction
SP2: Right.]
SP1: of Mannarino Street.

In the first overlap in this example SP1 has the floor. SP2 signals, using the discourse marker *Mela* (frequently used as a means of 'resetting', in preparation to initiate a new move), that he would like to take the floor: there is no overlap up to this point. SP1 however does not get the message, and continues giving directions (*fid-direzzjoni*). At this point, SP2 overlaps, and takes the floor specifically to say that he needs time to carry out the instruction he had been given. Once he has done this, he picks up from where he had interrupted SP1's instruction to *Itla 'fuq*, by saying *Titla' 'l fuq*. SP1 realises that he is ready to move on and overlaps with him once more, once again taking the floor and repeating the instruction *Titla' 'l fuq*. It is now clear he is ready to follow. There is one final overlap involving SP2 providing feedback, with no further change of speaker.

The examples illustrated above suggest that it may be too simplistic to suggest that overlap with change of speaker is always the result of competition for the floor, at least for the kinds of data, such as Map Task data, where speakers are engaged in a collaborative task. Or at least, competition for the floor here serves a different function than in conversational data, in that the speakers are eager to make sure that they understand each other in order to complete their joint task.

In an attempt at getting a preliminary indication of whether or not competitive overlap tends to be longer than non-competitive overlap, we examined overlaps in the data which exceeded (the arbitrarily chosen threshold of) 60s in duration. In line with the finding that the number of overlaps in the MAMCO data is greater than in the MapTask data, there were also more lengthy overlaps in the MAMCO data than in the MapTasks.

Preliminary findings do not, however, support the hypothesis of a greater tendency for longer overlaps to be competitive. Straightforward feedback-related overlap with no competition and no change of speaker occurred in more than half the cases examined (8/13). In two further instances of feedback-related overlap, a change of speaker occurred, but without competition. In the first of these, a (relatively long) pause (0.52s) followed the feedback – the speaker responsible for the overlap consequently felt the need to get the interaction going again. In the second instance a new element of information was provided following the feedback, with the speaker immediately relinquishing the floor once this information had been communicated.

Three of the 13 cases of longer overlap could, indeed, be classified as examples of competitive overlap. A complete analysis of the relation between length and competitive overlaps, however, presupposes functional labelling of all the examples in the corpora, a task which we leave for future research.

6 Conclusion and future work

In conclusion, we have shown that overlaps in both the corpora analysed are used (i) to provide feedback during the dialogues; (ii) to anticipate

answers to questions that are being asked, and (iii) to take the floor. The degree and length of overlapping is different in the two corpora, reflecting the different communicative situation involved.

As we were expecting, a larger degree of overlapping occurs in the free MAMCO conversations. In addition, slightly more overlaps in MAMCO result in a change of speaker, which also confirms the more dynamic nature of these conversations, in which neither speaker has a pre-defined role in the dialogue. However, the degree of overlapping is seen to decrease slightly as the conversations proceed, probably due to the participants adjusting to each other's turn taking mechanism.

Conversely, less overlapping and less change of speaker in connection with overlaps occur in the Map Task dialogues, where the underlying task and the roles assigned to the two participants provide for a more rigid structure. A peculiar feature of these dialogues, by contrast, is the occurrence of relatively long overlaps in which the dialogue participants try to recover from communication breakdowns in order to be able to complete their task.

In this paper, overlaps were studied only from the point of view of the speech contributions. In future, we would like to extend the analysis to non-verbal behaviour. For example Navarretta (2013) discusses how multimodal behaviour can be used to predict overlaps on the Danish NOMCO corpus, which, as was pointed out earlier, has a very similar setting to MAMCO. It would be interesting to compare her findings with similar observations from the Maltese data, in both the corpora described here.

Acknowledgments

We would like to acknowledge Sarah Agius, Marija Debono and Luke Galea, who transcribed the MAMCO conversations. This work was possible through funding from the University of Malta's Research Grant LINRP06-02.

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 Sixth 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.

Carletta, J., A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon and A. Anderson, The reliability of a dialogue structure coding scheme, *Computational Linguistics* 23 (1): 13-32, 1997.

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.

Navarretta, C. 2013. Predicting speech overlaps from speech tokens and co-occurring body behaviours in dyadic conversations. Proceedings of ICMI 2013: 157-164.

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.

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.

Vella, A. and P. Paggio. 2013. Overlaps in Maltese: a comparison between map task dialogues and multimodal conversational data, in Allwood et al. *NEALT Proceedings. Northern European Association for Language and Technology, 4th Nordic Symposium on Multimodal Communication*, November 15-16, Gothenburg, Sweden, pp. 21-29.

Cognitive processes and multimodal communication

in the parody of politicians

Isabella Poggi
Dipartimento di Filosofia, Comunicazione
e Spettacolo
Roma Tre University
Via Ostiense 234 – 00146 Rome
isabella.poggi@uniroma3.it

Francesca D'Errico
Facoltà di Psicologia
UniNettuno Telematic University
Corso Vittorio Emanuele II, 39
00186 Rome - Italy
f.derrico@uninettunouniversity.net

Abstract

To single out the cognitive processes implied in the production of a parody, viewed as a distorted imitation of a text or behavior aimed at eliciting laughter and mocking someone, a corpus of parodies of politicians has been collected and multimodal communication analyzed through a devoted annotation scheme. Analysis allows to distinguish between surface and deep parodies, to single out the steps required for making a deep parody when the bare imitation of the Target is not enough for the Parodist's satiric goals, and to see the intertwining of various modalities in conveying the crucial information of a parody: identification and characterization of a Target and of its flaws through allusion to some event.

1 Introduction

A common activity in everyday life, entertainment, and political satire, is to make parodies. Students make parodies of their teachers, humor writers make parodies of poems or songs, comedians perform parodies of politicians. This work explores the cognitive and communicative processes underlying the production of parodies in political satire.

2 What is parody

Holman and Harmon [1] define parody as an imitation intended to ridicule or criticize, that to be understood requires familiarity with the original object, and to be effective has to "sound true", that is, faithful to the original. Rose [2; 3] sees parody of literary works as the comic reworking of preformed material through their partial imitation or evocation in a comic manner that marks the ambivalence of the parodist's attitude to the object of criticism. Being a case of intertextual work, the parody contains two texts-worlds, and the reader must understand the comic satiric relationship between them. [3; 4; 5].

Parody is not a simple imitation, but an "approximation" to an original source, in which, like in sarcasm, "the subject is treated in a contradictory manner: elevated subjects are debased and low ones are elevated" [4; 5; 6]. Bachtin [7: 76] views the parodistic act as "an arena of conflict between two voices", split from one another in a hostile contrast, with the second voice representing a "semantic authority" with which the audience is expected to agree.

Rossen-Knil and Henry [8] mention four pragmatic aspects of parody: (1) the intentional verbal representation of the object of parody, (2) the flaunting of the verbal representation, (3) the critical act, and (4) the comic act.

The techniques used by the parodist to refashion an older text or image range from caricature to substitution, addition, subtraction [9], exaggeration, condensation, contrast, and discrepancy [5].

Luttazzi [10] attributes two goals to parody, informing and deforming, the latter often using "bodily reduction" to physical needs, with the aim of dissacrating and destroying hierarchies, mixing sacred and secular, and making fun of boasting characters and their arrogance in a blasphemous way.

Various authors [2; 3; 5; 9], stress how the parodistic act depends on the successful interaction between parodist and audience, that not only needs to acknowledge the Parodist's "authority" and moralistic intention, but also must know vices and virtues of the Target, especially when the parody is focused on his/her body and verbal features (tics, stuttering…) that are the trigger of the comic part. In brief, a verbal parody is a highly situated, intentional, and conventional speech act that re-presents some object but flaunts the re-presentation to convey humorous criticism [11; 12].

3 The Parody of politicians.

Based on [12] and [13], where ridiculization and mockery of politicians are viewed as "discrediting moves", we define a parody as a communicative act – a text or a verbal or 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 about either the behaviour or one who performs it. A text, a discourse, a rite, an institution, and finally a person may all be an object of parody. In the parody of a person, the Parodist P imitates a Target T by reproducing his/her traits and / or communicative or non-communicative behaviors, but in a distorted, for example an exaggerated or misleading way, that highlights the Target's flaws; to do so the parodist must single out the most characterizing features of T's physical traits or behaviors, and imitate them while exaggerating or subtly changing them in such a way as to make them appear ridicule. As mentioned, parody necessary makes use of allusion – the device of indirectly referring to something without explicitly mentioning it – in order for the Audience to recognize the Target and the reasons for the Parodist's criticism.
In political satire, a comedian (Parodist) performs a distorted imitation of a politician (Target) to make fun of him/her, aiming at cruel criticism or benevolent irony.

[12] posit four defining features of the parody of politicians: 1. Similarity to the Target; 2. Allusion, 3. Distortion of the similarity, aimed at stressing ridicule aspects of the Victim and eliciting laughter, 4. Induction of inferences implying a negative evaluation, that in the judgment of politicians may concern three criteria: *benevolence* (caring the electors' goals, not working on behalf of one's own interest, being trustworthy, honest, ethical), *competence* (expertise, knowledge, planning and reasoning skills), and *dominance* (capacity of winning in contests, influencing others, imposing one's will).

4 "Surface vs. "deep" parody.

Often the Parodist's imitation is not a faithful – albeit distorted – reproduction of the Target's actual visible or audible behaviours, but rather a "deep" imitation: the parodist extracts a – sometimes hyperbolic or surreal – submersed ridicule aspect of the Target's personality, and imitates the behavior that would stem out of it. A such device is exemplified in the comedian Maurizio Crozza's parody of Matteo Renzi, who in 2012, before becoming the Italian Prime Minister in 2014, was an emerging leader of the Italian Democratic party. Crozza impersonates Renzi as a young boy hopping around and jumping the rod. This alludes to Renzi's struggling against the old leaders of his party and presenting himself as an "enfant terrible" carrying new ideas and a new young atmosphere. Of course, Renzi has never shown while jumping the rod, but his general attitude can well be represented by that (fake but funny) image. In this case the Parodist does not reproduce actual visible or audible features of the Target's traits or behaviors, but ones that might plausibly be displayed by the Target, given his/her general attitude. To do so, P must find out the core of T's personality, and imitate those traits and behaviors that may plausibly stem out of it, even if T has never actually exhibited them. This distinguishes a "deep" from a "surface" parody.

5. Towards a cognitive model of Parody

Based on the above definition, we made a hypothesis about the cognitive processes implied in the production of a political parody.

5.1.Hypothesis

The sequence of steps gone through in making a parody can be split in two phases, devising what humorous aspects to highlight in the Target, and deciding how to communicate the humorous criticism devised.

5.1.1. Devising humorous aspects of the Target

The first phase of parody making is common to any kind of humorous behavior: before communicating humorous points, the Humorist must find them out. In political parody, the Parodist must find out some aspects of the Target that are not only worth being made fun of, but are so concerning some political criterion – according to our model, as far as the features of benevolence, competence, and dominance are concerned.

To illustrate this step with a real example, we may take the parody of Fabrizio Cicchitto by the comedian Max Paiella.

> The parliamentary member Fabrizio Cicchitto in 1980-90 was in the center-left Italian Socialist party and a devoted follower of its leader Bettino Craxi; since 1995 on he became a member of the Italian Parliament for the center-right party of Silvio Berlusconi, and one of his most devoted followers.
>
> In Paiella's parody, the background scene is the wide luxurious hall of the Italian Parliament. Cicchitto is represented by Paiella as a roman waiter, dressed with a long white pinny and talking, in a heavy popular roman accent, of Italian politics as if presenting the menu of his restaurant. This rendering Cicchitto as a waiter highlights his lack of dominance, namely his always being a devoted follower of some charismatic leader.

In this case, the process going from singling out a ridicule aspect of the Target subsumed to some

political criterion, to reproducing it to communicate this criticism, includes the following steps.

First the comedian must **choose a general political criterion** according to which a potential ridicule flaw can be found in the Target. Here the general criterion is Dominance: that particular politician does not show strong, autonomous and independent at all.

Second, to display a flaw according to that criterion, the parodist must **single out a specific flaw** in the Target, that may be subject to mockery. Cicchitto is seen as a *follower of leaders*, one submissive to important people.

Third, to embody the flaw of low dominance – to express it in a visible or audible way – the Parodist must devise a specific **"characterization"** of the Target. Cicchitto is characterized as a category of people who by definition must comply with another's commands: in a word, a servant. To find out the Target's characterization is, actually, the core of "deep" parody.

5.1.2. Communicating the humorous aspects of the Target

Once devised the Target's specific flaw, the parodist must communicate it, but must do so in a satiric way, that is, by highlighting its potential of ridicule, to make it an object of fun. In sum, conveying political criticism in parody requires that the following aspects are communicated, either directly or indirectly, in one or another modality:

a. the Target's identity
b. the event where the Target's flaw emerged
c. the specific flaw F (a negative property) attributed to T
d. (in some cases) the Target's characterization as a C, i.e., its attribution to category C, in which flaw F is embodied. In fact, attributing T the flaw F makes one characterize T as a C (as belonging to the negatively evaluated category C)
e. the humorous aspects of the Target's flaw F and/or characterization C

6. How the parody-crucial information is multimodally conveyed

To see how all the information is multimodally [14] conveyed by the visual and acoustic scene of the parody, we run a qualitative observational study.

6.1.Corpus

To find examples of how this information is conveyed in real parodies, we collected a corpus of 40 parodies of 30 Italian politicians and other public characters performed in Italian satire shows by 12 Italian comedians. We also included the parody of Hitler in Chaplin's *"The Great Dictator"* as an additional item.

6.2.Annotation scheme

To analyze the parodies of the corpus, we built an annotation scheme encompassing all the mechanisms that, according to our hypothesis, may be at work in making a parody. Table 1 shows the analysis of Max Paiella's parody of the Mayor of Rome Gianni Alemanno.

(http://www.youtube.com/watch?v=19ZACx63Vso). Here we report the scene and background knowledge necessary to understand the parody and its analysis.

> Gianni Alemanno, a former member of the Italian fascist party Alleanza Nazionale (National Alliance), was the Mayor of Rome from 2008 to 2013, often criticized for his having been a fascist drubber in his youth, and, when in the role of Mayor, for his familistic management of the Roman administration, having hired relatives and friends in the town bus company.

For his parody of Alemanno, Paiella picks up an episode in which the Mayor really made himself ridicule: the snow in Rome. Below we describe the context and background of this event, with words in bold describing the "allusion points", i.e., the objects and events the Parodist supposes to be known by the Audience, and to which he alludes in his parody.

> On February 3rd, 2012, snow came on Rome. Not used to snow, Rome is generally not prepared for this challenge, but in this case the disorganized management by Alemanno' staff turned a meteorological event into a disaster. A **newsletter** from the national Civil Protection warned that **35 millimeters** water were expected to come; actually, 1 mm. water corresponds to 1 cm. snow, but Alemanno and his staff did not know this, so they expected 3,5 centimeters of snow instead of the 35 that came in fact. No kind of prevention was undertaken: no **sal**t to prevent streets from freezing, no **snow chains** for buses; cars stopped, buses stopped for hours with romans inside. All that Alemanno did was to warn people to stay home, to buy **shovels** and distribute them to Romans recommending to clean up their doors (he also was videorecorded on TV while shoveling snow), and to say he would **call the army** to cope with the emergency. As the emergency was over, Alemanno was accused of disorganization and inefficiency, and to justify himself he appeared in all TV news and talk shows imputing the disaster to the Civil Protection, who had not warned how serious the situation was, and complaining that **he had been left alone** to confront the emergency.

In the annotation scheme of Table 1., Col.1 lists the modalities analyzed and Col. 2 the signals in the various modalities. Columns 3-5 refer to the topics on which the signal of col. 2 provides information: it may explicitly mention or allude to the Event relevant for the Target's judgment (col.3), contribute to the Target's identification (4), or characterization (5). In col. 6 we state if some stereotype is exploited in this characterization, and if so, which one; in col. 7, we write the specific flaw attributed to the Target through

the characterization of col. 6, and in col. 8 the political criterion – Benevolence, Competence or Dominance – with respect to which that is a flaw.

In Alemanno's parody, the relevant signals of the background scene are (line 1, col. 2) the *Coliseum* and the *flocks of snow*, the *shovel* and the *sheet of paper*, which provide information about details of the **Event** (col. 3): *Coliseum* and *flocks* allude to the snow in Rome, the *shovel* to Alemanno exhibiting while shoveling, the *sheet of paper* to the Civil Protection's misunderstood newsletter. While reminding the scene of the Event (*snow* and *Coliseum*) is potentially neutral, the other two allusions contribute to the characterization of the Target (col.5) and then point at his flaw (col. 7), finally classified in terms of a political criterion (col.8). Respectively, allusion to exhibiting while shoveling characterizes Alemanno as one who cares the image of doing things more than doing in fact, then a negative judgment of hypocrisy in terms of Benevolence; allusion to the misunderstood newsletter marks him as ignorant, a negative evaluation as to Competence.

On line 2, the *Roman centurion costume* (col.2) characterizes the Target (col.5) as the tourist operators at Coliseum dressed as centurions, generally connoted as underprivileged people from Roman slums, waiting for a tip after posing for a picture. This suit then conveys a social stereotype (col.6) of low socio-cultural level (col. 7): a negative evaluation in terms of Dominance or possibly Competence (col.8).

Line 3., with the morphological trait of a *square face* similar to Alemanno's, obtained by make-up (col.2), informs about Target **identity** (col.4).

Voice (line 4) here is relevant to highlight the **flaw**. Paiella imitates Alemanno's prosodic features of a voice sometimes tachilalic (very fast), even stuttering. This, besides contributing to Target identification (col.4), may also lead to infer anxiety or fear of the Audience judgment (col. 5), again implying too much care for one's image (col.7), a flaw as to Benevolence (col. 8.).

In the verbal modality (line 8.) Paiella literally repeats words really uttered by Alemanno in various news and talk shows. The literal quotations *"I have been left alone"* and *"I'll call the army"*, besides evoking the event of snow in Rome and the Mayor's behavior, characterize him as one who plays the victim (col.5) to reject accusation and criticism, highly caring his image (col.7): again a flaw in Benevolence (col. 8).

No relevant signal is found here as to gaze, facial expression or head movement (line 4.), gesture and body movement (5), conversational behavior (7) and name allusion (9), relevant in other parodies (see below).

6.3. Multimodal resources to communicate the bulk of parody

When devised the event, flaw, identification and characterization of the Target, how does the parodist distribute this information across modalities? We overview how corpus helps answer this question starting, in this Section from the information to convey (the columns of the annotation scheme), in Sect. 6.4 from the available communicative resources (the lines), showing the potentialities of each modality.

a. Target identification

When the Parodist impersonates the Target, to let the Audience understand who s/he is, it is sometimes sufficient for him to imitate the Target's suit and/or multimodal traits or behaviors. Max Paiella in his parody of the Mayor of Rome Alemanno exhibits a *square face*, similar to Alemanno's; in one of the subsequent Mayor Ignazio Marino, Paiella is dressed with a *sweater*, as Marino typically wears.

b. Event

Information concerning the event, is generally given by the scene background. That Cicchitto is a member of the Italian Parliament, and that he is talking of Italian politics, can be understood from the background scene, representing the *Parliament hall*. In the Parody of Alemanno, his being the Mayor of Rome is clear from the background of *Coliseum*.

c. Flaw

The flaw is less trivial to convey, being an abstract – not directly perceivable – property. How can the Parodist select just those physical features that characterize a concept, referent or property in such a way as to make it recognizable by the Audience? For example, how can one convey the concept of "servant"? This is where information d. may help.

d. Characterization

The Parodist characterizes the Target by assign him/her to a category that is stereotypically or prototypically distinguished by the flaw to convey. For example, what category of people is typically characterized by humbly complying with another's will? Waiters. In fact, Paiella characterizes Cicchitto as a popular roman waiter, presenting himself as a "humble servant", who manages the restaurant for politicians in the Parliament, and lists the present political events as items of a menu.

Both to invent a characterization that is a carrier of flaw c., and to find out a shared recognizable appearance of the devised category, the Parodist often resorts to the stereotypes or prototypes linked to that category.

Stereotypes

A stereotype is a schematic cognitive structure: a set of beliefs attached to some concept or category, that are socially shared in people of a given culture, and allow them to generate fast and easy inferences. [15; 16; 17]. The stereotypes that Jews are intelligent, that Italians eat spaghetti, or that Swedish are blonde allow people to generate expectations in case of interaction with people of those cultures. We have stereotypes concerning all categorization criteria – gender, age, social class, social role, culture, communicative behavior – and concerning the acoustic or visual appearance of people belonging to those categories [13]. So the Parodist may exploit stereotypes in finding both a category generally marked by a given (internal) flaw and the (stereotypical) multimodal features to represent that category. To convey the idea of "servant", Paiella characterizes Cicchitto as a waiter, and makes him recognizable by three stereotypical features of waiters in a popular roman restaurant: the textual one of *listing political topics in the form of a menu*, the visual one of the *long white pinny*, and the acoustic one of *roman accent*.

Prototypes

In some cases the characterization is so extremely stereotypical as to use a prototype. A prototype is a representative of some category that is characterized by its defining features to such an extent as to become an emblem, a primary exemplar of the whole category. A Parodist makes use of a prototype, not only of a stereotype, when the exemplar used for his characterization is a real person, embodying the defining features of that category so fully as to be a vivid and extreme example of them. Like in Maurizio Crozza's parody of Alan Friedman, an American journalist who conducts TV programs on Italian economy. In imitating him, Crozza adopts the peculiar accent of an American speaker, but to characterize it in an exaggerated, hence humorous way, he uses the unmistakable accent adopted by the famous Italian comic actor Alberto Sordi in his dubbing of Oliver Hardy: for Italians a prototype of the American man speaking Italian, an exemplar embodying all the most typical features of American accent.

e. Humor

According to most influential theories [15; 16], humor results from the violation of expectations consequent to the clash between two scripts, here represented by the distortion of the Parodist's imitation. In Cicchitto's Parody, the humorous effect is caused by the clash between the two scripts "Parliament" and "restaurant", the former evoked by the *Parliament hall*, the second by Paiella *dressed as a waiter*. The humorous intent is conveyed by the distortion – the exaggeration and stereotypicality – of the *pinny*.

6.4.The intertwining of modalities

From the analysis of our corpus it emerges that all modalities may be exploited to convey the various types of information relevant for a parody, but there is not a one-to-one relationship between types of modality and types of information. Let us take the modalities in the lines of our annotation scheme and see what types of information they provide in the parodies of the corpus.

A. Scene background

A first type of signal is the scene of the parody, generally informing about the event. In Paiella's parody of Alemanno, that the event concerned snow is alluded to by *flocks falling down* around Alemanno, while the location is revealed by the *Coliseum*.

At times, though, the location of the scene is used to identify the Target: the *Parliament hall* on the background of the man with a pinny listing his (political) menu helps recognizing that waiter as the Parliament member Fabrizio Cicchitto.

Music, a relevant acoustic aspect of the scene, sometimes helps Target identification, like when the *Italian national hymn* opens Crozza's Parody of the Italian President Giorgio Napolitano. Yet sometimes music makes part of the very criticism borne by the parody: like in the parody of Elena Boschi by Virginia Raffaele. Boschi is a young left-wing ministrer very close to Prime Minister Matteo Renzi, who deliberately made a government full of young politicians and women. In this parody, Raffaele / Boschi) is interviewed by a male journalist and, when asked politically embarrassing questions, her face performs a seductive behavior, while in the background goes the soundtrack of the movie *"A man and a woman"*, a music by itself evoking romantic and erotic contents: an allusion to Elena Boschi's seductiveness.

One more aspect of the scene is bystanders' behavior. In *"The Great Dictator"* this is a relevant pointer to political criticism in Chaplin's parody of Hitler. As soon as Hitler/Chaplin raises his hand in a sort of nazi salute, all *the crowd simultaneously starts clapping and booing*, while as he lowers it, the crowd *abruptly stops*: this alludes to the typical conformity and unanimity of totalitarian regimes, and therefore is part of the criticism.

B. Suit and general make up

Another signal helping to recognize the Target are, quite trivially, suit and general make up. This is the case with Hitler's *uniform* worn by Chaplin, but also with the *sunglasses* worn by Crozza's Flavio Briatore, which (stereotypically) characterize him as the tanned and vacuous millionaire he in fact is.

Sometimes, though, the suit, dress or costume is definitely part of the critical act. Cicchitto's *pinny* that is definitely a pointer to his being a servant, while Alemanno's *centurion costume* assimilates the Mayor of Rome to an unemployed of a Roman suburb in search for a tip from tourists.

On the other hand, in Crozza's parody of Roberto Formigoni, the right-wing Governor of Region Lombardia, famous for his bright color blazers, the extremely bright color blazer worn by Crozza alludes to Formigoni's crazy habit, or it makes fun of it; but it does not contribute to the political criticism.

C. Morphological traits

As predictable, visible morphological traits of the Target are often imitated, through make-up or fakes, to make the Target recognizable: *teeth* and *hair* in Crozza's parody of Matteo Renzi, the *square face* of Paiella's Alemanno, and the *moustache* of Chaplin's Hitler.

Only the exaggeration of morphological traits is sometimes used as humor point, like in Crozza's parody of Renato Brunetta, a right-wing Minister who is physically very characterized by his being very short. Crozza represents him by standing on his knees.

D. Facial expression

Communication through head, gaze and facial expression is not used very often to convey crucial information in parody. As predictable, different from morphological traits, that are by definition stable, it is never exploited for Target recognition. Yet, it is used as a "flaw pointer" in the parody of Elena Boschi, where her particular facial expression is crucial to convey the idea of a seductive she/politician.

E. Gesture, posture and body movement

Gestures, postures and body movements are frequently used, as predictable, to identify the Target, like in the parody of Matteo Renzi as a Prime Minister, where Crozza imitates his *loose and casual walk*, his talking with *hands in his pockets*, and his typical gestures of impatience. But both gesture and posture are also often exploited to point at the Target's ridicule flaws. Crozza counts various parodies of Umberto Bossi, the founder and first charismatic leader of the North League – a party struggling for the secession of the North from the Center and South of Italy. In a recent parody, that follows the fall of Bossi and the ascent of his former lieutenant Roberto Maroni, Crozza performs both roles, of Bossi and Maroni: the former as a very active, dynamic, still enthusiastic and provocative person, the latter as a rigid clerk, a white-collar, a burocrat; and to render this imagine of Maroni, Crozza represents him as a person always *dressed in brown*, with a *rigid posture*, his *head recessed in his shoulders*, always *still* and *looking forward*, much like a robot or a puppet. All this points to Maroni's lack of the vision, creativity and charisma necessary to a leader, as opposed to Bossi. Thus, Bossi's postures and movements evoke cheerfulness, while Maroni's are intended to raise criticism, derision, and laughter.

In general, what is imitated in parody – and actually what most characterizes the identity of a Target – is not so much a specific gesture or posture, but rather the "expressivity parameters" of the characters' movements [18], their amplitude, fluidity, velocity, repetition: for instance while gestures and movements of Crozza / Renzi are frequent, of high fluidity and medium amplitude, those of Crozza / Maroni are few, of minimum fluidity and low amplitude.

Only in rare cases are some specific gestures the marker of a Target: like for Brunetta, whose typical gesture of *raising both hands with extended index fingers* is repeatedly used in Crozza's imitation. Actually, Brunetta is generally very aggressive and arrogant, and the didactic and haughty attitude he generally adopts is often conveyed by that very gesture.

F. Voice

A classical and important part of imitation is voice. All parameters of voice are exploited by Parodists in our corpus in order to Target identification: voice quality, regional accent, typical intonation and prosody, including the Target's idiosyncratic temporal structure (fast or slow voice). Crozza imitates Renzi's *Florentine accent*, Paiella sometimes *speaks* as *fast* as Alemanno, Chaplin reproduces Hitler's *jerky rhythm of voice*.

In many cases, though, that particular vocal parameter is not only used to identify the Target, but also – or only – to convey the ridicule flaw and to solicit laughter: it is not easy to tell which of two functions is mainly aimed at by the Parodist. For example, Chaplin's jerky and loud German vocal onset not only identifies Hitler, but also the stereotype of the threatening German, and at the same time makes fun of him when the word onset turns into a cough.

G. Conversational behavior

Parodists in our corpus do not only imitate Targets' specific words, but also their "conversational behavior". Since Brunetta, when being interviewed, takes the role of the interviewer, chasing the actual interviewer, asking provocative questions and repeating them obsessively, Crozza imitates his aggressive and insistent sequence of behaviors.

H. Words, sentences, discourses

Verbal text is a relevant part of the Parodist's work, in which the Parodist performs both a surface and a deep imitation of the Target. A surface – yet, quite effective – imitation when s/he utters the very same words or sentences that have actually been used by the Target, becoming a "torment", a verbal emblem or griffe of that character. A such case is in Crozza's parody of Fausto Razzi, a senator who betrayed his left-wing party and passed to a right-wing party, thus avoiding the Government resignation, because if the legislature had been closed he could not have got his retirement fund. In an off-air personal dialogue with a colleague he justified his vote by saying: *"Fatti un*

poco li cazzi tua. … Dammi retta, te lo dico da amico" (Act in your own interest… Listen, I tell you this as a friend). These sentences became a legend, so much so that even a T-shirt was invented with this motto. Crozza parodies the off-air dialogue using Razzi's very words.

Another case of literal quotation of the Target's words is Crozza's parody of the millionaire Flavio Briatore, who often uses the adjectival idiom *da sogno* (dream-like). Crozza often uses this expression but does so in quite improbable combinations, like "dream-like frozen green peas", to elicit laughter.

Often, though, the Parodist's words do not literally draw on the Target's, but rather express the concepts – or a parody of them – generally conveyed by the Target: a case of "deep" parody.

For example, to make fun of Briatore's strange priorities, Crozza says: *"Io penso che l'altruismo sia molto importante. Per me l'altruismo è al diciottesimo posto"* (I think altruism is very important. To me altruism is in the 18° place). This points to the essence of Briatore as a person strongly oriented to business and money.

In sum, the imitation of words or sentences, both in their literal phrasing and in their simply expressing a concept typical in the Target's (communicative) behavior, is never used only in order to his identification, but to point at his ridicule flaws.

I. Names and puns

So far we have only found one case of distortion of the signal in our corpus: exaggeration. Crozza represents Brunetta as much shorter than he really is; Hitler's jerky accent in Chaplin's representation is exaggerated, as is, up to the paradox, the list of ethical priorities for Briatore. But an intriguing exploitation of words in parodies is word mangling, and more specifically, name mangling. Name mangling is a way to make a name just a little bit different and thus evoke a different meaning, but keeping track of its original meaning: what is generally done in puns.

Name mangling is a typical strategy of discredit is, considered, even by Freud, an insulting behavior, because it shows contempt toward the named person – not even worth to have his name reminded. – and possibly suggests some negative nuance of him/her. This strategy is used by Chaplin when, as Hitler, he mentions the German officials around him, *Harring* and *Gabitsch*, probably alluding to Goering and Goebbels: two names evoking the nouns "*herring*" and "*garbage*", and thus shedding a light of insult over them – a filthy animal and filthy stuff.

The same strategy is used by Crozza in his parody of Massimiliano Fucsas, a famous Italian architect and designer, who, to stigmatize his intellectualism, vacuity and the odd things he says, is called "Massimiliano *Fuffass*", a name connected to the jargon term *fuffa*, "vacuous, vague, imprecise stuff".

1 Conclusion

To describe the cognitive processes implied in producing parodies of politicians, we made a hypothesis about the types of information a Parodist necessarily conveys.

The Parodist does not only exhibit some ridicule features of the Target, but also make it recognizable resorting to all possible devices of imitation. Like for any imitation the Parodist must select which features to represent of the Target, but if those features are not per se ridicule, he sometimes characterizes the Target as belonging to an unexpected – hence laughter inducing – category, by finding out a "deep" aspect of it.

By analyzing a corpus of 41 parodies we described how all modalities in the Parodist's behavior and background scene intertwine in the crucial steps: identifying the Target, alluding to specific events, and highlighting the Target's flaws through characterization within an unexpected category.

Acknowledgements. Research partly supported by SSPNet Seventh Framework Program, European Network of Excellence SSPNet (Social Signal Processing Network), Grant Agreement N.231287.

References

1. Holman C. H. and Harmon W. 1986. *The handbook to literature.* 5th ed. New York, Macmillan.
2. Rose M. 1979. Parody//Meta-Fiction: an analysis of parody as a critical mirror to the writing and reception of fiction. London, Croom Helm.
3. Rose M. 2011. Pictorial Irony, Parody, and Pastiche: Comic Interpictoriality in the Arts of the19th and 20th Centuries. Bielefeld, Aisthesis Verlag.
4. Condren C., Milner Davis J., Phiddian R. and McCausland S. Defining parody and satire: Australian copyright law and its new exception, Part II – Advancing ordinary definitions. Media Arts Law Review, Vol. 13, No. 4, Dec 2008, 401-421.
5. Milner Davis J. Book review of "Margaret Rose: Pictorial Irony, Parody, and Pastiche: Comic Interpictoriality in the Arts of the19th and 20th Centuries. British Journal of Aesthetics Vol. 53 | Number 3 | July 2013 | pp. 365–376.
6. Kreuz R.J. Roberts R. 1993. On satire and parody: The importance of being ironic. *Metaphor and Symbolic Activity* 8(2): 97-109.
7. Bakhtin M. M. 1981. From the prehistory of novelistic discourse. In: Michael Holquist, ed., *The dialogic imagination*, 41-83. Trans. Caryl Emerson and Michael Holquist. Austin, TX: University of Texas.

8. Rossen-Knill D.F., Henry R. 1997. The pragmatics of verbal parody. *Journal of Pragmatics* 27(6): 719-752.
9. Rotermund E. 1964. Die Parodie in der modernen deutschen Lyrik. Berlin, Eidos Verlag.
10. Luttazzi D. 2001. *Satyricon*. Milano: Mondadori.
11. Hulstijn J., Nijholt A. 1996. (eds.). *Proceedings of the International Workshop on Computational Humour* (TWLT 12), University of Twente, Enschede, Netherlands.
12. Poggi I. D'Errico F. 2013. Towards the Parody Machine. Qualitative Analysis and Cognitive Processes in the Parody of a Politician, New Trends in Image Analysis and Processing – ICIAP 2013, Lecture Notes in Computer Science, Petrosino, Alfredo, Maddalena, Lucia, Pala, Pietro, Springer Berlin Heidelberg, 491-500.
13. Poggi I., D'Errico F., L.Vincze,. 2011. Discrediting moves in political debate. In F.Ricci et al. (eds) Proceedings of Second International Workshop on User Models for Motivational Systems: the affective and the rational routes to persuasion (UMMS 2011) (Girona) Springer LNCS.pp. 84-99, 2011.
14. Poggi I. 2007. Mind, hands, face and body. A goal and belief view of multimodal communication. Berlin: Weidler.
15. Gordon W. Allport. 1954. The Nature of Prejudice. Addison-Wesley, Cambridge, MA.
16. Susan T.Fiske. 1998. Stereotyping, Prejudice, and Discrimination. In The Handbook of Social Psychology, Daniel T.Gilbert, Susan T. Fiske, and Gardner Lindzey. Volume Two (4th ed.). McGraw-Hill, Boston, Mass.
17. Dirk Geeraerts. 2008. Prototypes, stereotypes, and semantic norms. In Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems, Gitte Kristiansen and René Dirven (Eds.), Mouton – De Gruyter, Berlin, 21-44.
18. Hartmann, B., Mancini, M., & Pelachaud, C. (2002). Formational Parameters and Adaptive Prototype Instantiation for MPEG-4 Compliant Gesture Synthesis. *Computer Animation 2002*, 111-119.
19. Ruch, W. (ed.) 1998. The Sense of Humor: Explorations of a Personality Characteristic. Mouton-de Gruyter, The Hague-Berlin.
20. Attardo S. 1994. Linguistic theories of humor. Walter de Gruyter, Berlin.

Table 1.The annotation scheme of parody

	1. Target and Modalities	2. Signal	3. Event	4. Target Identity	5. Characterization	6. Stereotype	7. Flaw	8. B/C/D
A	Alemanno							
1	Background	Coliseum Snow flocks Shovel Paper	Snow in Rome Showing while shoveling Misunderstood newsletter				 Image care Ignorance	 B C
2	Suit & makeup	Centurion costume			Tourist operator going for tips	Social	Low cultural level	D
3	Morph. traits	Square face		X				
4	Head face gaze							
5	Gesture & body movement							
6	Voice	Tachilalic		X	Performance anxiety		Image care	B
7	Conversational behavior							
8	Text	*I have been left alone* *I'll call the army*			Self-victimization		Image care	B D
9	Name allusion							

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

Abstract

This paper investigates cross-linguistic speech-gesture differences in a 'prototypical' satellite framed language (Danish) and a language with split system possibilities for lexicalization (Italian). It is often claimed that languages have specific inventories for lexicalizing motion which are reflected in gestural repertoires. But how are gestures expressed in a language that easily encodes PATH of motion either in verb roots or in satellites? Results from Danish and Italian show cross-linguistic differences in speech and gesture patterns and suggest that gesture production is more a result of the speakers' on-line utterance conceptualization processes rather than language-specific cognitive diversity.

1 Introduction

People often gesture when they speak. These speech-accompanying gestures are closely tied semantically and temporally with speech and language (Kendon 1980; McNeill 1992, 2005). Because of this tight link, speech and co-speech gestures are increasingly seen as planned and processed together at a conceptual level (McNeill 2005) although the nature of the link is still debated (de Ruiter 2007). One way to investigate the speech-gesture relation involves looking at cross-linguistic differences in semantic fields. Languages vary substantially in how meaning is expressed, especially when speaking about motion (Berman and Slobin 1994; Levinson and Wilkins 2006).

Motion is a frequent and everyday topic in human discourse and all languages have means for describing it. A motion event is basically a FIGURE in MOTION along a PATH in respect to a GROUND (for translational motion). Typologically different languages vary considerably in lexicalization of how the semantic features e.g. MANNER of motion (the way the FIGURE moves) and PATH of motion (the direction of the FIGURE) are mapped onto linguistic form (Slobin 2004; Talmy 1985, 1991). These lexicalization patterns are also claimed to influence the syntactic packaging at the level of a clause (Slobin 1996). Based on how and where PATH of motion is framed, languages are generally classified into at least two major categories (Talmy 1985): verb framed languages (e.g. Romance, Turkish, Japanese and Semitic) encoding directionality (PATH) in the verb root, and satellite framed languages (e.g. Indo-European languages except Romance) encoding directionality in satellites to the verb e.g. prefixes, verb particles, adverbs. This categorization leads to the assumption that speakers of different languages have different thinking-for-speaking patterns, also known as linguistic conceptualization (Cadierno and Ruiz 2006; Slobin 1987, 1991, 1996). The term *thinking-for-speaking* refers to the possible effects of language on the kind of thinking that occurs online while speaking a particular language. Cross-linguistic studies reveal striking differences in how speakers of specific languages

allocate attentional and other resources to features of events that the language they speak either foregrounds or provides readily accessible expressions for (Slobin 2004; Cadierno 2012).

This means that speakers of different languages do not attend to the semantic concepts of MANNER and PATH equally since their language does not weight these factors of motion in an equally salient way. Thinking-for-speaking "*involves picking those characteristics that (a) fit some conceptualization of the event, and (b) are readily encodable in the language*" (Slobin 1987, 435). Speakers therefore construe situations in terms of those dimensions that are privileged in their own language which leads to different patterns of lexicalization.

1.1 Cross-linguistic differences in gestures

Insofar as languages differ in how meaning is organized syntactically, co-speech gestures will often reflect these cross-linguistic differences (Gullberg 2011; Kita et al. 2007; McNeill and Duncan 2000; Stam 2006). The influence of lexical packaging of information on gestural output is demonstrated in a number of studies investigating gesture production in typologically different languages (Kita and Özyürek 2003; Brown 2007; McNeill and Duncan 2000). Speakers of satellite framed languages are claimed to focus more on the MANNER component (Slobin 2006) though targeting both MANNER and PATH in speech. They typically express both elements in a single spoken clause and produce a single gesture containing MANNER, PATH or MANNER *and* PATH together in one gesture reflecting the tightness of clause structure (Brown and Gullberg 2008; Negueruela et al. 2004; Kellerman and van Hoof 2003; Kita and Özyürek 2003). Speakers of verb framed languages, on the other hand, more often focus on, and target, PATH of motion in speech and to a minor extend MANNER. MANNER is an optional element and often omitted in speech, possibly due to smaller manner lexicons (Slobin 2003) and/or complexity in subordination of manner. Therefore speakers often distribute MANNER and PATH in two spoken clauses accompanied by two gestures i.e. one gesture per clause. Path gestures tend to align with path verbs and Manner or Manner-Path conflated gestures with MANNER subordinated clauses (Hickmann, Hendriks, and Gullberg 2011; Stam 2006; Kita and Özyürek 2003).

But gesture patterns seem not to be language-specific in the sense that a specific language holds a certain 'mode' of gesturing. Kita et al. (2007) investigated the effect of syntactic frames on gestural representation of MANNER and PATH in English, a satellite framed language, by manipulating elicitation material to elicit both one and two-clause constructions. They found that English speakers were more likely to alter gestural distribution relative to the syntactic construction of the motion description. When describing motion in two separate clauses, speakers were more likely to produce two separate gestures. They thus concluded that the speaker's choice of syntactic framing influences the packaging of information in gestures. Therefore gestural variation reflects the speakers' on-line utterance conceptualization process rather than a habitual cognitive diversity.

1.2 Danish and Italian

Very little attention has been paid to Danish and Italian regarding speech-gesture patterns. There are several reasons for why the languages in question are of interest. Danish and Italian belong to two different typological patterns (Cadierno and Ruiz 2006). 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 verb roots (e.g. *salire*, *scendere*, ascend, descend), with verb particle constructions (e.g. *andare su*, *salire su*, go up, ascend up) (Folli 2008; Iacobini and Masini 2006) or with manner verbs and directional adverbs (*rotolare su*, roll up). This variety of possibilities in a verb framed language show properties of a 'split system' typology (Talmy 2000, 64).

Italian is particularly interesting as verb particle constructions seem to be more frequently used in than previously thought (Slobin 2004) and as Italian is believed to possess a manner verb inventory that is more comparable in size to English (Iacobini 2010), but see Cardini (2008) for alternative perspectives.

In respect to gesture studies, Rossini (2005) investigated how different levels of lexicalization could affect gestural distribution and found that Italians do express PATH in satellites to the verb far more often than chance (58%) and synchronize gestures with either the lexical item or verb + satellite, but she fails to mention the semantic content of the co-expressive gesture on a quantitative level. Rossini hypothesizes that the distribution of gestures is the result of the tightness of lexicalization patterns i.e. whether the verb and the satellite are bound tightly or clearly separated by other grammatical constituents or prosodic features. Other studies also probe into speech-gesture differences in Italian (and English) finding significant differences in gesture rate and gesture space between the languages when narrating motion events, but do not explain whether these differences are due to habitual differences alone or to linguistic packaging of lexical items as well (Cavicchio and Kita 2013, 2013).

The question remains whether Danish and Italian have different gestural repertoire based on a preferred linguistic pattern of that particular language or if their co-speech gestures are a result of the on-line utterance choice and syntactic structuring of semantic elements.

1.3 Present study

This study investigates how different strategies for expressing MANNER and PATH in Danish and Italian influence the content of co-speech gestures. Since speech and gestures are increasingly seen as integrated in production we expect to see inter-typological differences (between Danish and Italian) and intra-typological in-language variations (for Italian). We ask whether Danish and Italian speakers generally have different gestural patterns based on their preferred typological pattern or whether gestures reflect the speaker's online strategy for lexicalization of motion events.

2 Method

2.1 Participants

Ten Danish (7 female) native speakers (M_{age} 38.3; SD 14.1, range 24-69) and ten Italian (6 female) native speakers (M_{age} 26.3; SD 6.38, range 19-42) participated in the study. They were all university students or postgrads from the University of Copenhagen/Copenhagen Business School and University of Rome (Roma Tre) respectively. All participants were individually shown the elicitation material on a laptop and narrated the events to a confederate listener. All participants were video recorded for further analysis. All data is analysed in Anvil 5.17 (Kipp 2004).

2.2 Material

The elicitation material in this experiment consisted of two sets of four motion events (eight in total). The first set; *the Tomato man movies* (Özyürek, Kita, and Allen 2001) [1] contained translational up and downwards motion (either rolling or jumping) as in figure 1.

The second set; *Boundary ball* (Wessel-Tolvig 2013) contained translational motion in or out (of a house) either rolling or jumping as in figure 2.

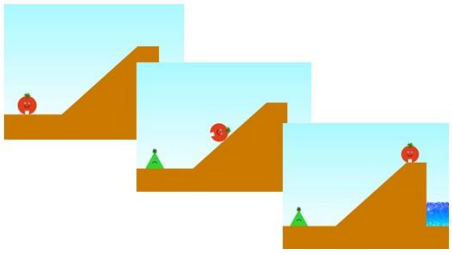

Figure **1**: Tomato man movies

[1] Used with permission from Sotaro Kita (University of Warwick).

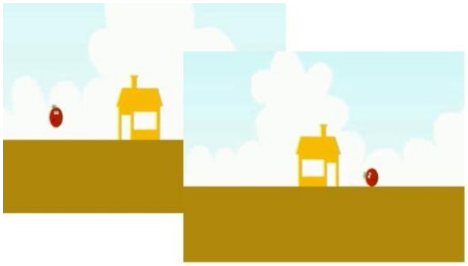

Figure **2**: Boundary ball

2.3 Coding

All participants saw and narrated all 8 motion events. For each narration the target event (the figure moving up and down the hill or in and out of the house) was selected for the analysis. Speech utterances were divided into clauses defined as "*any unit that contains a unified predicate*" (Berman and Slobin 1994, 660) as in (1) and (2). Gestures were coded using McNeill's coding scheme (1992, 377-378).

(1) The ball **[bounced down]** the hill

(2) The ball **[descended]** the hill | as it **[bounced]**

Thus speech clauses were coded for containing either PATH only (he ascends), MANNER only (he rolls), MANNER verb + PATH particle (he rolls up) or PATH verb + subordinated MANNER (he ascends rolling/as he rolls). Gestures were coded as to how they expressed MANNER information (the way the figure moved, gesture mainly depicting a rolling or jumping motion with no directionality), PATH information (the direction of the movement with no indication of how the figure moved), or MANNER and PATH conflated into one single gesture e.g. both MANNER and PATH information simultaneously (the manner of how the figure moved and the direction in respect to the background). The attributes for speech and gesture coding are shown in table 1. Gestures were linked to the concurrent speech clause to clarify how speech and gestures aligned in motion descriptions across the two languages.

Sometimes people do not gesture when speaking, but for data analysis, all motion events were collected whether speakers gestured or not.

Attribute	**Value**
Speech	PATH only MANNER only PATH verb + $\text{SubM}_{\text{ANNER}}$ MANNER verb + satellite
Gesture	Manner only Path only Manner and Path conflated

Table **1**: Annotation features

3 Results

Results show both inter-typological differences in speech patterns and gesture distribution and intra-typological or in-language differences in how gestures are distributed and aligned with speech elements in Italian.

3.1 Speech results

Speech was first transcribed and tokenized, each word constituting a token. The vocal elision of some Italian articles + nouns contracted in written form e.g. *all'interno* (within) was counted as two tokens. For the Danish participants 89 motion events were recorded, and 77 for the Italian speakers. The Danish speakers produced 638 words (only counted within the target motion) whereas the Italians produced 448 words. The Danish participants in the experiment produced more words per motion event than their Italian counterparts as table 2 shows. This can be due to the fact that many Italian motion descriptions only included PATH descriptions (lacking MANNER component), due to implicit subject in Italian, and/or the fact that Danish use directional adverbs + prepositions for PATH + GROUND descriptions.

	Motion events (ME)	Tokens	Tokens/ME Mean	SD
Danish	89	638	7.17	2.8
Italian	77	448	5.82	2.19

Table **2**: Speech results

Congruent with the typology proposed by Talmy (1985), Danish speakers expressed PATH in verb particles (op, ned, ind, ud –*up, down, in, out*) and encoded MANNER in the verb root in one clause constructions (e.g. ruller ned – *rolls down*) as can be seen in figure 3. Italians on the other hand described motion events using a variety of different syntactic constructions including PATH

in verb roots with or without MANNER subordination (entra nella casa | rotolando - *enters the house* | *rolling*) and MANNER verbs followed by PATH particles (rotola giù per la collina - *he rolls down the hill)*. Speech patterns are shown in figure 4 shows.

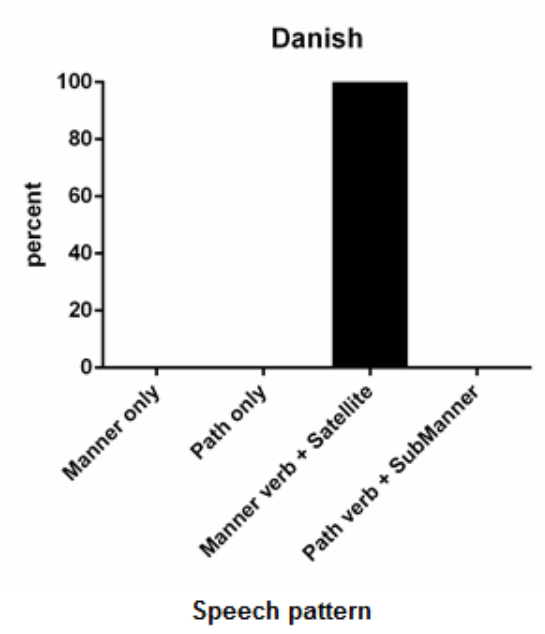

Figure **3**: Speech patterns in Danish

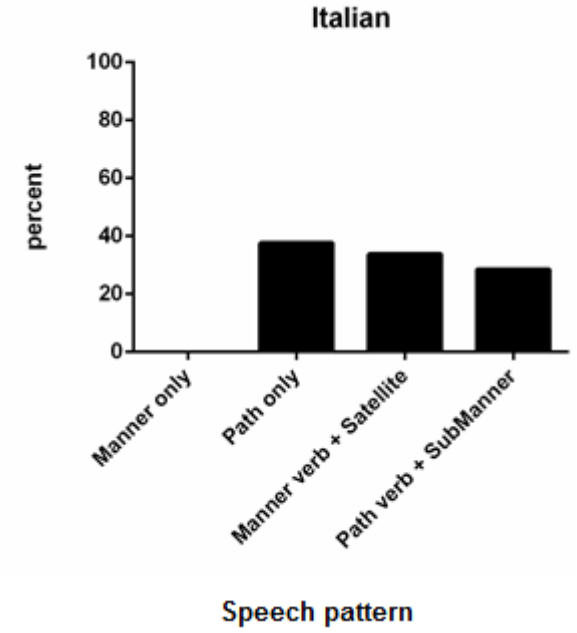

Figure **4**: Speech patterns in Italian

The Italian speakers show a preference for expressing PATH of motion in verb roots (66.2% of all occurrences), but to a high degree also PATH in verb particles (33.8%). The Italian participants often left out MANNER components and expressed only PATH in speech (37.7%). This is congruent with previous research results claiming, that speakers of verb framed languages often leave out MANNER because of syntactic complexity, subordination (Gullberg 2011; McNeill and Duncan 2000), and the fact that motion events can be completed without expressing MANNER e.g. "he enters the house [FULL STOP] | *Jumping*".

Our results for the Italian participants are similar to the findings from Rossini (2005) who also found MANNER verb + PATH particle constructions, but is not similar to the findings by Stam (2006) where none of the Spanish speakers (also verb framed) conflated MANNER and PATH in speech.

3.2 Gesture types

The distribution of gestures shows both similarities and differences between the groups. Extracting the motion events *with* gestures we exclude motion events where no gestures occurred. Many factors govern individual differences in the production of gestures e.g. extraversion/introversion, confederate or naïve listener, shyness, surroundings, repletion/practise etc.

The Danish speakers produced 62 motion events with gestures, which corresponds to 7.05 words per motion event ratio and a 1.05 gesture per motion event ratio, while the Italian speakers produced 68 motion events with gestures corresponding to 5.72 words per motion event ratio and a 1.13 gesture per motion event ratio as can be seen in table 3.

	Motion events + gesture	Gestures	Ratio W/ME	Ratio G/ME
Danish	62	65	7.05	1.05
Italian	68	77	5.72	1.13

Table **3**: Gesture results

The results actually show a very similar distribution of gesture types in the two languages. As seen in figure 5 and figure 6 the Danish and Italian participants produce roughly the same

amount of gestures expressing Manner only, Path only and Manner and Path conflated gestures. The most striking finding is perhaps the high number of gestures conflating MANNER and PATH in Italian, which contradicts other results obtained for other Romance languages.

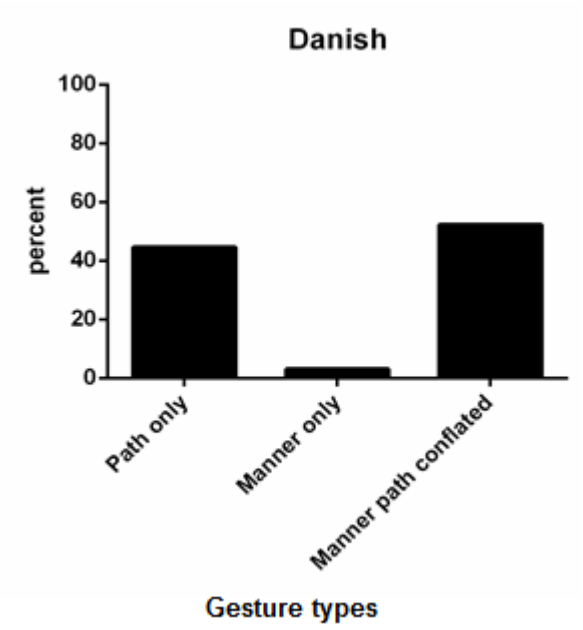

Figure **5**: Gesture types used by Danish speakers

.

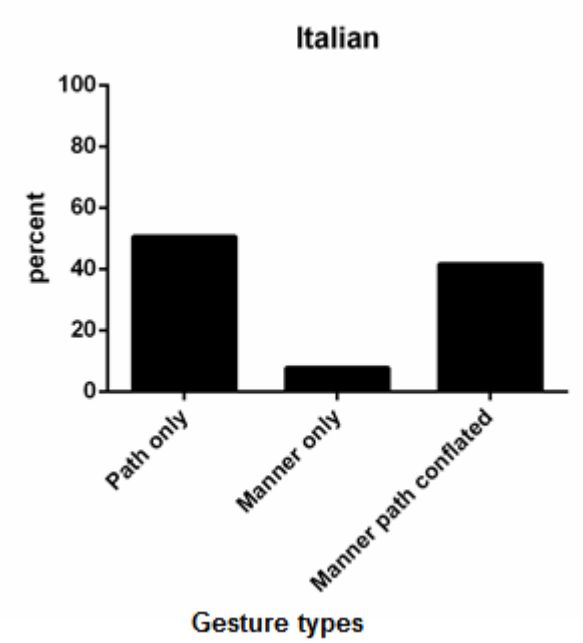

Figure **6**: Gesture types used by Italian speakers

For instance Hickmann, Hendriks, and Gullberg (2011) found none of their French adult speakers conflated MANNER and PATH in gestures. McNeill and Duncan (2000), on the contrary, found Spanish speakers to conflate MANNER and PATH in gestures when expressing motion events with no mention of spoken MANNER. The Manner gesture was therefore not directly tied to a linguistic component, but rather a sign of complementing the verb utterance as it lacked MANNER.

3.3 Gesture distribution

The distribution of gestures (figure 7 and 8) compared with lexical choice shows how Danish speakers in this experiment naturally align all gestures with a MANNER verb + satellite construction (100%) as all gestures in our material co-occur with this type of construction.

More interestingly for Italian the distribution of gestures shows that Path only gestures often align with PATH only speech utterances (MANNER omitted) (e.g. sale per la collina - *he ascends the hill*). This is in line with previous results stating that Path gestures often align with PATH expressions in e.g. Spanish (McNeill and Duncan 2000; Negueruela et al. 2004; Stam 2006).

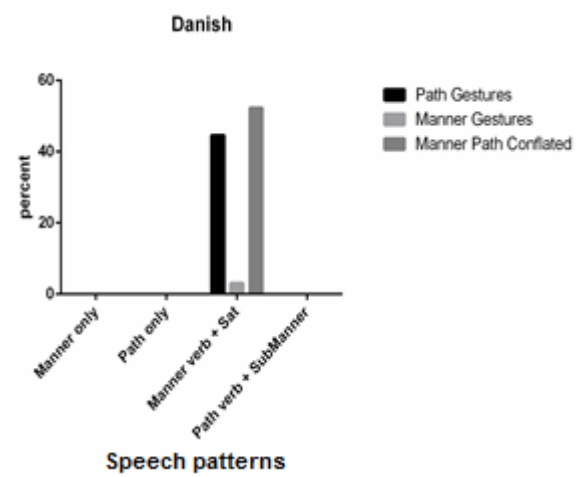

Figure **7**: Distribution of gestures on speech constructions in Danish

Results also reveal that Italian path verb + $SubM_{ANNER}$ constructions (entra nella casa | rotolando - *enters the house* | *rolling*) yield a large distribution of both Path only and Manner

and Path conflated gestures. When the Italians produce MANNER verb + PATH particle (as in satellite framed constructions) they produce Manner and Path conflated gestures.

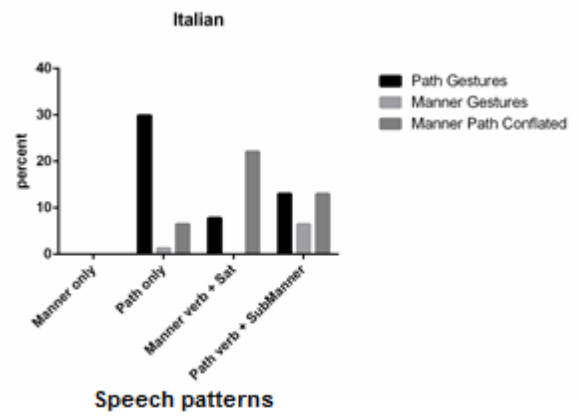

Figure **8**: Distribution of gestures on speech constructions in Italian

4 Discussion

Although the MANNER verb + PATH particle construction is not as frequent in Italian (33.77%) as in Danish (100%) the verb particle construction may be a more recurring element in modern Italian (Iacobini and Masini 2006; Korzen 2012) than previously thought (Slobin 2004). The "deviation" from the more standard verb framed forms is not as pervasive as found in Rossini (2005) who reported 58% of motion occurrences in a study of Italian to be satellite framed. This discrepancy could be explained by several factors like linguistic regional variation of north/south Italy, individual variation (and small datasets) and the fact that half of this study's elicitation material involved the animated character crossing a physical boundary. As indicated by Aske (1989) and Slobin and Hoiting (1994) crossing boundaries using PATH particles in Romance languages is not a possibility (boundary crossing constraint). MANNER of motion in verb framed languages can only be mapped onto the verb root in non-boundary crossing situations (Cadierno and Lund 2004) although there are indications in this material that Italians can *and do* express boundary crossing situations with MANNER verbs and PATH particle constructions. This variation in lexicalization makes Italian an interesting field of study regarding gesture representation of events. The data show how the frequent use of satellite framed constructions in Italian has an influence on the gestural representation. Lexicalizing motion with verb particle constructions promoted the use of Manner-Path gestures more frequently than when MANNER was omitted. The results seem to support the idea that gesture production is influenced by the choice of syntactic packaging of semantic elements, but not to the extent that we can speak about a certain habitual (Italian) way of gesturing. This is in compliance with the idea that gesture production is influenced not only by the preferred speech patterns of the particular language you speak, but more precisely by the on-line utterance planning and syntactic construction you choose for describing a motion event.

5 Conclusion

Although linguistic conceptualization is a well-studied area and also growing within gesture studies, only a handful of languages are recently investigated. The need for baseline data from other languages is vital, and especially data from languages not following standard verb or satellite framed forms (like Italian) can provide detailed information to a range of theoretical issues in language and gesture production and in cognitive studies. Future analyses will focus on how Danish learners of Italian learn to cope with the dual Italian strategy for expressing motion. Preliminary results indicate a re-organization of semantic representation and a shift in attention towards a uniform verb-framed system which is opposite of the Danish L1 system, but which does not really correspond to the reality of spoken Italian.

Acknowledgments

I thank Costanza Navarretta and Magdalena Lis for useful feedback, Isabella Poggi and Laura Vincze for helping with data collection in Rome and Patrizia Paggio for invaluable guidance. All mistakes and errors are mine and probably due to not following good advice.

References

Aske, Jon. 1989. Path predicates in English and Spanish: A closer look, at Proceedings of the Berkeley Linguistic Society 15.

Berman, Ruth Aronson, and Dan Isaac Slobin. 1994. *Relating events in narrative: A crosslinguistic developmental study*. Edited by L. E. Associates. Hillsdale, NJ: Psychology Press.

Brown, A. 2007. *Crosslinguistic influence in first and second languages: Convergence in speech and gesture*. Vol. PhD, *MPI Series in Psycholinguistics*. Nijmegen: Max Planck Institute for Psycholinguistics.

Brown, A, and Marianne Gullberg. 2008. Bidirectional crosslinguistic influence in l1-l2 encoding of manner in speech and gesture: A Study of Japanese Speakers of English. *Studies in Second Language Acquisition* 30 (02):225-251.

Cadierno, Teresa. 2012. Thinking for speaking in second language acquisition. In *The Encyclopedia of Applied Linguistics*, edited by C. A. Chapelle. Oxford: Wiley-Blackwell.

Cadierno, Teresa, and K. Lund. 2004. Cognitive linguistics and second language acquisition: Motion events in a typological framework. In *Form – meaning connections in second language acquisition*, edited by B. VanPatten, J. Williams, S. Rott and M. Overstreet. Hillsdale, N.J.: Lawrence Erlbaum.

Cadierno, Teresa, and L. Ruiz. 2006. Motion events in Spanish l2 acquisition. *Annual Review of Cognitive Linguistics* 4:183-216.

Cardini, Filippo-Enrico. 2008. Manner of motion saliency: An inquiry into Italian. *Cognitive Linguistics* 19 (4):533-569.

Cavicchio, Federica, and Sotaro Kita. 2013. Bilinguals Switch Gesture Production Parameters when they Switch Languages, at TiGeR in Tilburg, Holland.

Cavicchio, Federica, and Sotaro Kita. 2013. English/Italian Bilinguals Switch Gesture Parameters when they Switch Languages, at CogSci in Berlin, Germany.

de Ruiter, Jan Peter. 2007. Postcards from the mind: the relationship between speech, imagistic gesture, and thought. *Gesture* 7 (1):21-38.

Folli, Raffaella. 2008. Complex PPs in Italian. In *Syntax and Semantics of Spatial P.*, edited by A. Asbury, J. Dotlacil, B. Gehrke and R. Nouwen: John Benjamins Publishing Company.

Gullberg, Marianne. 2011. Thinking, speaking and gesturing about motion in more than one language. In *Thinking and speaking in two languages*, edited by A. Pavlenko. Bristol: Multilingual Matters.

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

Iacobini, Claudio. 2010. The number and use of manner verbs as a cue for typological change in the strategies of motion events encoding. In *Space in Language: Proceedings of the Pisa International Conference*, edited by G. Marotta, A. Lenci, L. Meini and F. Rovai. Pisa: Edizioni ETS.

Iacobini, Claudio, and Francesca Masini. 2006. The emergence of verb-particle constructions in Italian: locative and actional meanings. *Morphology* 16:155-188.

Kellerman, E., and A. M. van Hoof. 2003. Manual accents. *International Review of Applied Linguistics*, 41 (3):251-269.

Kendon, Adam. 1980. Gesture and speech: two aspects of the process of utterance. In *Nonverbal Communication and Language*, edited by M. R. Key. The Hague: Mouton.

Kipp, Michael. 2004. *Gesture Generation by Imitation - From Human Behavior to Computer Character Animation*. Florida: Boca Raton.

Kita, Sotaro, and Asli Özyürek. 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, Sotaro, Asli Özyürek, S Allen, A Brown, R Furman, and T Ishizuka. 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.

Korzen, Iørn. 2012. Endo- og exocentrisk verbaltypologi : En genlæsning af Talmy - nu med (god) grund. *Ny forskning i grammatik* 19:129-152.

Levinson, Stephen C., and D. P. Wilkins. 2006. *Grammars of space: Explorations in cognitive diversity*. Cambridge: Cambridge University Press.

McNeill, David. 1992. *Hand and mind: what gestures reveal about thought*. Chicago: University of Chicago Press.

McNeill, David. 2005. *Gesture and Thought.* Chicago: University of Chicago Press.

McNeill, David, and Susan Duncan. 2000. *Growth points in thinking-for-speaking*. Edited by D. McNeill, *Language and Gesture*. Cambridge Cambridge University Press.

Negueruela, E, J. P. Lantolf, Stefanie Jordan, and Jamie Gelabert. 2004. The 'Private Function' of Gesture in Second Languages Communicative Activity. A Study on Motion Verbs and Gesturing in English and Spanish. *International Journal of Applied Linguistics* 14 (1):115-159.

Rossini, Nicla. 2005. Phrasal verbs or words? Towards the analysis of gesture and prosody as indexes of lexicalisation, at On-line Proceedings of the 2nd ISGS Conference "Interacting Bodies" in Lyon, France.

Slobin, Dan Isaac. 1987. Thinking for speaking, at Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society.

Slobin, Dan Isaac. 1991. Learning to Think for Speaking: Native Language, Cognition, and Rhetorical Style. *Pragmatics* 1 (1):7-25.

Slobin, Dan Isaac. 1996. From ‘‘thought and language’’ to ‘‘thinking for speaking’. In *Rethinking linguistic relativity*, edited by J. J. Gumperz and S. C. Levinson. Cambridge: Cambridge University Press.

Slobin, Dan Isaac. 2003. Language and thought online: cognitive consequences of linguistic relativity. In *Language in mind: Advances in the study of language and thought*, edited by D. G. S. Goldin-Meadow. Cambridge, MA: MIT Press.

Slobin, Dan Isaac. 2004. The many ways to search for a frog: Linguistic typology and the expression of motion events. In *Relating events in narrative: Typological and contextual perspectives* edited by S. Strömqvist and L. Verhoeven. Mahwah, NJ: Lawrence Erlbaum Associates.

Slobin, Dan Isaac. 2006. What makes manner of motion salient. Explorations in linguistic typology, discourse, and cognition. In *Space in Languages: Linguistic systems and cognitive categories*, edited by M. H. S. Robert. Amsterdam/Philadelphia: John Benjamins.

Slobin, Dan Isaac, and N Hoiting. 1994. Reference to movement in spoken and signed languages: Typological considerations, at Proceedings of the Twentieth Annual Meeting of the Berkeley Linguistics Society.

Stam, Gale. 2006. Thinking for speaking about motion: L1 and L2 speech and gesture. *International Review of Applied Linguistics* 44:143-169.

Talmy, Leonard. 1985. Semantics and syntax of motion. In *Language typology and syntactic description, Vol. 3, Grammatical categories and the lexicon*, edited by T. Shopen. Cambridge: Cambridge UniversityPress.

Talmy, Leonard. 1991. Path to realization: A typology of event conflation, at Proceedings of the Seventeenth Annual Meeting of the Berkeley Linguistics Society.

Talmy, Leonard. 2000. *Toward a Cognitive Semantics*. Vol. II. Cambridge: The MIT Press.

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

Özyürek, Asli, Sotaro Kita, and S Allen. 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.