Proceedings of the 2nd European and the 5th Nordic Symposium on Multimodal Communication

August 6-8, 2014
Tartu, Estonia

Editors
Kristiina Jokinen & Martin Vels

Copyright

Linköping University Electronic Press
Linköping, Sweden, 2015
Linköping Electronic Conference Proceedings, No. 110
ISSN: 1650-3686
eISSN: 1650-3740
ISBN: 978-91-7519-074-7
URL: http://www.ep.liu.se/ecp_home/index.en.aspx?issue=110

Preface

The 2nd European and the 5th Nordic Symposium on Multimodal Communication

1 Introduction

Multimodal communication as a research area is growing rapidly. Both technological and social-linguistic approaches feature an increased interest in studying interactions with respect to communicative signals which do not only comprise of spoken language, but also hand gesturing, facial expressions, head movements, and body posture. Interactions at work places, school environments, health care and other services involve complex multimodal communication. Such embodied and situated communication is extended from human-human interactions to cover human interaction with intelligent agents such as robots and animated agents, as well as interaction with technological artefacts and ambient environments which affect communicative activities. The development of innovative computer interfaces, mobile media, and robotics provides new technical solutions to multimodal communication possibilities, while at the same time creating new challenges for communication research.

The proceedings contains the final versions of the accepted papers presented at the 5th Nordic and 2nd European Symposium on Multimodal Communication. The symposium took place 6-8 August, 2014, at the University of Tartu, and it followed the successful first symposium, organized at the University of Malta in 2013, which sought to broaden, with European dimension, the tradition established by the Nordic Symposia on Multimodal Communication held from 2003 to 2012. The symposia aim to provide a multidisciplinary forum for researchers from different disciplines studying multimodality in human communication and in human-computer interaction.

2 Research Training Course

The symposium was preceded by a research training course *Pointing to Gestures* which took place on 4-6 August, 2014, also at the University of Tartu. The training course focussed on gestures and their function in natural communication. Gestures have been identified as important signs that can appear in synchrony or non-synchronously with the speech, in order to identify references (deictic), describe speech content (iconic), and coordinate the communication in general (metalevel gesturing). Active research is being conducted concerning the use of gestures in natural conversations, as well as their automatic recognition and integration into interactive systems. Collection and analysis of high-quality video recordings has brought forward new accurate ways to study gestures as part of the interlocutors' communicative activity, and novel recognition devices, semi-automatic data analysis, and visualisation techniques enable investigations of various multimodal interaction phenomena in multidisciplinary framework.

The course provided an opportunity to study multimodal interaction phenomena, both from a theoretical and practical point of view, and allowed the student participants to discuss methodological and technical challenges related to their own work and in general, on research issues and data collection. The lecturers were the two invited speakers, Alan Cienki (VU University, Amsterdam and Moscow State Linguistic University) and Kirsten Bergmann (University of Bielefeld, Germany), as well as: Jens Allwood (University of Gothenburg), Elisabeth Ahlsen (University of Gothenburg), Patrizia Paggio (University of Copenhagen and University of Malta), and Graham Wilcock (University of Helsinki). The course and the symposium shared an excursion day on Wednesday, and the course participants were encouraged to attend the symposium as part of the course.

3 Symposium topics

In line with the preceding editions, the symposium accepted papers in a wide range of topics. However, this year the symposium was also linked to the research training course which concerned communicative gestures and their function in natural interactions, and thus the symposium encouraged interdisciplinary submissions, especially dealing with gesturing and gesticulation, automatic analysis of multimodal data, human-robot interaction, and multimodal processing. The submissions were evaluated by the international programme committee. After the symposium, extended contributions were invited for the post-proceedings of the symposium, and these submissions went through another reviewing process before being accepted in current volume. We believe that the ten submissions provide a good overview of the varied aspects of multimodal communication.

The conference also featured two invited speakers. Alan Cienki (VU University, Amsterdam, and Moscow State Linguistic University) focussed on human perspective in interaction and Kirsten Bergmann (University of Bielefeld) discussed about gestures and their modelling for human-robot interaction. Furthermore, the symposium provided a demonstration of the WikiTalk robot interaction system by Graham Wilcock (University of Helsinki).

4 Acknowledgements

We would like to extend our special thanks first to the programme committee members who promptly provided knowledgeable and constructive reviews of the submissions. Their work cannot be underestimated in preparing the interesting and multifaceted symposium programme and providing valuable comments to the final versions of the papers that then appear in the current proceedings.

The local organising team at the University of Tartu consisted of: Silvi Tenjes (chair), Kristiina Jokinen, Anne Kaaber, and Ingrid Rummo. Maria Gaiduk provided assistance for the webpage maintenance, while Martin Vels took care of preparing and finalizing the contributions to the right format for the proceedings.

The symposium was jointly organised by the Institute of Estonian and General Linguistics and the Institute of Computer Science at the University of Tartu, in collaboration with the Estonian Graduate School of Linguistics, Philosophy and Semiotics, the Estonian Science Foundation project MINT (Multimodal Interaction), and the European Regional Development Fund through the Estonian Center of Excellence in Computer Science, EXCS. We wish to express our sincere thanks to all of them for their generous support for the symposium, as well as for the research training course.

Kristiina Jokinen and Martin Vels

Institute of Computer Science
University of Tartu

5 Symposium organisers

Kristiina Jokinen, University of Tartu and University of Helsinki (chair)
Elisabeth Ahlsèn, University of Gothenburg
Jens Allwood, University of Gothenburg
Costanza Navarretta, University of Copenhagen
Patrizia Paggio, University of Copenhagen and University of Malta
Silvi Tenjes, University of Tartu

6 Local organisers

Silvi Tenjes, University of Tartu (chair)

Kristiina Jokinen, University of Tartu

Anne Kaaber, University of Tartu

Ingrid Rummo, University of Tartu

Maria Gaiduk (webpage maintenance)

Martin Vels (technical assistance)

7 Program committee

Elisabeth Ahlsèn, University of Gothenburg

Jens Allwood, University of Gothenburg

Loredana Cerrato, Trinity College Dublin

Rilla Chaled, University of Malta

Lin Chin, University of Illinois at Chicago

Onno Crasborn, Radboud University Nijmegen

Jens Edlund, KTH, Stockholm

Dirk Heylen, University of Twente

Kristiina Jokinen, University of Tartu

Michael Kipp, Hochschule Augsburg

Stefan Kopp, University of Bielefeld

Costanza Navarretta, University of Copenhagen

Patrizia Paggio, University of Copenhagen

Silvi Tenjes, University of Tartu

Laura Vince, University of Rome

Table of Contents

Gesture Use – From Real to Virtual Humans and Back

Kirsten Bergmann
Bielefeld University
Faculty of Technology, CITEC
P.O. Box 100 131
33501 Bielefeld, Germany
kirsten.bergmann@uni-bielefeld.de

When we are face to face with others, we use not only speech, but also a multitude of nonverbal behaviors to communicate with each other. A head nod expresses accordance with what someone else said before. A facial expression like a frown indicates doubts or misgivings about what one is hearing or seeing. A pointing gesture is used to refer to something. More complex movements or configurations of the hands depict the shape or size of an object. Of all these nonverbal behaviors, *gestures*, the spontaneous and meaningful hand motions that accompany speech, stand out as they are very closely linked to the semantic content of the speech they accompany, in both form and timing. Speech and gesture together comprise an utterance and externalize thought; they are believed to emerge from the same underlying cognitive representation and to be governed, at least in part, by the same cognitive processes (Kendon, 2004; McNeill, 2005). Despite this important role of co-speech gestures in communication, little is known, however, about the mechanisms that underlie gesture production in human speakers (cf. Bavelas et al. (2008), de Ruiter (2007)) as well as the functions gesture use fulfills in communication and even beyond, e.g. in educational or therapeutic contexts. My talk at the symposium showed how building computational simulation models of natural, communicative behavior and employing these models in virtual humans allows to address these research issues.

1 A computational model for iconic gesture production

With the GNetIc approach (Generation Networks for Iconic Gestures; Bergmann & Kopp (2009)) we proposed a computational framework to automatically generate novel gesture forms to be realized with a virtual human. Based on extensive empirical data from human-human interaction (SaGA corpus; Lücking (2013)), GNetIc accounts for a number of factors, identified in empirical corpus analyses, which can roughly be divided into three kinds. First, since the meaning of iconic gestures is explained by similarity or resemblance to their referent, the way how meaning is mapped onto gesture form is decisive. This implies that iconic gesture use is dependent on an underlying imagistic representation and rises the question how this representation is transformed into gesture form constrained by the use of different gestural representation techniques (e.g., placing, drawing, or posturing; cf. Kendon (2004)). Second, a gesture's form is also influenced by specific discourse contextual constraints as well as its linguistic context. For instance, speakers tend to employ gestures rather to introduce new information into the discourse, than to refer to what is already known and acknowledged by the dialogue partners (McNeill, 1992). And third, inter-individual differences in gesture use are quite obvious, reflecting an individual, speaker-specific gesture style. Individual speakers differ obviously with respect to gesture frequency, handedness (one-handed vs. two-handed gestures), preference for particular handshapes etc.

The GNetIc approach takes all these factors into account and allows to derive gesture forms on the basis of characteristics extracted from the imagistic representation of a referent object. Going beyond a straightforward meaning-form mapping, contextual factors like the given communicative goal, information state, or previous gesture use are also taken into account. In particular, different gestural representation techniques are considered, mediating the meaning-form mapping. By combining rule-based and data-based models, GNetIc can simulate both systematic patterns shared among several speakers, as

K. Jokinen and M. Vels. 2015. Proceedings of The 2nd European and the 5th Nordic Symposium on Multimodal Communication.

well as idiosyncratic patterns specific to an individual. That is, GNetIc can produce novel gestures as if being a certain speaker. Further, building and comparing networks from different speakers allows to gain insights into how production processes might differ from individual to individual.

2 How do human observers judge virtual humans using gestures?

In an evaluation study of the GNetIc model, human observers were provided with an object description given by a virtual human (Bergmann et al., 2010). We manipulated the agent's gestural behavior and addressed in how far different GNetIc models (individualized ones and an 'average' one learned from the aggregated data of several speakers) as well as control conditions (no gestures; randomized gesture use) affect the perception of human observers. Spoken words remained the same across all conditions. Results showed that the two individual GNetIc conditions outperformed the other conditions in that gestures were perceived as more helpful, overall comprehension of the multimodal presentation was rated higher, and the agent's mental image was judged as being more vivid. Similarly, the two individual GNetIc conditions outperformed the control conditions regarding agent perception in terms of likeability, competence, and human-likeness. Moreover, the aggregated GNetIc condition was rated worse than the individual GNetIc conditions throughout. And finally, the no gesture condition was rated more positively than the random condition. That is, it seems even better to make no gestures than to randomly generate gestural behavior. Overall, this study provides evidence that building generative models of co-verbal iconic gesture use, instead of using pre-defined gestures from a lexicon or 'gesticon', can yield encouraging results with actual users.

3 A virtual human as a tutor in second language learning

Beyond beneficial effects of virtual humans' gesture use regarding presentation quality and agent perception, virtual tutors also have potential to support learners' performance, e.g., in learning linguistic materials. In an empirical study we addressed whether the memory-supporting effect of iconic gesture imitation in vocabulary learning, so far shown for real human tutors (e.g. Macedonia et al. (2011)), is also valid for virtual tutors. The study employed a within-subject design manipulating the type of training in terms of (1) gesture-based training with human stimuli, (2) gesture-based training with agent stimuli, and (3) a control condition without any gestures. A total of 32 participants learned 45 vocabulary items on three consecutive days. Training materials comprised 45 nouns in German and 'Vimmi' (an artificial corpus created for experimental purposes to avoid associations and to control for different factors that might constrain the memorization of particular vocabulary items; see Macedonia et al. (2011)).

Short-term learning performance was measured the next day prior to the training session, respectively. The long-term effect of information decay was measured additionally four weeks after training was finished. In every test session, participants conducted a free and thereafter a cued recall test. Results showed, for both types of long-term measures (free and cued recall), better memory performance for items learned in the virtual human condition over items learned in the control condition. The same effect was present for short-term measures of free recall. Notably, in all tests performed, there was a trend for the virtual agent leading to better memory performance than training with a human (for details see Bergmann & Macedonia (2013)). These findings doubtlessly need follow-up studies to elucidate which factors and constraints constitute the beneficial effect of the virtual character. Nevertheless, they clearly substantiate the view that virtual pedagogical agents might play a crucial role in future language learning.

Acknowledgements

This research is supported by the Collaborative Research Centre SFB 673 "Alignment in Communication" and the Cluster of Excellence in "Cognitive Interaction Technology" (CITEC), both funded by the German Research Foundation (DFG).

References

J. Bavelas, J. Gerwing, C. Sutton, and D. Prevost. 2008. Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, 58:495–520.

K. Bergmann and S. Kopp. 2009. GNetIc—Using Bayesian decision networks for iconic gesture generation. In Z. Ruttkay, M. Kipp, A. Nijholt, and H. Vilhjalmsson, editors, *Proceedings of the 9th International Conference on Intelligent Virtual Agents*, pages 76–89. Springer, Berlin/Heidelberg.

K. Bergmann and M. Macedonia. 2013. A virtual agent as vocabulary trainer: Iconic gestures help to improve learners' memory performance. In *Proceedings of the 13th International Conference on Intelligent Virtual Agents*, pages 139–148, Berlin/Heidelberg. Springer.

Kirsten Bergmann, Stefan Kopp, and Friederike Eyssel. 2010. Individualized gesturing outperforms average gesturing–evaluating gesture production in virtual humans. In J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, and A. Safonova, editors, *Proceedings of the 10th Conference on Intelligent Virtual Agents*, pages 104–117, Berlin/Heidelberg. Springer.

J.P. de Ruiter. 2007. Some multimodal signals in humans. In *Proceedings of the 1st Workshop on Multimodal Output Generation*, pages 141–148. CTIT.

A. Kendon. 2004. *Gesture—Visible Action as Utterance*. Cambridge University Press.

A. Lücking, K. Bergmann, F. Hahn, S. Kopp, and H. Rieser. 2013. Data-based analysis of speech and gesture: the bielefeld speech and gesture alignment corpus (SaGA) and its applications. *Journal on Multimodal User Interfaces*, 7(1-2):5–18.

M. Macedonia, K. Müller, and A.D. Friederici. 2011. The impact of iconic gestures on foreign language word learning and its neural substrate. *Human Brain Mapping*, 32:982—998.

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

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

The Dynamic Scope of Relevant Behaviors in Talk:
A Perspective From Cognitive Linguistics

Alan Cienki
Vrije Universiteit/VU,
Amsterdam, Netherlands &
Moscow State Linguistic
University, Russia
a.cienki@vu.nl

Extended Abstract

Within the field of gesture studies, the many ways in which speech and gesture interrelate in the production of utterances has been the object of research for decades (at least since Kendon 1980). Nevertheless, it has only been in recent years that more linguists have begun to think seriously about the implications of this research for their theories of language. In cognitive linguistics, the framework of Cognitive Grammar (CG) is one that provides a means of taking account of gesture; this is a consequence of usage-based (Langacker 1988) nature of the theory. To be consistent with this framework, analyses in CG should work from the ground up to see what form-meaning (phonological-semantic) associations are abstracted by language users, that is: which ones become schematized and entrenched from communicative usage events. The phonological pole of a linguistic sign could conceivably include signals that are not just the audible sounds of spoken language, but also behaviors concomitant with language-based expression, such as gesture (Langacker 2008: 457), to the degree that they should also become schematized and entrenched in sufficiently consistent association with concepts (the semantic pole of a sign). The theory thus allows for linguistic signs to be multimodal (audio-visual in the case of speech and gesture) to varying degrees, based on the extent of schematization and entrenchment.

However, this elegant theoretical characterization, that appears to quite validly capture what language users do in practice, is a difficult one for linguists to do justice to in their analyses. In traditional studies within cognitive linguistics, language is treated as if it were a discrete category, separate from gesture, and the assumptions behind this are not problematized – this despite the fact that in CG, linguistic categories on various levels (phonemes, constructions, word meanings, etc.) are considered to involve continua and prototype categories rather than categories with strict boundaries.

The present study begins by considering the different degrees to which speakers' manual gestures are conventionally communicative. Kendon (1988) and McNeill (1992) discuss a continuum of gesture types, from more to less conventionally communicative signs. On one end are "emblem" gestures (e.g., think of a "thumbs up" gesture to show a positive evaluation of something). These are the most word-like in the stability of their sign status. However, on the other end is "spontaneous gesticulation": the manual movements that may relate to the contents of the speech in idiosyncratic ways, such as depicting selected aspects of the forms of entities one is talking about. Such depiction can take place in more detailed or more schematic ways, and can vary greatly in form across speakers and across usage events. Is there any way in which such gestures can sensibly be accounted for in a theory such as CG?

I propose the notion of a dynamic scope of relevant behaviors (SRB) as a way to take into account the varying kinds of sign-status that gestures can have. Elsewhere (Cienki 2012) I consider how the SRB can also help handle the variable relation of other behaviors to talk, such as the use of non-lexical utterances (such as uh and mm in English) and the use of hummed intonation contours without speech. The proposal is that in a communicative context, there is a dynamic scope of relevant behaviours, the scope of which may differ for the producer of the behaviours and for anyone paying attention to him

K. Jokinen and M. Vels. 2015. Proceedings of The 2nd European and the 5th Nordic Symposium on Multimodal Communication.

or her (the “attender[s]”) at any given moment. The scope has a focus and a periphery. In face-to-face communicative usage events between hearing people, spoken language is the default focus of the scope of relevant behaviors. A producer can flexibly make use of a smaller or larger scope of expressive behaviors, and an attender’s focus can also be narrower or broader and can change in size over the course of a conversation. The SRB is thus dynamic in terms of zooming in (taking just one behaviour into account) or zooming out (including more than one behaviour at a time) and in terms of its shifting focus, determining which behaviour(s) is/are in focus.

The phenomenon of self-repetition (of the verbal or gestural part of an utterance) provides a useful context for exploring how the properties of the construct proposed above operate in practise. Self-repetitions within topic units from a set of interviews in English from an American television talk show were analysed. Verbal repetitions were counted when they consisted of more than one word repeated, and gestural repetitions consisted of repeated use of more than two form features of the set of four that have become customary in gesture analysis since McNeill (1992), namely: hand shape, palm orientation, location, and movement type. The analysis revealed a number of patterns according to which the SRB may be expanded and/or contracted in the process of talk on the time scale of seconds or minutes. The types were interpreted as follows:

- Repetition involving contraction of the SRB: An idea goes from being presented with multiple articulators (spoken words produced orally and gestures produced manually) and/or more elaborate use of one or more form features to being presented with one mode of expression. The repetition thus involves reduction of behaviours on the part of the producer.
- Expansion of the SRB: An idea first expressed with a conventional linguistic sign is reintroduced using multiple articulators and/or more elaborate use of one or more form features. One can see this as a temporary loosening of symbolization.
- Expansion and then contraction of the SRB: An idea first presented verbally is reintroduced using multiple articulators (speech and gesture), with the speaker settling on one form of expression; momentary elaboration into a multimodal sign ensemble is followed by stabilization on the use of a monomodal sign.
- Maintenance of an expanded SRB: An idea first expressed with multiple articulators is repeated and part of it is symbolized in gesture for a small stretch of discourse – using what McNeill (1992) calls a gestural “catchment”. Information that was originally temporally integrated in what Enfield (2009) calls a “composite utterance” becomes decompressed, as it were, as relevant behaviours (speech and gesture) continue to serve somewhat different expressive functions over perhaps a few minutes of talk.

The SRB that we make use of varies, and this variation occurs not only in different ways but also along different time scales. Consequently, symbolization processes can also play out along different time scales, e.g., within topic units of less than a minute (microsymbolization via gestural catchments), across topics within a usage event (in the use of ad hoc words or gestures spontaneously ascribed a symbolic function within a conversation) and across genre events (in the gradual codification of functions/meanings with certain lexico-grammatical and/or gestural forms within communities of practice)

In conclusion, this research helps us reflect theoretically about how linguistic signs function in practice. Behaviours that repeatedly occur in the SRB paired with certain functions should be more likely to become more entrenched as symbolic structures or signs. But the present study suggests that signs need not be considered static entities with rigid boundaries, but rather may have relatively stable centers and variability in their boundaries. This also allows us to explain the overlap between the semiotic systems of spoken language and gesture (and other concurrent behaviours) as communicative signs in face-to-face interaction – an overlap which varies in degree over the time course of any usage event of talk.

Acknowledgements

Thanks to Suwei Wu for collection and transcription of the video data. This research has been supported by Russian Science Foundation grant #14-48-00067.

References

Cienki, Alan. 2012. Usage events of spoken language and the symbolic units we (may) abstract from them. In J. Badio & K. Kosecki (eds.), *Cognitive processes in language*, 149–158. Bern: Peter Lang.

Enfield, N. J. 2009. *The anatomy of meaning: Speech, gesture, and composite utterances.* Cambridge: Cambridge University Press.

Kendon, Adam. 1980. Gesticulation and speech: Two aspects of the process of utterance. In: M. R. Key (ed.), *The relation between verbal and nonverbal communication*, 207–227. The Hague: Mouton.

Kendon, Adam. 1988. How gestures can become like words. In Fernando Poyatos (ed.), *Cross-cultural perspectives in nonverbal communication*, 131–141. Lewiston, NY: C. J. Hogrefe.

Langacker, Ronald W. 1988. A usage-based model. In: Brygida Rudzka-Ostyn (ed.), *Topics in cognitive linguistics*, 127–161. Amsterdam: John Benjamins.

Langacker, Ronald W. 2008. *Cognitive grammar: A basic introduction*. Oxford: Oxford University Press.

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

Gestures Used in Word Search Episodes – by Persons with and without Aphasia

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

Abstract

This study investigates recurring patterns of gesturing during episodes of word search and own communication management, i.e. choice and change operations in speech. Two databases of word search episodes, one of person with aphasia and one of persons without aphasia, each containing 100 episodes, were analyzed. An extensive set of recurrent features of gesture and meaning connections was identified. In general, the same recurrent features were found for persons with and without aphasia. The recurrent gesture patterns can be used for educating health personnel and families of persons with aphasia in order to enhance understanding or aphasic gesturing.

1 Introduction

1.1 Background on aphasia and gesture

How gestures interact with words in conveying meaning is a question, which has attracted attention for some time (cf. Kendon 1983, 2004). Theories differ with respect to the role of gesturing for persons with aphasia. Some researchers claim that speech and gesture are co-generated and cannot be generated separately – if speech is disturbed, so is gesture (e.g. McNeill 1985, 1992, 2000, 2007). Others claim that speech and gesture are separate systems – then gesture can be used if speech is disturbed (e.g. Hadar and Butterworth 1997, Beattie and Shovelton 2000, 2002, 2011). Most likely is perhaps that gesture and speech are closely linked but to also some extent independent, which would also make compensatory gesturing possible. This has been assumed in a number of studies, that have shown compensatory gesturing by persons with aphasia in spontaneous communication (e.g. Ahlsén 1985, 1990,1999, Feyereisen 1991, Lott 1994, LeMay et al. 1988, Macauley and Handley 2005).

A related question is whether gesturing helps word finding and communication more in general for the speaker. A number of studies have claimed that this is the case, in general (Kita 2000, Kraus et al. 2000, Melinger and Kita 2006, Rauscher et al. 1996) and for persons with aphasia (e.g. deRuiter 2006). Studies of brain activity have also addressed this question (Willems et al. 2007, Wu & Coulson 2007).

A recent study showed that persons with and without aphasia use iconic-illustrating gestures accompanying verbs and nouns, much in the same way in spontaneous conversations. There were some differences, i.e., how much eye contact was maintained, the number of one vs. two hand gestures and to some extent the "complexity" of gestures. We also found that clearly iconic/illustrating gestures occurring during word search episodes can be found in at least 25-30% of the cases, for persons with and without aphasia (Ahlsén and Schwarz 2013).

Departing from these results, the general questions investigated in the present study are: (i) What are the features of iconic/illustrating gestures occurring during word search episodes, i.e., when a speaker is struggling to find the right words to express something, giving rise to hesitation and/or self-interruption and reformulation henceforth called OCM = Own Communication Management, see further below)? and (ii) What are the features of the rest of the gestures occurring with word search episodes?

K. Jokinen and M. Vels. 2015. Proceedings of The 2nd European and the 5th Nordic Symposium on Multimodal Communication.

1.2 Background on word search and OCM (Own Communication Management)

Behaviour during own communication management can be of two main types. The first type is related to choice operations. Such behaviour can be pause, hesitation sounds, like “eh”, lengthening of continuants or OCM phrases like “what’s it called”. The second type involves change of expression, e.g. self-interruption, self-repetition and/or reformulation. Combinations of several of these phenomena of either type or of both types are common. In Allwood, Nivre and Ahlsén (1990) a typology of OCM phenomena is presented, where OCM units can be either a single type of feature or a combination of features. Single features can be either basic OCM expressions (pause, hesitation sound, OCM expression or OCM phrase) or basic OCM operations (lengthening, self interruption and self repetition). These features were mainly identified in relation to speech, whereas the present study concerns multimodal communication, including gesture in a wide sense.

1.3 Research questions

The specific research questions to be addressed in this study are:

1) What gesture types and functions occur in episodes of OCM/word search/reformulation for persons with and without aphasia?

2) What are the applications of our findings for understanding communication of persons with aphasia?

2 Method of the empirical study

2.1 Database

The database consisted of gestures in 100 word search episodes for persons with aphasia and- 100 word search episodes for persons without aphasia, videorecorded in informatal conversation or narration. Each of the two parts of the database contained data from 10 different persons. The databases were collected in earlier projects and the episodes were chosen by extracting consecutive sequences of relevance. The participating speakers with aphasia had morderate to mild aphasia, according to the BDAE (Goodglass and Kaplan 1973). All participants were between 20 and 60 years old, the gender distribution was equal in both data sets and the types of conversation and narration were comparable, i.e. they were all collected in studio or studio-like environments during informal conversation on various evdryday topics. All videorecordings had been transcribed and annotated, however, the gesture analysis was made in the present study.

2.2 Gesture coding

The gestures occurring in OCM episodes (word search/reformulation episodes), where they occurred were coded with respect to the following features:

- Topic of discussion,
- Preceding, accompanying and succeeding speech and kinetic context,
- Gesture components for hand gestures (see specification below),
- Iconic and illustrative features of the gestures.

The coded components of hand gestures were:

- Configuration
- Plane
- Orientation
- Straight line movement or curved movement
- Laterality vs. symmetry
- Localization relative to body
- Localization in space
- Accompanying eye gaze, head movement
- Recurrence

Each of the gesture episodes was also coded for choice (i.e. pause, hesitation) or change function (i.e. self-interruption, reformulation.

2.3 Analysis

A number of patterns of gesture features were identified. After this, typical examples of the different patterns were selected for qualitative microanalysis in context. A number of types and functions related to specific gesture features were suggested. Finally, the features were compared between the corpora for aphasic and non-aphasic communication. All gesture annotation was checked by two coders and determined by joint re-analysis, if there were discrepancies.

3 Results - Recurrent patterns of illustrating gestures

3.1 Reference to person (in a wide sense)

If we start by the hand movements showing direction in space and reference to person, we find recurrent hand directions. Moving one hand towards your own body is used for reference to "I", "me", and "one" (both by persons with and without aphasia). Moving a hand towards one's head is used with reference to mental processes, "think", "read" (in the aphasia group), "remember", "forget" (in both groups) and also for "generalize" (in the non-aphasic data). One hand towards the interlocutor is used as reference to the interlocutor, for turn giving and for asking for help; it also occurs with "as you know" (in both groups).

3.2 Spatial orientation (other than person reference)

Hand away from body
Moving one hand away from the body and forward is associated with many related meanings, like: "out", "away", "ahead" and metaphorically "promote" (in both groups). If the movement is pushing or moving up and down with the palm directed forward, this is associated with "marking a limit", "protect" and with the utterance "he is dead".
When one hand is moved sideways with the palm directed to the side, this is connected to "throw away", "get rid of" and "reject" (in both groups), and also to the more abstract "confrontation" (in the non-aphasic data). If the hand is moved sideways, with the palm instead directed downwards, this is connected to reference to "landscape", "fields", "ground." and the more abstract "exploitation" (all in the non-aphasic data).

Other hand and finger movements in the air
The hand circling (or cyclic hand movement) occurs in the context of "progress", "action", "forward", "start" and "eagerness" and when the circling movement is directed towards the speaker's own body: "own emotion" (in both groups).

Finger movements during another hand gesture indicate turn keeping, especially during word search (in both groups). Both hands are moving upwards with: "raise" and "lift" (in the non-aphasic group), while both hands move downwards with "dead" (also in the non-aphasic group). Both hands come together in relation to words like "summary", "join", "agree", "group" and "totality" (in the non-aphasic data).

Hand towards table
Moving the hand towards the table is connected with a number of actions that are done on a table, e.g. to writing, sorting etc. Some examples are: "stamp" (simple movement) or small movements: for "categorize", "classify" (in the non-aphasic data, and "read", "one by one", "lines", and "map out every word" (aphasia data).

3.3 Raised hand (choice), shaking movement of hand or head (change)

There are a number of gesture features indicating own communication management, i.e. choice and change operations, when a person is searching for words, hesitating or self-interrupting and repeating or reformulating their speech. The most typical word choice and hesitation gesture is hand raised, palm up, with small movements or shaking. In both groups, this gesture occurs with phrases like "What's it called?", involving "choice" operations, e.g. word finding problems. See examples 1 and 2 below.

Example 1. Raised hand – choice (non-aphasic data)
upload ... eh n/ like one
[left hand small movements up and down, palm up]

Example 2. Raised hand – choice (aphasia data):
then it was… eh then: it was eh … eh what's it called
[palm up, fingers moving quickly, ending with fingers still and hand moving up-down]

A shake of a raised hand with the palm turned up as well as a shake of the head are used by both groups in a "change" situation, i.e. when an error is corrected by self-interruption and reformulation. The shaking movement is, thus, connected to cancellation/denial of one's own speech production and self-correction.

Example 3. Head shake – change
invoi+ eh ... classification system
[gaze down, head shake]

Example 4. Head shake – change.
then it so/ ... flashed or what it is called
[head shake]

Example 4. Head shake – change.
it was I myself that myself the oth/ it they didn't notice
[head shake]

Pointing
Pointing indicates direction, and place (in both groups).

3.4 Examples of illustrating gestures

If we look more closely on iconicity in gesturing, pantomimes occur frequently, in different forms. In both groups, It is also common that the concrete part of an action is performed, or the outline of an object or the concrete part of an object is given, often with the use of both hands (if applicable and possible).

Example 5. Index-Icon Illustrating mental process (reading and listening to the same text)
the more eh ... directions ... it ... comes from
[left hand index finger pointing, semicircles from different directions towards own head
that eh ... more it sticks
[left hand towards left ear, making "pushing" movements with fist]

Example 6. Gesture indicating interest with asynchronous movements of the left and the right hand:
Like eh … like eh … the text … is interesting then you know
[both hands moving in small circles, not synchronized]

Example 7. Gesture indicating activity, dynamics using both hands in synchrony.
s*o so eh ... so you can ... get started*
[both hands, palms directed toward each other, small circular movements]

4 Conclusions

4.1 Recurrent gesture patterns carrying meaning

A number of recurrent gesture patterns with systematic relations to meaning were identified in this study. This means that gestures in context of word search and own communication management, which often precede or replace corresponding spoken words, convey information to the listener about several aspects of the meaning of a multimodal contribution. As we have seen above, aspects which were identified were reference to objects and persons in space, reference to different persons in the interaction, information about which type of action, property or object is being referred to – if it is fast or slow, sudden or even, if it is stretched out along a line or surface or an on-going process, if it involved symmetric or asymmetric movement etc. In relation to word finding problems, gestures also convey important information about the speaker's attitude to what he/she has just said, especially if it should be retracted, changed or disregarded. Furthermore, there are features of typical word search gestures, i.e. raising one hand, which provide cues about the actual word search process and possibly also about the intended target words. A repertoire of more specific recurring gesture features related to more specific meanings exist, as described and exemplified above.

4.2 How different is aphasic gesturing?

In general, the same types of gesturing were found in aphasic as in non-aphasic gesturing, which indicates that there is no major difference of the kind that would suggest a gesture disorder comparable to the speech-language disorder of aphasia. As identified by Ahlsén and Schwarz (2013), there are, however, a few differences, which can, at least to some extent, be considered as secondary effects, caused by other primary motor or aphasic difficulties.

(i) One difference is the more frequent use of only the left hand by the persons with aphasia, which is even more pronounced in contexts of own communication management. This is in most cases an effect of earlier right-hand hemiplegia, which has caused residual weakness of the right arm and hand and/or a change of habitual gesturing that has remained. Since a) bimanual gesturing has been considered more complex (it also gives wider possibilities) and b) left hand gesturing can be less precise than right hand gesturing, the frequent use of only the left hand easily leads to less complex gestures. Less complex gestures can, therefore, not unambiguously be ascribed to less complex semantics, although there may be a disorder directly affecting gestures, to some extent.

(ii) Another difference is that in aphasic gesturing there is quantitatively more marking of word search and more pointing to the interlocutor than in non-aphasic gesturing. There is, however, most often some content feature or features in the gestures produced by PWA. The gestures, thus, reflect difficulties in speech-language.

(iii) The aphasic gesturing also involves considerably more gaze aversion during word search, both compared to other word production in the same persons and compared to word search in persons without aphasia. This strongly points to a higher cognitive load involved in word search, caused by anomia.

5 Discussion

Proponents of a very tight relation, co-generation and interdependence between words and gestures have inspired a view that assumes that if a disorder of word finding also means a corresponding disorder of gesture finding. In aphasia with anomia, especially fluent, Wernicke-type aphasia and global aphasia, the frequent hand movements and pointing have not been considered as illustrating or indicating content. In our data, however, we find very few cases of what has been called "indecisive hand waving" during OCM/word search/ reformulation. Gestures usually refer to something and/or fill an interactive or own activation function. This applies to persons with aphasia as well as persons without aphasia. There are patterns of recurrent gesturing indicating just search for words and lack of expression, but most often some indication of the word searched for actually occurs. The patterns found in this study for Swedish were in many cases similar to those found by Bressem and Müllller (in press) and Ladewig (in press) for German.

Reference conveyed by gesture may be both concrete and abstract – where the concrete part is being shown. It can be vague and general or precise and specific. Indexical gestures are common; some iconic feature is often included. Pantomimes also occur. Access to the linguistic and situational context is, however, essential for the interpretation of recurrent but often polysemous gesture patterns.

6 Possible applications

If gesture patterns during word search episodes are, to some extent, consistent, they give clues to the intended word. These clues can be systematically learned by people communicating with persons with aphasia – to some extent they can be intuitive, but they could be more consciously applied in interpreting communication attempts. Tutorials could, for example, be made, in order to make conversation partners of persons with aphasia aware of recurring patterns of gestures. The persons in the study often produced also verbal clues or the actual target word after the gesture.

Other persons with more severe aphasia than the ones on our data might not be able to do this. Still, the gesture patterns might be preserved, and, in these cases, could be very helpful for communication. More systematic studies of the recurrent features of gesturing in persons with severe aphasia, taking the full context into consideration, could investigate this possibility.

Acknowledgments

The research that has led to this work was supported by the project "Semantic processes in word finding and gesture", which was funded by the Swedish Research Council, grant number 2006-1598.

References

Elisabeth Ahlsén. 1985. *Discourse patterns in aphasia. Gothenburg Monographs in Linguistics, 5,* University of Gothenburg, Department of Linguistics.

Elisabeth Ahlsén. 1991. Body communication and speech in a Wernicke's aphasic – A longitudinal study. *Journal of Communication Disorders*, 24:1-12.

Elisabeth Ahlsén. 2002. Speech, vision and aphasic communication. In Paul Mc Kevitt, Sean. O'Nualláin, and Colwyn Mulvihill (Eds.), *Language, vision and music* (pp. 137-148). John Benjamins, Amsterdam.

Elisabeth Ahlsén and Anneli Schwarz. 2013. Features of aphasic gesturing – an exploratory study of features in gestures produced by persons with and without aphasia. *Clinical Linguistics and Phonetics* Oct-Nov;27(1011): 823-836.

Jens Allwood, Elisabeth Ahlsén, Johan Lund, and Johanna Sundqvist. 2007. Multimodality in own communication management. In Juhani Toivanen and Peter Juel Henrichsen (Eds.), *Current trends in research on spoken language in the Nordic countries,* Vol. II (pp. 10–19). Oulu University Press, Oulu, Finland.

Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. 1990. Speech management – on the non-written life of speech. *Nordic Journal of Linguistics*, 13:3-48.

Geoffrey W. Beattie and Heather K Shovelton. 2000. Iconic hand gestures and predictability of words in context in spontaneous speech. *British Journal of Psychology*, 91:473–492.

Geoffrey W. Beattie and Heather K. Shovelton, 2002. An experimental investigation of some properties of individual iconic gestures that mediate their communicative power. *British Journal of Psychology,* 93:179–192.

Geoffrey, W. Beattie and Heather K. Shovelton. 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.

Jana Bressem and Cornelia Müller. A repertoire of German recurrent gestures with pragmatic functions. In Cornelia Müller, Alan Cienki, Ellen Fricke, Silva H. Ladewig, David McNeill, and Jana Bressem, (Eds.). In press. Volume 2. *Body – Language – Communication: An International Handbook on Multimodality in Human Interaction.* For the series *Handbücher zur Sprach- und Kommunikationswissenschaft (HSK).* De Gruyter Mouton, Berlin.

Jan P. De Ruiter. 2006. Can gesticulation help aphasic people speak, or rather, communicate? *Advances in Speech-Language Pathology,* 8:124–127.

Pierre Feyereisen. 1991. Communicative behavior in aphasia. *Aphasiology,* 5 (4-5): 323-333.

Uri Hadar and Brian Butterworth. 1997. Iconic gesture, imagery and word retrieval in speech. *Semiotica,* 115: 147–172.

Adam Kendon. 1983. Gesture and speech: How they interact. In John M. Wiemann & Randall P. Harrison (Eds.), *Nonverbal interaction* (pp. 13–45). Sage Publications, Beverly Hills, CA.

Adam Kendon. 2004. *Gesture: Visible action as utterance.* Cambridge University Press. Cambridge, UK.

Sotaro Kita. 2000. How representational gestures help speaking. In David McNeill (Ed.), *Language and gesture: Window into thought and action* (pp. 162–185). Cambridge University Press, Cambirdge, UK.

Robert M. Krauss, Yihsiu Chen, and Rebecca F. Gottesman. 2000. Lexical gestures and lexical access: A process model. In David McNeill (Ed.), *Language and gesture* (pp. 261–283). Cambridge University Press, New York.

Silva H. Ladewig. Recurrent gestures.. In Cornelia Müller, Alan Cienki, Ellen Fricke, Silva H. Ladewig, David McNeill, and Jana Bressem, (Eds.), In press. Volume 2. *Body – Language – Communication: An International Handbook on Multimodality in Human Interaction.* For the series *Handbücher zur Sprach- und Kommunikationswissenschaft (HSK)*. De Gruyter Mouton, Berlin.

Amanda LeMay, Rachel David, and Andrew P. Thomas. 1988. The use of spontaneous gesture by aphasic patients. *Aphasiology*, 2:137–145.

Petra Lott. 1999. *Gesture and aphasia.* Peter Lang, Bern.

Beth L. Macauley and Candace, L. Handley. 2005. Conversational gesture production by aphasic patients with ideomotor apraxia. *Contemporary Issues in Communication Sciences and Disorders,* 32:30–37.

David McNeill. 1985. So you think gestures are nonverbal? *Psychological Review*, 92:350–371.

David McNeill. 1992. *Hand and mind.* The University of Chicago Press, Chicago.

David McNeill. 2000. *Language and gesture.* Cambridge University Press, Cambridge, UK.

David McNeill. 2007. *Gesture and thought.* University of Chicago Press, Chicago.

Alissa Melinger and Sotaro Kita. 2006. Conceptual load triggers gesture production. *Language and Cognitive Processes*, 22:473–500.

Frances H. Rauscher, Robert M. Krauss, and Yihsiu Chen, Y. 1996. Gesture, speech and lexical access: The role of lexical movements in speech production. *Psychological Science,* 7:226–231.

Roel M. Willems, Azli Özürek, and Peter Hagoort. 2007. When language meets action: The neural integration of gesture and speech. *Cerebral Cortex*, 17:2322–2333.

Ying Choon Wu, and Seana Coulson. 2007. How iconic gestures enhance communication: An ERP study. *Brain and Language*, 101:234–245.

Verbally Assisted Haptic Graph Comprehension: The Role of Taking Initiative in a Joint Activity

Özge Alaçam
Department of Informatics
University of Hamburg
Hamburg/Germany
alacam@informatik.uni-hamburg.de

Christopher Habel
Department of Informatics
University of Hamburg
Hamburg/Germany
habel@informatik.uni-hamburg.de

Cengiz Acartürk
Cognitive Science
Middle East Technical University, Ankara/Turkey
acarturk@metu.edu.tr

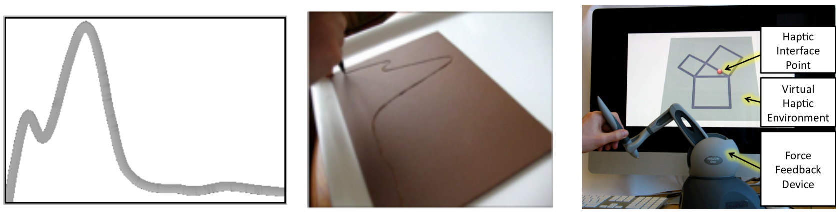
Abstract

Statistical graphs are tools for multimodal communication in daily life settings. For visually impaired people, haptic interfaces provide perceptual access to the information provided by the graph. Haptic comprehension can be facilitated by audio and verbal assistance. We investigate the circumstances under which verbal assistance facilitates haptic comprehension of graphs. For this, we focus on cases where unaided haptic graph comprehension has limitations, such as when describing a global (rather than a local) maximum. In an experiment that employed a joint activity setting, we observed two major factors for facilitating haptic-graph comprehension by providing verbal assistance: First, the results revealed significant impact of the explorer as a dialogue initiator during the course of haptic exploration. Second, verbal assistance led to more successful graph comprehension when it was enriched by modifiers.

Keywords: haptic line graph comprehension; sketching, verbal assistance, turn-taking

1 Haptic Audio Line-Graph Exploration

Presenting and representing information in visuo-spatial formats, such as graphs, maps or diagrams, is important, as well as successful, for thinking, problem solving and communication (Hegarty, 2011). Their usage covers science and education settings, and also the news media and economy bulletins. Thus, there have been continuous efforts for the inclusion of blind and visually impaired people in using these visuo-spatial representations. For example, users can explore haptic graphs (Figure 1.a) by hand movements following graph lines engraved in a (real) physical plane (Figure 1.b) or by using a force-feedback device, for instance a Phantom Omni® (recently Geomagic® Touch^{TM}, see Figure 1.c), to explore virtual graph lines, i.e. lines engraved in a virtual plane. Comprehension of haptic line graphs is based on exploration processes with the goal to collect information provided by the geometrical properties of the line explored; in particular, shape properties and shape entities—as concavities and convexities, maxima and minima, corners and smooth turning points—have to be detected.



(a) (b) (c)

Figure 1: (a) Sample haptic graph, (b) Exploration of a physical haptic map and (c) Phantom Omni® device and visualization in the domain of geometry

Haptic graphs may be conceived as efficient interfaces in providing access to those spatial and conceptual structures through the haptic sensory modality. On the other hand, haptic information access has a

K. Jokinen and M. Vels. 2015. Proceedings of The 2nd European and the 5th Nordic Symposium on Multimodal Communication.

lower *bandwidth* compared to visual information access, and additionally, haptic exploration is sequential, while visual perception allows the perception of both local and global information about a graph at one glance.[1]

In order to bridge this bandwidth gap between haptic exploration and visual exploration of graphs, and in order to provide sufficient information access to haptic graph readers, haptic graphs should be accompanied by alternative modalities, in particular modalities that provide additional verbal or audio information. Therefore, in addition to pure haptic graphs, haptic-audio interfaces have been developed to provide perceptual access to spatial representations, thus facilitating comprehension of spatial displays by visually impaired (see, e.g. Yu and Brewster, 2003; Zhao et al., 2008; Abu Doush et al., 2010). But there still remains much need for further development of specific types of haptic-spatial interfaces and the need for research that focuses on peculiar characteristics of the interface design.

Statistical graphs do not only present data but they also provide perceptual access to second-order entities, such as extrema, trends and trend changes. In particular, the properties of the line shape allow distinguishing a global maximum from a set of local maxima, or detecting inflection points that depict trend changes. We propose that verbal assistance may facilitate overcoming—for example—the problem of distinguishing local maxima from the global maximum, and problems similar to the local maxima problem, by providing necessary information through the auditory channel (see, e.g., Alaçam et al., 2013a). Our long-term goal is to realize an automatic, i.e., computational, verbal assistance system that has the capability to provide instantaneous support for haptic-graph explorers during their course of exploration. For developing such an assistance system—beyond realizing components considering natural language generation and user-system interaction—empirical studies are needed to understand the underlying principles of haptic graph exploration, of conceptualizing graphs, and of communicating about graphs.

According to this, a specific challenge is faced for designing haptic graphs: It is necessary to determine which concepts depicted by the graph—or by the segments of the graph—are appreciated as *important*. This challenge becomes significant when designing haptic line graphs with several local maxima, as opposed to simple graph lines with a single global maximum, because in contrast to visual exploration, a local maximum cannot be recognized as a global maximum during the course of haptic exploration.

To investigate the circumstances under which verbal assistance facilitates haptic comprehension of graphs, we designed an experimental setting, in which two participants perform joint activity for graph exploration, thus performing verbal assistance for haptic graph exploration (Alaçam et al., 2013a). It is a task-oriented joint activity (Clark, 1996) of two agents, a (visually impaired or blindfolded)[2] explorer (E) of a haptic graph and an observing assistant (A) providing verbal assistance, as depicted in Figure 2.

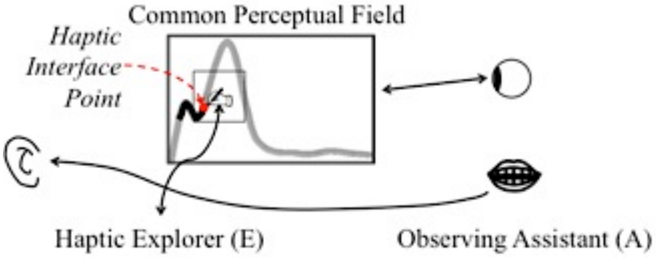


Figure 2: Assisted haptic graph exploration, a joint activity

[1] See, Loomis et. al.,1991, and Loomis and Klatzky, 2008, on the concept of *sensory bandwidth*, i.e. the information-carrying capacity of the sensory modalities, and on its role for cross-modal integration. Yu and Brewster, 2003, discuss this topic from the haptic-interface perspective.

[2] In our experiments either blindfolded sighted or visually impaired participants explorers participated (no mixed groups). Whereas blindfolded explorers have experience in graph reading, many blind and visually impaired people have less experience and competence with respect to graphs. Acartürk, Alacam, and Habel (2014) discuss some differences and similarities between these groups. We assume, that the tasks (1) giving verbal assistance for late-blind graph explorers and (2) giving verbal assistance for early-blind graph-reading novices only partially overlap.

A successful communication through graphs and language usually requires the integration of information contributed by both graphical entities and verbal entities so that the reader arrives at integrated conceptual and spatial representations. We have investigated various aspects of such integrated conceptual and spatial representations both from a theoretical perspective and in empirical studies (Habel and Acartürk, 2007, Acartürk, 2010). In the present paper we focus on haptic graph exploration as a collaborative activity between two humans, A and E: They share a common field of perception, namely the haptic graph, but their perception and comprehension processes differ significantly. For example, while (visually impaired or blindfolded) E explores the highlighted, black segment of the haptic graph (Figure 2), A perceives not only this segment but also the global shape of the graph. In particular, A is aware of shape landmarks and line segments. Similarly, when E starts exploring the first local maximum followed by a local minimum, E does not have any information about the global maximum, which is already part of A's knowledge. Thus, the haptic explorer E and the assistant A have different internal representations of the graph line, and A's referring to the graph could augment E's internal model substantially. At this moment (corresponding to the position depicted in Figure 2) A can take the initiative, and starts uttering "Now you have reached the heights of the last peak" to provide E with additional information. Another suitable comment would be "You are in the increase to the population maximum", or even "You are in the increase to the population maximum of about 90, that was reached in 1985". Alternatively, E can take the initiative before A, by asking for advice. Initiating a dialogue has a significant role in constructing the alignment in the dialogue because it aims at making explicit the missing information (or information that is difficult to comprehend during the course of exploration), which is necessary for achieving the alignment. The joint activity in our case can be considered as an asymmetric dialogue, namely haptic-explorer-dominant, because the participants in the experiment (see below) were instructed that for efficient and effective verbal assistance systems, the haptic explorer initiates the help request and the verbal assistant provides help based on the explorer's need. This asymmetry is due to the asymmetry of participants' roles during the course of joint activity: assisted person (E) – assisting person (A) [corresponding, e.g. to 'stranger' – 'local' in route-instruction dialogues].

The success of the joint activity of the explorer E and the observing assistant A in general, and in particular the success of their task-oriented dialogue, depend on the alignment of the interlocutors' internal models, especially on building implicit common ground (Garrod and Pickering, 2004). E's internal model of the activity space, i.e. the haptic graph and E's explorations, is perceived via haptic and motor sensation, whereas A's internal model of the same space is built up by visual perception. Therefore similarities and differences in their conceptualization play a central role in aligning at the situation-model level. To be assistive, A should provide E verbally with content which is difficult to acquire haptically. This—haptically difficult to be built up—content has to be combined with haptically-explored content in the same sentence (or phrase) to fulfill the given-new contract (Clark and Haviland, 1977). In our study, the verbal assistant is expected to provide most helpful and relevant information for the haptic explorer at that particular moment among all possible information that can be derived from the representation, by taking into account haptic explorer´s previous actions on the graph and previous utterances. The motivation that underlies this expectation is that the content of the verbal assistance has the potential to influence the alignment process, thus leading to a better or worse comprehension of the haptic graphs. As described in more detail below, we focus on the role verbal assistance content, as well as the role of dialogue initiating in the present study.

In the investigation of haptic graph comprehension, we have employed various methodologies including analysis of gestures, referring expressions and haptic exploration movements. Our previous research (Alaçam et al., 2013b) on accompanying gestures produced during verbal description of haptic graphs showed that in haptic graph comprehension, speech-accompanying gestures are usually produced with expressions that highlight relevant shape properties. Moreover, the analysis of referring expressions in the dialogues provides insight about how the haptic explorer comprehend the data, parse it for naming, and recognize which graphical elements are salient and which are hard to distinguish from the others. In addition to the analyses of gesture and referring expression production, sketch analysis of explored graphs is an appropriate methodology to evaluate the conceptualization of the events represented by the graph. As stated by Tversky (1999), "drawings reveal people's conceptions of things, not their perceptions of things". In particular, sketches provide complementary data for the analysis, because they can reveal details that the graph reader skips in verbal description for various reasons (e.g., the concept may be hard to express verbally or it may be considered as redundant by the reader). Accordingly, we employ the analysis of post-exploration sketches in the present study.

To sum up, we focus on two factors highly relevant for designing a computational verbal assistant for haptic exploration of line graphs: (1) the role of the haptic explorer as dialogue-initiator and as a consequence the assistants' competence to interact cooperatively, and (2) the role of the type and richness of content the verbal assistance provides. The empirical study presented in the paper was conducted with blindfolded participants to investigate modality-dependent differences by keeping other variables (i.e. the modality used to acquire prior graph-domain knowledge) constant and also to clarify which pieces of content should be provided by the modalities.

2 Experiment

In order to investigate the contribution of verbal assistance in haptic graph comprehension, we conducted the experiment in two conditions and performed a comparative analysis of the results. The first condition examined haptic exploration of line graphs by single, blindfolded participants, in the absence of verbal assistance. In the second condition, participant pairs (a blindfolded haptic explorer and a verbal assistant who was able to observe the haptic exploration) collaborated in exploring the haptic line graphs of Condition 1. In both conditions, each single session took approximately one hour. Haptic explorers were presented a warm-up session to get familiarized with the equipment (in this case, Geomagic Touch, formerly Sensable Phantom Omni®, Figure 1). Then they were presented with the stimuli. The stimuli included five haptic line graphs with smooth edges (Figure 3, two additional graphs were employed for familiarization with haptic line graphs). The participants were informed that they were presented bird-population graphs. The graphs were presented in randomized order. During the course of the experiment session, the explorer-participants performed haptic graph exploration by moving the handle of the haptic device, which can be moved in three spatial dimensions (with six degree-of-freedom). The line (proper) in the graphs was represented by engraved concavities on a horizontal plane; therefore the graph readers perceived the line as deeper than the other regions of the surface. Due to the interface limitations in the haptic representation, the numerical data labels were not presented. After the exploration of each graph, the participants produced single-sentence verbal descriptions of the graphs to a hypothetical audience; their spontaneous speech accompanying gestures produced during verbal descriptions were also recorded (these data are not in the focus of the present study). After the verbal description, they produced a sketch of the graph with paper and pencil. Two raters scored the sketches (all raters were blind to the goals of the study) for their similarity to the stimulus-graphs by using a 1 (least similar) to 5 (most similar) Likert Scale. The graphs employed in this study were taken from a publicly available consensus report (PRBO, 2012) and redesigned for the purposes of the study. The line graph stimuli were selected to represent a variety of patterns in terms of the number and polarity of curvature landmarks, length and direction of segments (Figure 3).

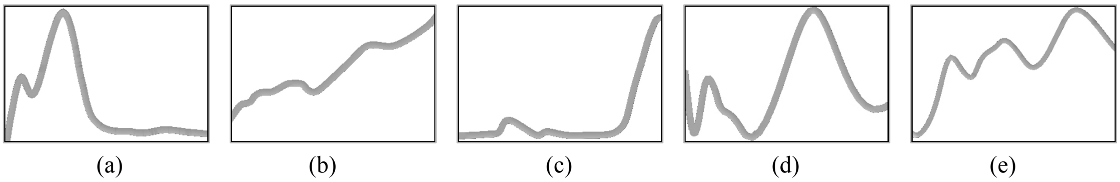


Figure 3: Stimulus-graphs (from graph-a to graph-e)

In Condition 1, nine university students (four female, *Mean age* = 25.0, *SD* = 6.3) performed haptic exploration of the stimuli without verbal assistance, and they completed the experimental protocol presented above. The participants of Condition 2 were pairs of students (13 pairs, 11 female, *M*=25.3, *SD*=3.27). Each pair was composed of a haptic explorer (E) and a verbal assistant (A). The haptic explorers were informed that the goal of the study was to design efficient and effective verbal assistance systems and the haptic explorer initiates the help request. The task of the verbal assistant was to provide the necessary information in a short description when required by the haptic explorer. The haptic explorer and the verbal assistant were located in separate rooms, so they communicated through speakers. While E explored the graph haptically throughout the experimental session, A had visual access to both the graph and E's exploration process (i.e. the current position E on the graph), which was displayed, on the computer screen together with the graph. Both E and A followed the experimental protocol employed in Condition 1 (i.e., a single-sentence verbal summary and sketch production).

The results of Condition 2 showed that out of 65 experimental protocols of graph stimuli, 28 protocols involved at least one request from the verbal assistant. In other words, the haptic explorers were dialogue initiators in 28 of 65 experiment protocols of the graph stimuli. Since this corresponds to almost half of the pairs in Condition 2, we decided to conduct a further analysis of data by dividing the protocols of Condition 2 into two groups: (1) Dialogue-initiated protocols (henceforth, dialogue-initiator protocols) (2) The protocols that involved no dialogue initiative, thus no verbal assistance (henceforth, no-initiator protocols). In this study, we focused on the content of the dialogue and the similarity of participants' sketches to the stimulus-graphs as a performance measure. For the analysis of the sketches, inter-rater reliability between the two raters was assessed using a two-way mixed, consistency average-measures ICC (Intra-class correlation). The resulting ICC=.69 can be classified as "good reliability" (Cicchetti, 1994; p. 286).

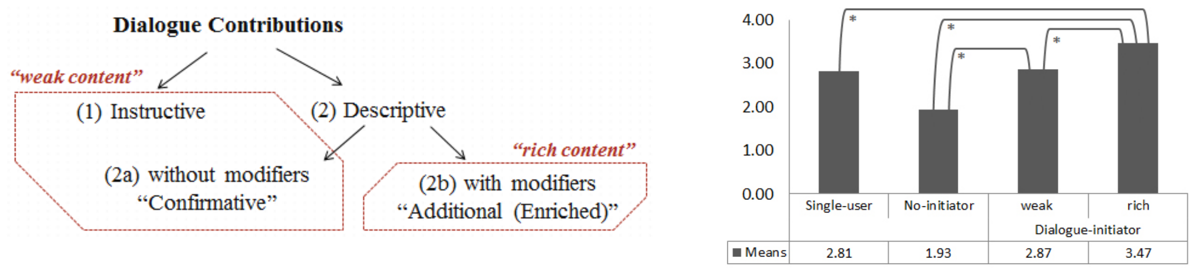
2.1 Utterances by the Verbal Assistant

Providing information in response to explorer´s question or statement has the potential to enhance E's comprehension of the graph at that particular moment of verbal assistance. In other words, E makes an inference by using the information provided by A's utterance and forms a potentially more correct representation of the graph or a graph segment by combining this information with his/her exploration. During the course of the dialogue between the haptic explorer and the verbal assistant, each of the explorer's utterances can affect the explorer's mental representation of the graph. In the present study, we focus on the communicative goal of the dialogues. The assistant's contributions to the dialogue can be classified as follows: (1) instructive (i.e. navigational, such as 'go downward from there'), or (2) descriptive. Descriptive utterances include, (2a) confirmative assistance (specifying exploration events or graph entities without using modifiers - such as 'there is a decrease'), and (2b) additional assistance (specifying properties of exploration events or graph entities using modifiers, such as 'there is a steep decrease'). Based on this scheme (see Figure 4a), we classified the dialogues into two major groups. Firstly, we identified *weak content* dialogues, which were less informative. These were the dialogues that contained assistance focused on (or restricted to) 'basic spatial properties' of the currently-explored region (i.e. the location or polarity of the graph segments). Secondly, we identified *rich content* dialogues, in which the verbal assistant also provided additional properties about the region explored (e.g., information about the steepness or length of the graph segments). Two coders performed the classification task. Interrater reliability between coders was calculated by Cohen's kappa. The results revealed a value of .80 that indicates substantial interrater agreement.

In addition to applying dialogue classification with respect to a communicative goal, which is investigated in this paper, the semantic annotation and classification by referring expression (namely representing each referring expression by its attribute-value pairs, cf. the attribute-set approach, Dale & Reiter, 1995) introduced advantages, as well, due to its fine-graininess and systematicity in the investigation of how the haptic explorers comprehended haptic landmarks and segments. The results which focuses on the semantic annotation method and the production of haptic ostensive expressions and on the perspective alignment during joint activity, were reported elsewhere (Alaçam et. al. 2014a; 2014b).

2.2 Analysis of Post-Exploration Sketches

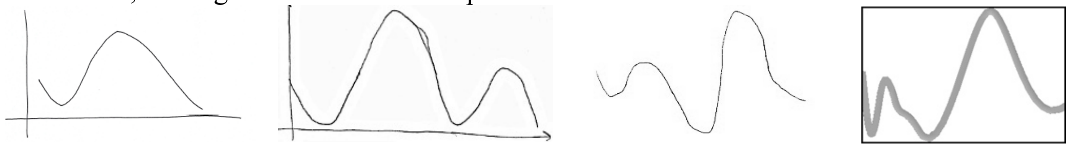
We analyzed participants' sketches in terms of their similarity to the stimulus-graphs. A statistical analysis using Kruskal-Wallis test revealed a significant difference between the ratings, $\chi 2(2, N=108)=23.3, p<.01$, among single-user protocols, no-initiator protocols, and dialogue-initiator protocols. Post-hoc testing of contrast using Mann-Whitney by using Bonferroni correction (so all effects are reported at a .0167 level of significance) showed that the sketches in no-initiator protocols ($M=1.93$, $SD=0.90$) received lower similarity scores both compared to the sketches in the single-user protocols ($M=2.81$, $SD=1.16$), $U=410.0$, $p<.01$ and the dialogue-initiator protocols ($M=3.17$, $SD=1.14$) $U=170.5$, $p<.01$, without a significant difference between the latter two. A further Mann-Whitney test (with Bonferroni correction) was conducted by taking into account the information content of the utterances. The results showed that the utterances that contained *rich content* resulted in higher similarity scores for the sketches (*M*=3.47, *SD*=.72) in the dialogue-initiator protocols than the sketches in the single-user protocols (*M*=2.81, *SD*=1.16), *U*=236.50, *p*<.05 and the other conditions, see Figure 4b.



(a) (b)

Figure 4: (a) Classification scheme for dialogue contribution and
(b) Ratings for sketches in five point Likert Scale (1: least similar and 5: most similar)

This indicated that the dialogues that contained more specific information (such as slight increase, biggest curve etc.) resulted in better sketch production as an indicator of more complete conceptualization of the event, see Figure 5 for sketch samples.



(a) (b) (c) (d)

Figure 5: Sketches after the protocols (a) without verbal assistance, (b) with *weak content* verbal assistance (c) with *rich content* verbal assistance and (d) Haptic graph-stimuli.

3 Discussion

In order to generate automatic verbal assistance with adequate content, it is necessary to identify the types and roles of individual utterances, as well as the structure of the dialogue content by using empirical studies. We conducted an experiment with two-conditions to investigate the contribution of verbal assistance in haptic graph exploration. The results of the first condition (single–user protocol without verbal assistance) and the second condition in two types of protocols (no-initiator protocols without assistance and dialogue-initiator protocols with assistance) were compared in terms of the analysis of sketches produced by the participants. High similarity ratings of the sketches were correlated with the richness of the content provided by verbal assistance. The sketches for dialogue-initiator protocols were significantly more similar to stimulus-graph compared to single-user and no-initiator protocols. The results also demonstrated considerable effect of verbal assistance content. In other words, the dialogues that contained modifiers (cf. *rich content*) were helpful to the explorer. Modifier presence made the assistance more elaborate; it helped the participant to notice the features of the event, which were currently explored (e.g. steepness of the curve and length, relation with another curve).

4 Conclusion

In this paper, we investigated the role of the haptic explorer as dialogue-initiator (or no-initiator) and the role of the content of the verbal assistance in a collaborative activity. Although haptic explorer and visual assistant share a common field of perception, their perception and comprehension processes differ significantly. The empirical results pointed out that haptic graph readers benefit from the verbal assistance to achieve more successful conceptualization of the events that are represented by graph lines. This is because the verbal assistant has a more complete mental representation of the graph (both global and local information on the graph) from the onset of the partner's haptic exploration. The results also revealed valuable insights about how the comprehension is affected by the provided language content. Combining these information with haptic exploration patterns before/during/after the explorer's help request will provide a concrete base to build up automatic detection of the need for verbal assistance. The detection of what a graph reader wants to know at a particular time during the course of exploration, by means of the analysis of his/her current position of exploration, previous exploration movements, and referring utterances (the referred locations and how these regions were referred) would yield a more effective design of (learning) environment for the graph reader compared to presenting all possible information to the graph reader at once. In addition, the information content and the need highlighted by the assistance request is another crucial topic that we addressed in this study. To sum up, our results indicated that taking initiative for requesting help and having adequate verbal assistance enriched by modifiers, rather than just confirmation of the basic spatial properties in response, seems a superb combination for a successful joint activity that inherently requires asymmetric dialogues between two users with different roles; haptic explorer and verbal assistant.

The experimental findings reported above we currently exploit in designing and realizing an automated verbal assistance system. Monitoring of the haptic explorations, in particular the recognition of exploration events, by the verbal assistant system, on the one hand, and selecting and providing that content which is necessary and sufficient for successful graph comprehension on the other hand, are two crucial processes that ground alignment and make the communication between haptic explorers and verbal assistants efficient and effective: A first prototype, called ObservingAssistant, analyzes the users' exploration movements based on rule-based methods, and triggers reactively canned text, which is realized by the MARY text-to-speech system (Schröder and Trouvain, 2003; Kerzel et al., 2014). The current step is to enhance this system by focusing on the *when to say*-matter so that the system would predict the need of assistance based on haptic explorer's exploration patterns, such as increase in the back-and-forth movements, the speed of the movement etc. (Acartürk et al., 2015).

Acknowledgments.

The study reported in this paper has been partially supported by DFG (German Science Foundation) in ITRG 1247 'Cross-modal Interaction in Natural and Artificial Cognitive Systems' (CINACS) and by METU BAP-08-11-2012-121 'The Study of Cognitive Processes in Multimodal Communication'.

References

Iyad Abu Doush, Enrico Pontelli, Tran C. Son, Dominic Simon, and Ou Ma. 2010. Multimodal presentation of two-dimensional charts: an investigation using Open Office XML and Microsoft Excel. *ACM Transactions on Accessible Computing*, 3(2), Article 8.

Cengiz Acartürk. 2010. Multimodal comprehension of graph-text constellations: An information processing perspective. *Doctoral dissertation, University of Hamburg*, Hamburg.

Cengiz Acartürk, Özge Alaçam, and Christopher Habel. 2014. Developing a verbal assistance system for line graph comprehension. *In A. Marcus (Ed.): Design, User Experience and Usability (DUXU/HCII 2014)*, Part II, LNCS 8518. (pp. 373–382). Berlin: Springer-Verlag.

Cengiz Acartürk, Özge Alaçam, and Christopher Habel. 2015. Haptic exploration patterns in virtual line-graph comprehension. *In A. Marcus (Ed.): Design, User Experience and Usability (DUXU/HCII 2015)*, Berlin: Springer-Verlag. (accepted)

Özge Alaçam, Christopher Habel, and Cengiz Acartürk. 2013a. Towards designing audio assistance for comprehending haptic graphs: A multimodal perspective. *In Constantine Stephanidis and Margherita Antona (Eds.), Universal Access in Human-Computer Interaction*. (UAHCI/HCII 2013), Part I, LNCS 8009. (pp. 409–418). Berlin: Springer-Verlag.

Özge Alaçam, Christopher Habel, and Cengiz Acartürk. 2013b. Investigation of haptic line-graph comprehension through co-production of gesture and language. *Tilburg Gesture Research Meeting*, Tilburg.

Özge Alaçam, Cengiz Acartürk, and Christopher Habel. 2014a. Referring expressions in eiscourse about haptic line graphs. *In Proceedings of the 18th Workshop on the Semantics and Pragmatics of Dialogue. SemDial 2014 – DialWatt.* Verena Rieser & Phillipe Muller (eds.). (pp. 7–16)

Özge Alaçam, Christopher Habel, and Cengiz Acartürk. 2014b. Perspective alignment during the course of verbally assisted haptic graph comprehension. *In the online Proceedings of the RefNet Workshop 2014 on Psychological and Computational Models of Reference Comprehension and Production*. Edinburgh, UK. 31 August 2014.

Domenic V. Cicchetti. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4), 284–290.

Herbert H. Clark and Susan E. Haviland. 1977. Comprehension and the given-new contract. *In Roy O. Freedle (Ed.), Discourse Production and Comprehension*. Hillsdale, NJ: Erlbaum.

Herbert H. Clark. 1996. *Using Language*. Cambridge: Cambridge University Press.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2), 233-263.

Simon Garrod and Martin J. Pickering. 2004. Why is conversation so easy? *Trends in Cognitive Sciences*, 8. (1), 8-11.

Christopher Habel and Cengiz Acartürk. 2007. On reciprocal improvement in multimodal generation: Co-reference by text and information graphics. *In Ielka van der Sluis, Mariet Theune, Ehud Reiter and Emiel Krahmer (Eds.), Proceedings of Workshop on Multimodal Output Generation*, 69-80.

Christopher Habel, Özge Alaçam, and Cengiz Acartürk. 2013. Verbally assisted comprehension of haptic line-graphs: referring expressions in a collaborative activity. *In Proceedings of the CogSci 2013 Workshop on Production of Referring Expressions*, Berlin.

Mary Hegarty. 2011. The cognitive science of visual-spatial displays: implications for design. *Topics in Cognitive Science*, 3, 446–474.

Matthias Kerzel, Özge Alaçam, Christopher Habel, and Cengiz Acartürk. 2014. Producing verbal descriptions for haptic line-graph explorations. *In Proceedings of the 18th Workshop on the Semantics and Pragmatics of Dialogue. SemDial 2014 – DialWatt.* Verena Rieser & Phillipe Muller (eds.). (pp. 205 – 207).

Jack M. Loomis and Roberta L. Klatzky. 2008. Functional equivalence of spatial representations from vision, touch, and hearing: Relevance for sensory substitution. *In John D. Rieser, Daniel H. Ashmead, Ford F. Ebner, & Anne L. Corn (eds). Blindness and brain plasticity in navigation and object perception*, (pp. 155-184). New York: Erlbaum.

Jack M. Loomis, Roberta L. Klatzky, and Susan J. Lederman. 1991. Similarity of tactual and visual picture recognition with limited field of view. *Perception*, 20, 167-177.

PRBO. Waterbird Census at Bolinas Lagoon, MarinCounty, CA. *Public report by Wetlands Ecology Division, Point Reyes Bird Observatory (PRBO) Conservation Science*. (2012) http://www.prbo.org/cms/366, retrieved on January 29, 2012.

Marc Schröder and Jürgen Trouvain. 2003. The German Text-to-Speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology, 6,* 365-377.

Barbara Tversky. 1999. What does drawing reveal about thinking?. *In John S. Gero, Barbara Tversky (eds.) Visual and spatial reasoning in design (VR99), Key Centre of Design Computing and Cognition*, University of Sydney.

Wai Yu and Stephen A. Brewster. 2003. Evaluation of multimodal graphs for blind people. *Journal of Universal Access in the Information Society*, 2, 105-124.

Haixia Zhao, Catherine Plaisant, Ben Schneiderman, and Jonathan Lazar. 2008. Data sonification for users with visual impairment: a case study with georeferenced data. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 15(1).

Facing Nadine's Speech.
Multimodal Annotation of Emotion in Later Life

Catherine Bolly
CNRS, UMR 7023 SFL and
Université catholique de Louvain
catherine.bolly@uclouvain.be

Anaïs Thomas
University of Paris Ouest Nanterre
anais2.thomas@gmail.com

Abstract

This pilot study aims at reconstructing the empathic profile of conversation participants from their interaction in real-world settings. It addresses the question of how verbal and nonverbal modes converge in conveying information about the emotional and attitudinal behavior in everyday communication. In particular, the empathic ability of older people is explored studying physiological patterning from nonverbal resources, in relation to emotions expressed through the face. In addition, the IRI psychometric test of empathy provides the participant's overall empathic profile. The data is taken from the CorpAGEst multimodal corpus, and focuses on the language of four healthy very old women who obtained a normal score on the MoCA cognitive test. Preliminary results indicate that, despite the highly idiosyncratic use of nonverbal resources, some inter- and intra-individual tendencies seem to emerge.

1 Introduction

The main objective of this paper is to investigate the extent to which we can (or cannot) reconstruct the emotional and attitudinal profile of conversation participants from their interaction in real-world settings. At the core of the study, the following question is addressed: "To what extent do the verbal and nonverbal modes converge in the information they convey about the emotional and attitudinal behavior of people in their everyday communication?" In particular, the empathic ability of very old healthy people is explored by analyzing their physiological patterning from nonverbal resources, in relation to emotions expressed through the face. In addition to the corpus-based approach, the recourse to the IRI psychometric test of empathy provides a more precise picture of the participant's overall empathic profile. Taking for granted the multidimensionality of empathy and the multimodality of emotions (Martin et al., 2006), the present study is a tentative effort to access the complexity of human beings' communication through the lens of complementary approaches to language in interaction.

2 Background

2.1 Pragmatic competence in later life

To date, only very little attention has been paid to the study of *pragmatic competence* – that is, the ability to use language resources in a contextually appropriate manner (Kasper and Rose, 2002) – of healthy older people from the angle of language production in a natural environment. Yet, the existence of pragmatic features specific to communication mode in the older people is recognized, which shows change in the interlocutors' behavior and increased (off-target) verbosity. On the one hand, it has been observed that speakers often adjust their way of speaking and gesturing to accommodate to the older people speech, and switch from their common way of speaking to a so-called "elderspeak" (Harwood, 2007). On the other hand, the Pragmatic Change Hypothesis (James et al., 1998) argues that the decrease in coherence – which goes together with an increase in amount of speech (*viz.* verbosity) in the older people – would result from a strategy to adapt their speech style according to communicative goals and social context.

K. Jokinen and M. Vels. 2015. Proceedings of The 2nd European and the 5th Nordic Symposium on Multimodal Communication.

2.2 Embodied emotion and empathic ability in aging

Emotions are grounded in the here-and-now experience of conversation participants. As such, they form part of our pragmatic competence: “Emotions are defined as short-term, biologically based **patterns** of perception, subjective experience, physiology, and action (or action tendencies) that constitute responses to specific physical and social problems posed by the environment” (Niedenthal et al., 2005: 22). It is worth stressing that the recognition of these emotional patterns is a complex process depending on the situational context, on the affective state, social and cultural identity of the participants (Russell et al., 2003: 334). What research in interpersonal pragmatics should therefore include in its scope goes far beyond the level of discourse, and must also address the embodied dimension of emotion as being part of the fuller context of interaction (Niedenthal, 2007). Facial expressions are particularly recognized as a major conveyance of both affective and cognitive stance, that is, of intersubjective evaluation, positioning, and alignment of language users in a situation of collaborative interaction (Englebretson, 2007). They may also have an emotion-regulating function (in the communicating person) and provoke empathic inferences (in the interlocutor).

Empathy is generally defined as the cognitive and affective ability to understand others’ emotions and points of view, as well as to be in-tune with their emotional states (Eisenberg et al., 2014). We will distinguish here between *empathy*, seen as the result of both affective and cognitive processes that are self- and other-oriented, and *sympathy*, defined as the “other-oriented desire for the other person to feel better” (Eisenberg and Fabes, 1990: 132). In the domain of aging and neuropsychology, results indicate that the healthy subjects’ advancing age may be accompanied by a loss of empathic ability (Bailey and Henry, 2008), liable to affect their ability to successfully engage in social interaction.

2.3 The CorpAGEst project: pragmatics, aging and language in use

The present paper is part of the CorpAGEst project (“A corpus-based multimodal approach to the pragmatic competence of the elderly”), which aims to establish the gestural and verbal profile of very old people in aging, looking at their pragmatic competence from a naturalistic perspective. The CorpAGEst assumption is that multimodal (inter)subjective markers of stance are highly relevant cues for the measurement of empathic ability of the older people: “Since evaluation is tied to affect, stance taking in the here-and-now of interaction serves to link affect to aspects of ideological systems and their expressions, including language, gesture, body practices, rhetoric, socialization, prior text, arts, aesthetic artifacts, and more” (Du Bois and Kärkkäinen, 2012: 437).

3 Objectives and research questions

This paper investigates the extent to which we can (or cannot) reconstruct the emotional and attitudinal profile of healthy very old people in their everyday interactions. The focus is on the synchronic and individual aspect of language competence in later life, without any longitudinal perspective or comparison between age groups at that stage. Hence, this pilot study has to be considered as a preliminary step for further investigation in a longer-term perspective within the framework of the project CorpAGEst: (i) the ongoing annotation of hand moves and body gestures, together with the functional analysis of pragmatic markers and gestures in the corpus, will strengthen the multimodal scope of the study; (ii) the longitudinal approach will allow for the detection of any individual change in the old persons’ way of speaking and gesturing with advancing age; (iii) the widening of the sample of study subjects will make the results more likely to be generalized, at least to some reasonable extent.

For the purpose of the present study, the following questions are addressed: “To what extent do the verbal and nonverbal modes converge in the information they convey about the emotional and attitudinal behavior of people in their everyday communication?” And, more precisely: “What can verbal and nonverbal emotional or attitudinal markers reveal about the empathic ability of the very old person?” Such markers may consist in verbal pragmatic markers of stance (e.g., *enfin* ‘well’; *tu sais* ‘you know’) or in nonverbal resources that have an expressive or interactive function (e.g. a wide opening of the eyes to indicate surprise; a gaze towards the interlocutor to maintain his or her attention).

4 Data

4.1 Study subjects and tasks

The CorpAGEst corpus (Bolly, 2013) is comprised of semi-directed, face-to-face conversations between an adult and a very old subject (75 y. old and more) living at home or in a residential home, which have been audio and video recorded. The corpus is two-fold: the cross-tasks corpus currently comprises 18 interviews (9 subjects; mean age: 85; duration: 16.8 hrs.); the longitudinal corpus (still in progress) will comprise interviews based on a shortened protocol from reminiscence tasks (see http://corpagest.org). Contextual independent variables are part of the corpus design, namely environment (private *vs.* residential home), the social tie between the participants (familiar *vs.* unknown interviewer) and the task type (focusing on past events *vs.* present-day life) (see Table 1). Metadata provide information about the interaction situation (e.g., date, place, quality of the recordings) and the participants (e.g., sex, education, profession, mother tongue, geographic origin, etc.).

Task Type	*Interview N°1* *(with a familiar person)*	*Interview N°2* *(with an unknown person)*
Task A: Focus on past events	Task 1A: Milestones in aging	Task 2A: Milestones in progress
Task B: Focus on present-day life	Task 1B: Self-perception of aging	Task 2B: Self-perception of every-day environment

Table 1. Tasks for the transversal corpus data collection

4.2 Clinical evaluation

Clinical evaluation scales were used to serve as a basis for methodological comparison and validation: the *Montreal Cognitive Assessment* test (MoCA, Nasreddine et al., 2005); and the French version of the *Interpersonal Reactivity Index* (F-IRI, Gilet et al., 2013). The IRI test takes the form of a questionnaire that takes into account four components of empathy, the first two being part of the cognitive dimension of empathy, and the last two being part of its affective dimension: (i) *Fantasy* is defined as "the tendency to imaginatively transpose oneself into fictional situations"; (ii) *Perspective-Taking* relates to "the tendency to spontaneously adopt the psychological view of others in everyday life"; (iii) *Empathic Concern* corresponds to "the tendency to experience feelings of sympathy or compassion for unfortunate others"; (iv) *Personal Distress* concerns "the tendency to experience distress or discomfort in response to extreme distress in others" (Davis, 1994: 55-57). Among the nine old people from the CorpAGEst corpus, only the four who obtained a normal score at the cognitive test (equal or more than 26/30) were selected for the present study (n: 4; sex: F; mean age: 80 – see Table 2).

Recordings	*hh:mm:ss*	*Speaker ID Code*	*Pseudo*	*Age*	*Birth*	*Sex*	*Education (n years)*	*Cognition (MoCA)*	*Empathy (F-IRI)*
ageBN1r-1 ageBN1r-2	1:01:14 0:49:02	ageBN1	Nadine	75	1938	F	12	29/30	64 %
ageLL1r-1 ageLL1r-2	1:13:41 1:14:25	ageLL1	Louise	79	1933	F	12	26/30	66 %
ageBM1r-1 ageBM1r-2	0:59:02 0:50:36	ageBM1	Anne-Marie	82	1932	F	12	28/30	61 %
ageDA1r-1 ageDA1r-2	0:59:07 0:52:41	ageDA1	Albertine	84	1929	F	14	29/30	61 %

Table 2. Main characteristics of the study subjects by chronological age (transversal corpus)

5 Method

Taking for granted the multidimensionality of empathy and the multimodality of emotions, the perspective adopted combines notions and methods inherited from various disciplines. About 1 hour of video data was fully annotated on the basis of facial physiological parameters (section 5.2) and emotional states (section 5.3). In addition, the data were partly analyzed in terms of multimodal relationship with speech (section 5.4). All annotations were done by one investigator and partly crosschecked by the other one, mainly during the learning phase, in order to develop, improve and stabilize the coding scheme. The second investigator also served as control for uncertain and ambiguous cases.

5.1 Multimodal approach

This pilot study corresponds to the very first step of the annotation procedure within the framework of the CorpAGEst project, which aims *in fine* at a better understanding of the way in which the verbal and gestural dimensions interact to make sense in real-world settings. Starting with mono-modal analyses (gesture *vs.* speech) and focusing on one group of articulators at a time within each modality (*viz.* face, gaze, head, shoulders, torso, hands, legs, and feet), the annotation procedure next moves to multimodal analyses. Consequently, the present study mainly concentrates on facial displays, gaze, and emotions perceived from the face. A first insight into the interaction of physiological and emotional parameters with contextual and discursive cues is given at the end of the paper (see section 6.4).

The text, sound and video data were aligned using the *ELAN* software (Wittenburg et al., 2006). The multi-level annotation of the audio and video samples (3*5 min. per interview) was performed as follows: (i) annotation of the physiological parameters for the face; (ii) annotation of emotions expressed through the face (no recourse to the sound signal); (iii) annotation of the relation between the tagged emotion and the contextual information (taking into account gestures and linguistic information).

5.2 Facial expressions and gaze

In line with form-based approaches to gesture (Müller et al., 2013) and mainly inspired from the *MUMIN* project (Allwood et al., 2007), the ELAN annotation scheme dedicated to the physiological description of facial expressions is comprised of 7 parameters (see Table 3 below).

Articulator	*Variable*	*Values / Labels*
Eyebrows	Form	Frowning, Raising, Other
Eyes	Form	Exaggerated Opening, Closing-Both, Closing-One, Closing-Repeated, Other
Gaze	Direction	Forward-Front, Forward-Right, Forward-Left, Up-Front, Up-Right, Up-Left, Down-Front, Down-Right, Down-Left, Other
	Target	Addressee, Other participant, Vague, Object, Body part, Camera, Other
Mouth	Openness	Open
	Lips' corners	Up, Down, Other
	Lips' shape	Protruded, Retracted, Other

Table 3. Articulators and physiological parameters for facial expressions

Facial displays (including gaze) were identified according to their location in the face (eyebrows, eyes, gaze, mouth) and then annotated in terms of physiological features (e.g., *closed-both* for the eyes, *corners up* or *retracted* for the lips). The annotation was made independently of the sound signal to avoid any interpretive bias in the semiotics of gesture at this stage in the analysis. Movements were identified according to the following principle: the left boundary of each annotation – that is, the beginning of the move – has been assigned to the first frame that corresponds to a visible change in the face, mostly on a blurred image (e.g., when the eyes begin to close, not when they are completely closed); in the same manner, the right boundary of facial expressions – that is, the end of the move – has been put on the frame corresponding to the absence of any visible change, mostly a fixed image (e.g., when the eyes are fully open again). It is of great importance here to stress some methodological issues. First, although the beginnings of facial moves were quite easy to detect, many of them were disappearing with a fading effect. In those cases, the right boundary has been put on the frame corresponding to the recovered neutral position. Other physiological features (e.g., wrinkling the forehead while eyebrow raising) were also used as support to detect the very end of such fading moves. Secondly, openness of the mouth and moves from the lips were only taken into account when not accompanying speech production. For these exceptions only, the sound signal was activated to distinguish between the two possibilities (with or without speech). Thirdly, we chose to annotate gaze all along the samples (with the obvious exception of closing eyes) rather than delimiting so-called "gaze-units", in order to highlight transitions in gaze direction and nature of the target.

5.3 Attribution of emotions

In the present study, emotions were annotated by looking at facial expressions. We followed the Plutchik's multidimensional model (1980) based on eight primary emotional dimensions, which are organized in polarity dyads (e.g., *ecstasy* as opposed to *grief*), declined into several combinations (e.g., *optimism* resulting from the combination of *anticipation* and *joy*), and nuanced according to their degree of intensity in tryads (e.g., *acceptance – trust – admiration*, from weakest to strongest) (Figure 1).

Anchored in biological and neurobiological grounds, the model includes 32 emotion labels that are said to be both discrete and gradual, insofar as intensity and polarity are considered to be central criteria for distinguishing between the emotions at stake. Three more recurrent emotions (*nervousness*, *disappointment* and *nostalgia*) emerged from the video data analysis and were therefore *a posteriori* added to the model. The closed list gives the advantage of providing a rich set of labels, which seems to be more accurate for the study of naturalistic data than other models strictly based on the 6 Ekman's basic emotions (Ekman, 1992). Moreover, all emotions in the Plutchik's model can be reduced to intrinsic positive and negative values. This is in line with the view that the broad bipolar dimensions of emotions (positive *vs.* negative valence) are the best (if not the only), most efficient way to distinguish between emotions from the face (Russell et al., 2003: 334).

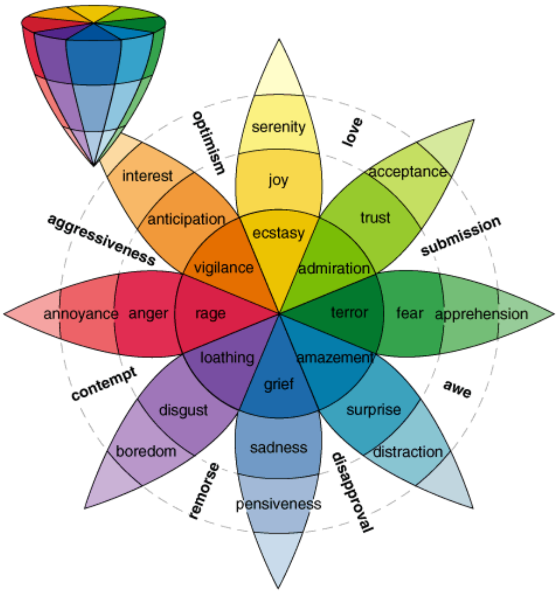


Figure 1. Plutchik's circumplex and wheel of emotions (reproduced from www.6seconds.org)

It is worth noting that the annotation of emotions was done without any access to the previously annotated physiological parameters. The boundaries of the emotion tags were determined on the basis of a holistic perception of the emotion expressed through the face, independently of the existing segmentation at the physiological level. Using emotion tags as a filter in the next step allowed for a bottom-up, relatively objective approach to the data. As a consequence, the boundaries of emotion tags do not correspond to physiological tags and emotions are mostly comprised of several physiological tags (e.g., one single emotion may include successive eye-closing moves and changes in gaze direction).

5.4 Contextual disambiguation of emotions

Emotional and attitudinal expression can be transmitted through multiple modes of communication (e.g., face, voice, words, and gestures) and may therefore result in complementary, redundant or even conflicting information (Gendron et al., 2012). According to that view, the semantic relation between emotions perceived from the face and their context of appearance (including the whole body, the linguistic and extralinguistic context) has been recognized to be either: (i) redundant: a similar emotion (even though not necessarily synchronous) is expressed from the face and from the linguistic context; (ii) complementary: the facial emotion is compatible with and adds some value to the linguistic information conveyed (e.g., modalization, emphasis, hedge, specification, elaboration, etc.); (iii) contradictory: the facial emotion is not compatible with the linguistic information conveyed; (iv) independent: there is no relation between the two modes, which fulfill their proper function in the language interaction; (v) accordant: facial emotions are in accordance with information transmitted by the extralinguistic context at large (e.g., as reaction to external stimuli such as noises). This classification is mainly inspired by Colletta et al. (2009), following the pioneering classification in the field by Poggi and Magno Caldognetto (1996). These relationships were attributed to emotion tags in the data of one speaker only, namely Nadine (3 samples; 16 min. 24 sec.), with an additional focus on the relation between emotions and discourse markers in one of the three samples.

Discourse markers can briefly be defined as short linguistic items, which have no or little referential meaning and are not syntactically connected to their host clause. They serve pragmatic purposes by guiding the addressee in the decoding of the information conveyed: they "can connect to the speaker or addressee, provide information about the attitude of the communicator, introduce assumptions, or provide information about the context of interpretation" (Brinton, 2008: 5). Once transcribed and aligned to the sound signal, discourse markers were semi-automatically retrieved from speech and aligned to the video signal in the ELAN files.

6 Results

Preliminary results from the study indicate that, despite the highly idiosyncratic use of nonverbal resources, some inter- and intra-individual tendencies emerge.

6.1 Empathic ability

Results from the empathy test (F-IRI, see above) suggest that the healthy subjects obtain a relatively homogeneous global score of empathy (from 61% to 66%). This seems to posit that their empathic ability is relatively well preserved. Yet, a highly significant variability has been observed in the individual profiles with respect to the four subscales of empathy ($X^2 = 30.94$; df = 9; $p < 0.001$) (Table 4).

Speaker ID Code	*Pseudo*	*Fantasy (F)*	*Perspective-Taking (PT)*	*Empathic Concern (EC)*	*Personal Distress (PD)*	*F-IRI Score (%)*
ageBN1	Nadine	60 **[+1.93]**	57 **[-2.05]**	86 [-0.08]	51 [+0.77]	64
ageLL1	Louise	51 [+0.31]	80 [+0.23]	91 [+0.05]	43 [-0.69]	66
ageBM1	Anne-Marie	29 **[-2.4]**	94 **[+2.59]**	91 [+0.8]	31 **[-1.98]**	61
ageDA1	Albertine	46 [+0.1]	66 [-0.75]	77 [-0.77]	57 **[+1.9]**	61

Table 4. Subscales of empathy in percent [with standardized residuals]

These results partly confirm Gilet et al.'s findings (2013), who stressed *Fantasy* as being the most age-sensitive subscale. As a matter of fact, our data range from 60% in the youngest (Nadine, 75 y. old) to 46% in the oldest (Albertine, 85 y. old) for *Fantasy*, with the lowest score in Anne-Marie (29%). Moreover, *Empathic Concern* – which has been evidenced in several works to be genre-specific – shows a very high score in every participant (from 77% to 91%). Looking at intra-individual differences, even more striking results were found in Anne-Marie's profile (82 y. old) who seems to be more likely to experiment feelings of sympathy and compassion (EC: 91%), as well as to adopt the point of view of others (PT: 94%), than to transpose herself into fictional characters (F: 29%) or to feel concerned by stressful situations (PD: 31%). In addition, participants highly differ in their ability to cognitively adopt the point of view of somebody else (PT), ranging from the lowest ability in Nadine (57%) to the highest in Anne-Marie (94%).

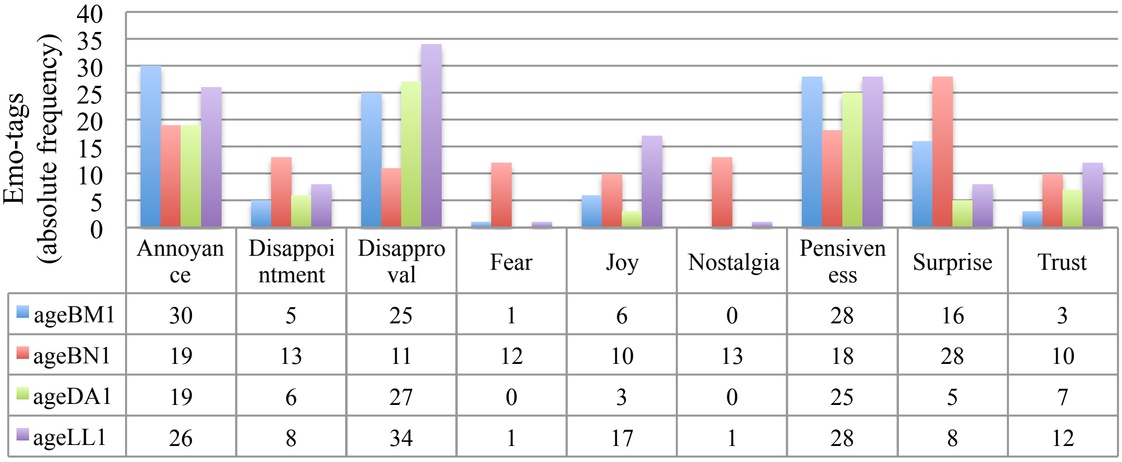
6.2 Emotional variety and richness

From the 581 emotions identified in the corpus data (including 8 undetermined emotions labeled as "Other"), it appears that the four subjects slightly differ with respect to their facial emotional richness, measured in terms of types of expressed emotions within the samples (Type/Token Ratio). Only 9 categories of emotion for a total of 108 annotated emotions were counted in Albertine's speech samples [TTR = 0.08], while a wider emotional panel of facial expressions was observed in Louise's speech (20 types for a total of 161 emotions tagged in the samples [TTR = 0.124]). Anne-Marie and Nadine obtained intermediate scores, with respectively 14 types for 143 emotions tagged [TTR = 0.097] and 19 types for 169 emotions tagged [TTR = 0.112]. Even though these results were not statistically significant, we would like to highlight the fact that only 23 types among the 35 emotion tags available in the Template were identified as such, from a wider variety in Louise and Nadine (with more than 50%

of the tags used), to much less diversity in Anne-Marie (40% of the available tags used) and Albertine (only 26% of the available tags used). Interestingly, some emotions were quite infrequent in the data (e.g., only 1 to 3 cases of *amazement*, *boredom*, *contempt*, *ecstasy*, and *nervousness*), while others seem to be specific to one single participant. For instance, *fear* and *nostalgia* were mostly recognized from Nadine's face (with 12 and 13 out of 14 cases, respectively), while *attention* is mainly attributed to Albertine (with 9 out of 11 cases) and *disgust* to Anne-Marie (7 out of 9 cases). The most frequent emotions will be analyzed in the next section, by crossing emotional tags and physiological features.

6.3 Physiological patterning

Looking at physiological patterning from face and gaze expressions with regard to frequent emotions in the corpus, no clear physiological pattern could be considered specific to one emotion or another, neither to one speaker or another. However, some regularity was noticed from a closer investigation of the nine most frequent emotions (equal to or more than 10 occ. in the speech of at least one participant): *pensiveness* (99 occ.), *disapproval* (97 occ.), *annoyance* (94 occ.), *surprise* (57 occ.), *joy* (36 occ.), *trust* (32 occ.), *disappointment* (32 occ.), *fear* (14 occ.) and *nostalgia* (14 occ.) (see Figure 2).



	Annoyance	Disappointment	Disapproval	Fear	Joy	Nostalgia	Pensiveness	Surprise	Trust
🟦 ageBM1	30	5	25	1	6	0	28	16	3
🟥 ageBN1	19	13	11	12	10	13	18	28	10
🟩 ageDA1	19	6	27	0	3	0	25	5	7
🟪 ageLL1	26	8	34	1	17	1	28	8	12

Figure 2. Distribution of the most frequent emotions across participants

For instance, results for the *annoyance* emotion showed some variance between individuals when comparing cases of frowning and raising eyebrows. Notably, the recognition of this emotion in Anne-Marie' face [ageBM1] is mainly correlated with eyebrow frowning (55% of her eyebrow moves, with a positive standard residual of 3.21 for 15 cases), whereas the other three participants preferably raise their eyebrows (with 94% of eyebrows' raising moves in Nadine [ageBN1], 88% in Albertine [ageDA1], and 60% in Louise [ageLL1]). When looking at eyes' moves, a difference is also observed between participants, showing (i) much complex and repeated closing of the eyes in Louise (in one out of two cases with a positive standard residual of 1.51), (ii) a lower proportion of eyes' moves linked to *annoyance* in Anne-Marie's face, noticeable through the absence of any eyes' move in 44% of the cases (with a standard residual of 2.68), and (iii) a specific use of exaggerated opening of the eyes in Nadine (with a standard residual of 2.28), by comparison with the other participants and the other types of eyes' moves. In spite of these individual differences, overall results showed a well-balanced use of single and double closing of the eyes corresponding, respectively, to 20% and 16% of the 91 *annoyance* tags. To sum up, we can say that the expression of *annoyance* tends to be more idiosyncratic in Anne-Marie's face, as she mostly frowns without any other characteristics in closing or opening the eyes, by contrast to the other three participants who mainly raise their eyebrows either with many more eye-closings (cf. Louise) or with exaggerated opening of the eyes (cf. Nadine).

Again, we could reasonably expect a strong correlation between exaggerated openings of the eyes and eyebrow raisings, as a means to express *surprise*. But, even though this combination is relatively frequent to express *surprise* (23% of the cases), it is above all true for *fear* (57% of the cases). In a much less remarkable degree, it also applies to *disappointment* (9% of the cases), *annoyance* (7% of the cases), *nostalgia* (7% of the cases), *joy* (3% of the cases), *disapproval* (2% of the cases), and *pensiveness* (2% of the cases). Going a step further, it appeared that this combination of physiological parameters was specific to one single participant: among the 37 cases of "exaggerated-opening/eyebrow-raising" pattern, 32 were identified in Nadine's face, as conveying one of the above-mentioned emotional states. More in-depth and exhaustive analyses, which would embrace all the physiological parameters (including the whole body) and examine the way they combine in every participants, would undoubtedly help distinguish between individual and shared uses of gestural patterns with regard to the emotional and attitudinal states of people interacting in real-world settings.

6.4 Multimodality and the speech-gesture interface

A closer look at Nadine's speech, with a focus on the first 5 min. of interaction, allows for a better understanding of the relationship between (frequent and infrequent) perceived emotions, the facial expressions, and the linguistic context. Results give a first insight on the role of discourse markers in the multimodal expression of emotional states. From a context-sensitive angle, results showed that emotions usually appeared to be congruent with the contextual and linguistic information. Adding some compatible semantic or pragmatic value to the meaning conveyed in language use, facial emotions were mostly identified as being *complementary* (42 out of 74 emotions, see the red bars in Figure 3). Yet, facial emotions sometimes contradict the information conveyed by the context (in 14 out of the 74 cases). For instance, most of the time, the annotation of *joy* does not mirror the information expressed by Nadine and should be disambiguated thanks to the linguistic context.

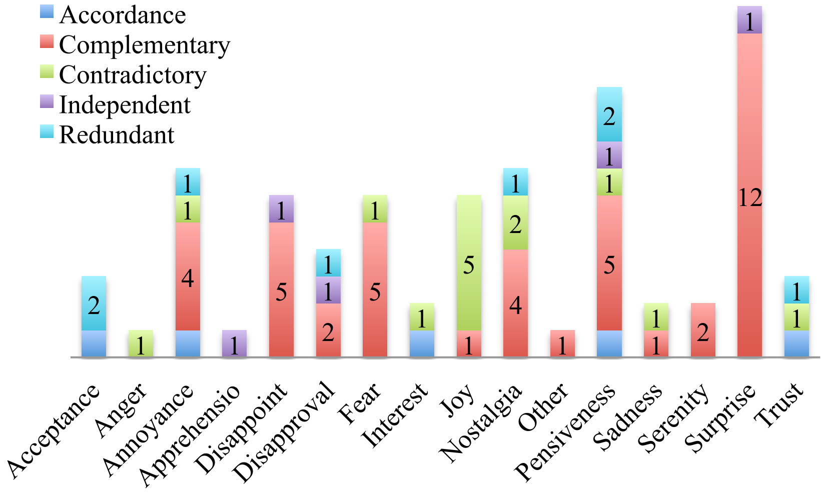


Figure 3. Semantic relationships between emotions perceived from the face and the contextual meaning (sample n°1 of Nadine's video data)

As the screen capture from the ELAN annotation file shows (Figure 4), the emotion perceived – independently of any contextual cues at the first step of analysis – is *joy*, but what the speaker is actually saying concerns a painful episode in her childhood (*c'était **un peu** je,une **quoi hein** j'ai été **un peu** malheureuse **là*** 'I was **a little bit** too young **well you know** I have been **quite** unhappy **there**').

The further question is to explore how far the global emotional and attitudinal state can be inferred from speech and from nonverbal resources. Our hypothesis is that Nadine is smiling here to mitigate the pain she is remembering (another interpretation would be that she is smiling because of being embarrassed to talk about an intimate and painful experience). Facial displays would then be redundant with the modal marker *un peu* 'a bit', notably repeated once. Concerning the function of discourse markers in the synchronous co-text of the emotion tag (*viz. quoi* 'well', *hein* 'he', and *là* 'there'), their intersubjective function (Kärkkäinen, 2006) could be seen as stressing the need to share the speaker's painful experience with the interlocutor or as reassuring that full attention is paid to what she says. As Russell et al. emphasized (2003: 242), smiles can be spontaneous "reliable signs of positive feelings toward a specific receiver" (expressive function), but they can also be produced in a controlled manner as "volitional smiles" which seek appeasement or help in the addressee (interactive function).

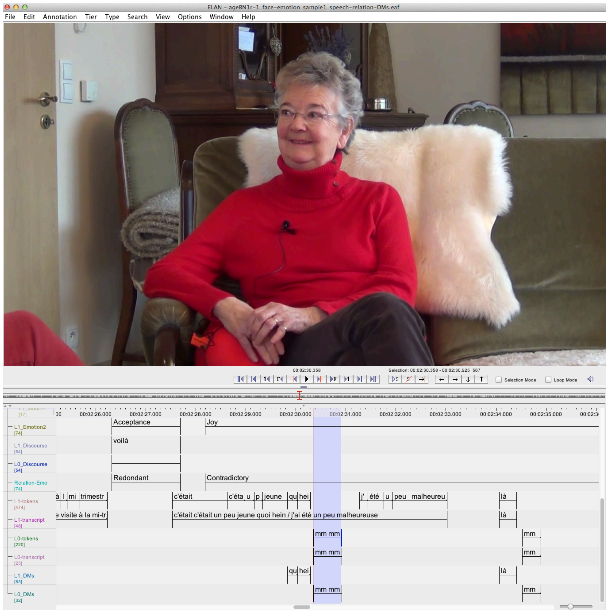


Figure 4. Example of contradiction in the emotion recognized from the face and the linguistic context.

7 Conclusion

Nonverbal language resources are recognized as a major channel of emotional expressivity and interactivity in the communicating person. But, due to their ambiguous and complex structure, emotional states are extremely challengeable to detect, even more in the natural context of language production (Douglas-Cowie et al., 2003: 36-38). Starting from the annotation of facial expressions and emotion recognition in authentic video data, it has been evidenced (however quite unsurprisingly) that the visual mode, if taken alone, was not sufficient to understand what kind of information the speaker is actually transmitting to the interlocutor. Rather, as the linguistic level of communication often needs to be contextualized, the nonverbal level of communication also needs more “words” in order to be interpreted in accordance with the speaker’s intention.

This study represents the very fist step of the CorpAGEst research project, which aims at developing a multimodal model for the annotation of pragmatic functions in speech and gesture, as a means to detect any change with advancing age in the pragmatic competence of very old people. The study has given a first insight into what we can infer from emotional and attitudinal expressions of very old healthy people by means of “naturalistic” corpus data. Even though providing only part of the big picture, the approach, we assume, allowed for a better understanding of the way older people show and express their emotions in real language use. It seems obvious that the pragmatic part of language communication is not of little interest in the field of aging research, and would need further investigation moving from experiments in the laboratory towards empirical studies “into the wild”.

Acknowledgment

The research leading to these results has received funding from the European Union Seventh Framework Programme ([FP7/2007-2013]) under grant agreement n° [PIEF-GA-2012-328282]. The authors are grateful to the two anonymous reviewers and L. Crible for their useful review of the paper. C. T. Bolly would like to warmly thank all the participants in the CorpAGEst project, who will certainly recognize themselves. She also thanks A. Sáfár, A. Thomas and D. Belin, for their help in the annotation procedure, and D. Boutet, A. Lacheret and L. Degand, for numerous inspiring discussions and their confidence in the project.

References

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 phenomena. *Language Resources and Evaluation*, 41(3-4), 273-287.

Bailey, Phoebe E. and Julie D. Henry. 2008. Growing less empathic with age: Disinhibition of the self-perspective. *Journal of Gerontology: Psychological Sciences*, 63B(4): 219–226.

Bolly, Catherine T. 2013. *CorpAGEst. Multimodal corpus for the elderly's language*. Louvain-la-Neuve, Paris: Université catholique de Louvain (Valibel - Discours et Variation) and CNRS (UMR 7023 Structures Formelles du Langage), <http://corpagest.org>.

Brinton, Laurel J. 2008. *The comment clause in English: Syntactic origins and pragmatic development*. Cambridge : Cambrige University Press.

Colletta, Jean-Marc, Kunene, Ramona N., Venouil, Aurélie, Kaufmann, Virginie and Simon, Jean-Pascal. 2009. Multi-track annotation of child language and gestures. In Michael Kipp, Jean-Claude Martin, Patrizia Paggio and Dirk Heylen (eds.), *Multimodal corpora* (*Lecture Notes in Computer Science* 5509). Berlin, Heidelberg: Springer, 54–72.

Davis, Mark H. 1994. *Empathy: A Social Psychological Approach*. Boulder: Westview Press.

Douglas-Cowie, Ellen, Campbell, Nick, Cowie, Roddy, and Peter Roach. 2003. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40: 33–60.

Du Bois, John W. and Elise Kärkkäinen 2012. Taking a stance on emotion: Affect, sequence, and intersubjectivity in dialogic interaction. *Text & Talk*, 32(4): 433–451.

Eisenberg, Nancy and Richard A. Fabes. 1990. Empathy: Conceptualization, measurement, and relation to prosocial behavior. *Motivation and Emotion*, 14(2): 131–149.

Eisenberg, Nancy, Shea, Cindy L., Carlo, Gustavo, and George P. Knight 2014. Empathy-related responding and cognition: A "chicken and the egg" dilemma. *Handbook of moral behavior and development: Research*, 2: 63–88.

Ekman, Paul. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3/4): 169–200.

Englebretson, Robert (ed.). 2007. *Stancetaking in discourse: Subjectivity, evaluation, interaction* (Pragmatics & Beyond New Series 164). Amsterdam, Philadelphia: John Benjamins Publishing.

Gendron, Maria, Lindquist, Kristen, Barsalou, Lawrence, and Lisa F. Barrett. 2012. Emotion words shape emotion percepts. *Emotion*, 12(2): 314–325.

Gilet, Anne-Laure, Mella, Nathalie, Studer, Joseph, Grühn, Daniel and Gisela Labouvie-Vief. 2013. Assessing dispositional empathy in adults: A French validation of the Interpersonal Reactivity Index (IRI). *Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement*, 45(1): 42–48.

Harwood, Jake. 2007. *Understanding communication and aging: Developing knowledge and awareness*. Thousand Oaks: Sage.

James, Lori E., Burke, Deborah M., Austin, Ayda and Erika Hulme. 1998. Production and perception of "verbosity" in younger and older adults. *Psychology and Aging*, 13(3): 353–367.

Kärkkäinen, Elise. 2006. Stance taking in conversation: From subjectivity to intersubjectivity. *Text & Talk*, 26(6): 699–731.

Kasper, Gabriele and Kenneth R. Rose. 2002. *Pragmatic development in a second language*. Mahwah: Blackwell (Also *Language Learning : Supplement 1*, 52).

Martin, Jean-Claude C., Niewiadomski, Radoslaw, Devillers, Laurence, Buisine, Stéphanie and Catherine Pelachaud. 2006. Multimodal complex emotions: Gesture expressivity and blended facial expressions. *International Journal of Humanoid Robotics*, 3(3): 269–291.

Müller, Cornelia, Jana Bressem and Silva H. Ladewig. 2013. Towards a grammar of gestures: A form-based view. In Cornelia Müller, Alan Cienki, Ellen Fricke, Silva H. Ladewig, David McNeill and Sedinha Teßendorf (eds.), *Body – language – communication: An international handbook on multimodality in human interaction*. *Vol. 1*. Berlin, Boston: Mouton de Gruyter, 707–733.

Nasreddine, Ziad S., Phillips, Natalie, Bédirian, Valérie, Charbonneau, Simon, Whitehead, Victor, Collin, Isabelle, Cummings, Jeffrey L. and Howard Chertkow. 2005. The Montreal Cognitive Assessment (MoCA): A brief screening tool for Mild Cognitive Impairment. *Journal of the American Geriatrics Society*, 53: 695–699, <http://www.mocatest.org/default.asp>.

Niedenthal, Paula M. 2007. Embodying emotion. *Science*, 316(5827): 1002–1005.

Niedenthal, Paula M., Barsalou, Lawrence W., Ric, François and Silvia Krauth-Gruber. 2005. Embodiment in the acquisition and use of emotion knowledge. In Lisa Feldman Barrett, Paula M. Niedenthal and Piotr Winkielman (eds.), *Emotion and consciousness*. New York: The Guilford Press, 21–50.

Plutchik, Robert. 1980. *Theories of Emotion. Vol. 1: Emotion: Theory, Research, and Experience*. New York: Academic.

Poggi, Isabella and Magno Caldognetto, Emanuela. 1996. A score for the analysis of gestures in multimodal communication. In L. Messing (ed.), *Proceedings of the Workshop on the Integration of Gesture and Language in Speech*, *Applied Science and Engineering Laboratories*, Newark and Wilmington, Delaware: 235–244.

Russell, James A., Jo-Anne Bachorowski and José-Miguel Fernández-Dols. 2003. Facial and vocal expressions of emotion. *Annual review of psychology*, 54(1): 329–349.

Wittenburg, Peter, Brugman, Hennie, Russel, Albert, Klassmann, Alex and Hans Sloetjes. 2006. Elan: A professional framework for multimodality research. In *Proceedings of LREC 2006*, Fifth International Conference on Language Resources and Evaluation. Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands, <http://tla.mpi.nl/tools/tla-tools/elan/>.

Teachers and Learners Constructing Meaning for Vocabulary Items in a Foreign Language Classroom

Eva Ingerpuu-Rümmel
University of Tartu
Institute of Social Studies
Estonia
eva.ingerpuu-rummel@ut.ee

Abstract

This paper presents a study of how teachers and learners use two semiotic resources – verbal expression and gestures – to construct meaning for words and expressions in French and Estonian language classrooms. For the purposes of the research, university seminars where Estonian and French were taught as foreign languages were videotaped. A micro-level multimodal discourse analysis of the videos was then conducted. The results show that each semiotic resource has its specific functions in the process of constructing meaning.

Keywords: meaning construction, gestures, foreign language learning, multimodal discourse analysis

1 Introduction

The paper presents a study of how teachers and learners use gestures to construct meaning for words and expressions in a foreign language classroom. Learners acquire elements of language and culture that are first partially or completely incomprehensible to them. A new word or expression ('word' hereafter) may be explained by the teacher alone or in collaboration with learners. The teacher and/or the learner(s) may introduce new words and expressions in a multimodal manner: verbal expression (words and grammar), vocal expression (e.g., stress, volume, pausing), gestures (body movement and position) as well as aids (e.g., texts, figures, drawings, videos) are used. Speech (verbal expression and vocal expression) and gestures collaborate in the co-construction of discourse (Kendon, 2004).

The explanation of words as a multimodal phenomenon and the role of gestures in this have not been extensively researched in the context of foreign language learning. Anne Lazaraton (2004) illustrates how a teacher explains words in English as a second language classroom. She underlines the role that gestures can have in how a teacher explains words. The present paper shows how the cooperation of semiotic resources (verbal expression and gestures) is important both in teachers' and learners' explanations. For the purposes of the research, university seminars where Estonian and French were taught as foreign languages were videotaped, and a micro-level multimodal discourse analysis of the videos was conducted. Communication in the seminars was entirely held in the target languages.

The analysis indicates that the explanation created for a word is not a definition that is clearly formulated as a sentence. The work involved in constructing meaning may be distributed between different semiotic resources. Teachers and learners may use two semiotic resources – verbal expression and gestures – together. The results show how each semiotic resource has its specific functions and how pieces of information are organized into an explanation through engaging several semiotic resources. For example, gestures provide information which is not included in the verbal expression and vice versa. The paper introduces examples of how teachers and learners use gestures and verbal expressions in collaboration when explaining words.

K. Jokinen and M. Vels. 2015. Proceedings of The 2nd European and the 5th Nordic Symposium on Multimodal Communication.

2 Background

The study involves multimodal discourse analysis and the approach of the study is based on authors whose areas of research are very different. This section presents some authors who are interested in gestures (Gullberg, 1998; Kendon, 2004), whose focus is mainly on body movements in communication; authors who are interested in classroom interaction (Hall, 2009; Mondada and Doehler, 2004; Shepherd, 2010); researchers who are interested in language use by people who have limited ability to express themselves verbally (Goodwin, 1995; Rummo and Tenjes, 2011; Jokinen et al., 2013); and researchers who classify themselves as engaging in multimodal discourse analysis (Drissi, 2011; Kress et al., 2001a; Kress and Leeuwen, 2001b; Lim Fei, 2011; O'Halloran, 2011).

Classroom interaction usually has a specific topic and specific goals. Teacher and learners meet in predetermined room for predefined time period. In order to better understand classroom interaction, it is useful to know which patterns and norms can be found in communication. The manner in which the participants' specific patterns of behaviour affect language learning has been explored through the analysis of classroom interaction in terms of turn taking and sequences of verbal expression (e.g., Lerner, 1995; Hall, 2009). Mondada and Doehler (2004) have analysed French lessons and show how language learning occurs in group communication and how tasks are (re)organized in cooperation during the class.

Even though the first purpose of foreign language learning is to learn to use accurately verbal and vocal semiotic resources in a target language, gestures are intrinsically part of classroom interaction. In order to understand body movement and positions, one needs a working knowledge of Adam Kendon's (2004) discussion of the classification of gestures by various authors. This will help understand body movements and positions from their physical performance to the construction of meaning. Kendon (2004: 104–106) discusses multiple continua of gestures proposed by different authors (e.g. Kendon, 1988; Gullberg, 1998; McNeill, 2000). In conclusion, he assumes that many authors agree that gestures can be pointing, depictive or enactive and "displaying aspects of a logical structure of a speaker's discourse" (Kendon, 2004: 107).

Kendon (2004: 80–82) also presents different possibilities concerning functions of gestures on the basis of previous researches on gestures. He considers that on one side, gestures may facilitate verbal expression and the thought processes, and that on the other side, gestures may have communicative purpose – they provide another person with information, for example, about their ideas and intentions. Researchers have taken different approaches to investigating the use of body movements and positions as well as the use of space and objects. For example, Shepherd (2010) has explored classroom discourse, focusing on the use of a specific movement – the raising of one's hand with the intention to be given the word. The use of body movement in language learning has been semi-experimentally researched by Gullberg (1998). Gullberg (1998) describes research in which people who were learning Swedish and French as foreign languages had to retell the story of a cartoon both in their native and in the foreign language. Gullberg (1998) provides an overview of the use of different types of gestures used in telling the story in different languages and notes that gestures may help overcome difficulties with verbal expression.

The closest research to the present one is Lazaraton (2004) in which teachers' gestures in explaining words are analysed. Lazaraton (2004) focuses on situations in which explaining words is not planned. She presents a table that shows that in 14 situations out of 18, the teacher used non-verbal means (hand gestures in 12 instances and the whole body in two) (Lazaraton, 2004: 94). She concludes that the use of hand gestures in explaining words is a very important tool for teachers. At the same time, she admits that her research does not indicate why the teacher uses gesture – is it because they cannot find the right words to express the meaning or because they wish to make the meaning clearer to learners, or both (Lazaraton, 2004: 108–109).

In the context of language learning, Gullberg (1998) and Lazaraton (2004) discuss the compensatory use gestures can have in the case of difficulties with verbal expression. Thus, researchers of language learning are also interested in people whose ability to speak is limited. Goodwin's article (1995) shows us how meaning is created in cooperation between participants, which is also part of constructing the meanings for new words in language classes. Goodwin (1995: 23) also explains how a person with aphasia who is only able to say three words uses "the full expressive powers of his body (intonation, gesture, affective displays of his face and body)" when communicating.

Jokinen et al. (2013) present in their article a study on the communication possibilities of a person with Patau syndrome. They declare that meaning is created in interaction and that there are different aspects which need to be considered in meaning creation – roles, relationship between participants, shared knowledge, and contextual information (Jokinen et al., 2013: 75). Jokinen et al. (2013) reveal with the examples of communicative situations how the meaning is negotiated and how the intended meaning of a person with Patau syndrome becomes more precise in collaboration of the participants in interaction.

This paper studies the meaning construction for vocabulary items. The term *meaning construction* represents the idea that the meaning is created by use of semiotic resources and in collaboration of the participants in the interaction. This paper relies also on semiotic principles presented by Kress and Leeuwen (2001b) – the provenance of signs and the experiential meaning potential. By the provenance Kress and Leeuwen mean that human beings ""import" signs from other contexts" (Kress and Leeuwen, 2001b: 10). The experiential meaning potential refers to the idea that human being is able to "extend [his/her] practical experience metaphorically, and to grasp similar extensions made by others".

Researchers of multimodal discourse think that in the process of participants' constructing meaning, all possible simultaneously used semiotic resources need to be considered (in the classroom, this can, for instance, include using boards and additional material, people's movement and position in the room). Kress et al. (2001a) analyse multimodal communication in the science classroom, including the combination of body movement and verbal expression, the use of objects (e.g., chalk, distilled water, the anatomical model of the human body, figures on the board), the location and movement of bodies in the room in connection to verbal expression. Lim Fei (2011) studies how two teachers use language, gesture, positioning and movements in the lessons in English and outlines five categories to map in communication: time, lesson genre and lesson microgenre, gesture, space, language. Drissi (2011) analyses how French language learning takes place by video conference and distinguishes three modalities – audio, typing on the keyboard and video. Multimodal discourse analysis gives us the opportunity to take into account the effect of various semiotic resources on the study process. In the present paper the analyses is based on the semiotic resources which are involved in the construction of meaning for a vocabulary item – on verbal expression and gestures.

3 Data

The research is based on the sub-corpus of interactive communicative situations from the database of multimodal communication. More specifically, the analysis focuses on language classes videotaped at an Estonian university in 2009–2010. The aim was to collect data from university-level foreign language classes. The videos include two Estonian and two French classes. Two cameras were used. While I was taping the classes, I did not interfere with the activities of the seminar and the participants did not address me verbally during the classes.

Each seminar lasted for 90 minutes. The activities of the seminars were based on the teachers' plans; I had no input in the structure or content of the classes. The Estonian seminars had different teachers (marked EEA and EEB in Table 1, next page), both native speakers of Estonian. The French seminars had the same teacher (marked PRC in Table 1). All teachers were female. The students participating in the seminars had different native languages; the language taught was a foreign language for all of them. There were 27 episodes of communication in Estonian classes and 24 episodes in French classes in which the meaning of a word was constructed through verbal expression and gesture.

4 Method

From each seminar, only those episodes of communication were chosen in which words were explained. Such situations occurred in all four seminars. The process of explaining a word was initiated by the teacher (if the teacher had asked the students whether they knew the word and noticed that some or all did not) or by the students (who pointed out that they did not know the word). Words were explained by the teachers alone, by the teachers and students, or by the students alone.

Class	Teachers	Number of students	Students' native languages	Number of the episodes of communication in which the meaning of a word was constructed through verbal expression and gesture
Estonian class 1	EEA	5	Russian (4), Finnish (1)	11
Estonian class 2	EEB	10	Russian (8), Hungarian (1), Ukrainian (1)	16
French class 1	PRC	8	Estonian (5), Russian (1), Italian (1), Turkish (1)	4
French class 2	PRC	8	Estonian (7), Russian (1)	20

Table 1. Statistical overview of analysed classes.

In transcribing the audiovisual material, I observed not only participants' (teachers' or students') verbal expressions but also their gestures (body movements and positions), movement in the classroom, use of objects, vocal expression (e.g., stress, volume, pausing). The analysis below is based on 51 instances of the meaning of words being explained (see Table 1) in a multimodal manner, that is, through the use of both words and gestures.

The video material was transcribed using Jefferson's (2004) transcription system which has been adapted and partially modified in order to achieve a clear presentation of the episodes. Additionally, the transcriptions employ letters to mark the participants and use *l* and *r* to mark *left* and *right* (see Appendix 1). Some iconic gestures are labelled with a descriptive name instead of a lengthy phrase. Such gestures are accompanied by photographs and descriptions and are referred to in the transcriptions.

Combining the observation of the audiovisual material and the reading of transcriptions, I analysed the communicative acts of the teachers and learners as one complex multimodal phenomena in which the participants work towards constructing a meaning for a word and the semiotic resources serve as tools or aids.

5 Results

The data encompasses 51 episodes of communication in which the meaning of a word is constructed through verbal expression and gesture (see also Table 1). Multimodal construction of meaning of words is used in all classes (27 episodes in Estonian seminars and 24 episodes in French seminars). The analysis allows us to conclude that the explanation created for an unknown word is not a definition that is clearly formulated as a sentence. The work towards constructing the meaning is distributed between two semiotic resources – verbal expression and gestures. The results show that verbal expression and gestures have their specific functions and that pieces of information are organized into an explanation through the engagement of both semiotic resources. For example, gestures provide information which is not included in the verbal expression and vice versa (Ex. 1).

Example 1. PRC1's class. The teacher is explaining the word "remuer" (*to move*).

1. T: remuer (…)   j'sais pas je vous parle quand vous avez une tasse de café
   to move (...)   *i don't know i tell you if you have a cup of coffee*
   ((the r hand stirs, see Figure 1))((swings, flails the r hand))
2. T: vous mettez du sucre et vous (…) on dit touiller avec une petite cuiller remuer aussi
   you put sugar   *and you (...) we can say to stir with a little spoon to move also*
   ((the r hand puts in))   ((the r hand stirs, then turns the gaze towards A, B, C, D))

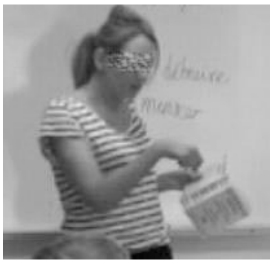


Figure 1. Example 1. Line 1 in the transcript. The teacher's right hand makes a stirring motion.

In a French class, the teacher is explaining the word "remuer". The fact that a cup of coffee and the adding of sugar are involved is made clear verbally (lines 1–2). She then adds a synonym to the word ("touiller"), saying: "on dit touiller avec une petite cuiller remuer aussi" (*we can say to stir with a spoon*) (line 2). When the teacher begins explaining the word by saying the word "remuer", a momentary silence follows during which she motions the act of stirring with her hand (line 1) (see also Figure 1). She repeats the same motion while saying the synonym for the word "remuer" ("touiller") (line 2). Prior to repeating the stirring motion, the teacher demonstrates the lifting and placing of something by touching the tips of her fingers and, at the same time, saying "vous mettez du sucre" (*you put sugar*) (line 2). The motion of stirring (lines 1–2) iconically signals the meaning of "remuer" before the verbal expression is given. This motion can help understand the meaning if "touiller" is not known to the students. The verbal expression "vous avez une tasse de café vous mettez du sucre" (*you have a cup of coffee you put sugar*) specifies the usage of "remuer" (lines 1–2).

Example 2 is taken from an Estonian class in which the words "soo" (*mire*) and "raba" (*bog*) are explained. The participants explaining the words are student A and teacher T.

Example 2. EEB's class. The teacher and the student A are explaining the word "soo" (*mire*) and "raba" (*bog*).

1. A: raba on kõrgsoo ütleme nii
a bog is a raised bog
((nods))((inclines head twice to the r))
T:
((looks at A))

2. A: bioloogias niimoodi õpetatakse
they teach it in biology

3. T: jah nii et (…)
right so (...)

4. A: ta on kõrgsoo
it is a raised bog
((inclines head to the l))

5. T: just
exactly

6. A: tal on need (…) igasugused need
it has these (...) all these
((shows a hill, separating hands in the middle, see Figure 2))

7. T:
((indicates with l palm towards A and remains waiting with an opened palm))

8. A: tal on niimoodi et all on nagu see vesi
it has like that so that like the water is below
((draws a hollow shape twice, see Figure 3))

9. A: ja peal pool on siis on siis
so on the upper side there is there is
((draws a hill with hands together and apart))

10. T: noh
well
((a nod))

11. A: kuidas öelda need
how to say these

((smiles and glances at T and then at E))
12. T: just (.) aga soo on siis see märg maa
exactly (.) but a mire is then this wet land
((draws a flat surface with l palm, fingers repeatedly touch the thumb, see Figure 4))
13. A: soo on lihtsalt märg jah
mire is simply wet yes
((draws a flat surface with r hand, see Figure 5))
T: see ei kasva ülesse kõrgemaks eks ole
it does not grow taller right
((draws a hill with l hand, see Figure 6))

In explaining the word "raba" (*bog*), student A takes the lead and says that it is taught in biology and that a bog is a raised bog (lines 1–4). Thus, "kõrgsoo" (*raised bog*) is offered as a synonym of "raba". The student continues by explaining that a bog is elevated compared to surrounding areas and has water below, all the while drawing a hill (see Figure 2) and a hollow shape in the air with hands (lines 6–9) (see also Figure 3).

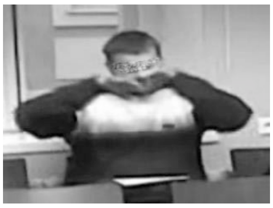


Figure 2. Example 2. Line 6. Student A draws a hill.

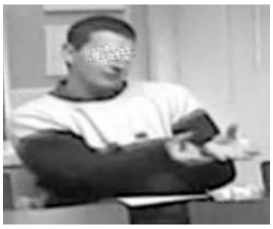


Figure 3. Example 2. Line 8. Student A draws a hollow shape demonstrating the water below the surface.

The student is looking for a suitable word to denote the top surface (line 6) but cannot find it and, instead, gestures with hands to complete the explanation. The teacher adds to this explanation by pointing to the difference between "raba" and "soo", commenting that "soo" (*mire*) is wetland as she draws a flat surface with the palm of her hand (line 12) (see also Figure 4).

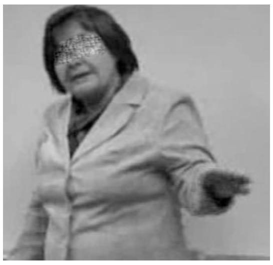


Figure 4. Example 2. Line 12. The teacher draws a flat surface (the palm of the hand moves horisontally).

Here also, one can see how the word and body movement cooperate and complement each other. The gestures outline the layers characteristic to bogs and mires. The words tell us that, with each type of surface, one is dealing with wet areas. However, in the case of a bog, the water is below the surface, whereas the surface of mire is itself wet. At the end of the episode, the teacher and student seem inspired by each other's use of gestures, as they repeat their respective motions (lines 12–13) (see also Figure 5).

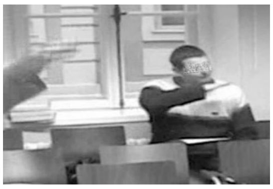


Figure 5. Example 2. Line 13. The teacher draws a hill (the hand at the left side of the Figure 5) and student A draws a flat surface.

Student A repeats the teacher's motion of drawing a flat surface and the teacher uses one hand to draw a hill (lines 12–13) (see also Figure 6), just as the student had done previously with two hands (line 9).

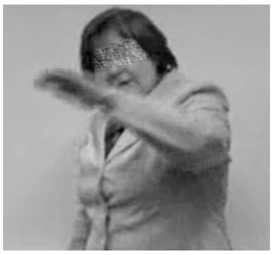


Figure 6. Example 2. Line 13. The teacher draws a hill.

The gestures are accompanied by affirmative words: the teacher uses "just" (*exactly*), "eks ole" (*right*) and the student says "jah" (*yes)* (lines 12–13). This is how they convey being in agreement to the other students and this is also how student A receives feedback that the teacher agrees with the explanation offered.

The meaning emerges through the use of two semiotic resources – verbal expression and gestures. Both examples illustrate how a gesture that reveals the meaning of a word can occur right before a longer explanation where verbal expression and gesture combine (French teacher (Example 1, line 1) and student A (Example 2, line 6)). In both examples, gestures present one visual aspect of the object explained with a word – the gestures are iconic. The verbal expressions in both instances include synonyms of the main word: in Example 1, the teacher offers the synonym "touiller" for the word "remuer", and in Example 2, the student offers "kõrgsoo" as a synonym for "raba". In the first instance, the verbal expression specifies the context of the motion and adds the substances involved (coffee and sugar). In the second, the verbal expression describes the context in which the word would be used (biology) as well as the location of water (in the surface or below it) that need not be visible. Whereas gestures present one possible visual aspect of the object described; verbal expressions create the context and add aspects that need not be visible to the eye.

6 Conclusion

The paper has shown how the construction of meaning occurs in the combination of verbal expression and gesture. Two examples were selected from the database encompassing 51 episodes of communication in order to demonstrate that the meaning emerges through the use of two semiotic resources and becomes audible and visible – as a puzzle, piece by piece – for other participants in interaction. Both semiotic resources – verbal expression and gesture – fulfil specific roles and work towards creating a meaningful whole. Both students and teachers use verbal expression as well as gesture in explaining a word. In the examples, the analysis illustrated how the gesture constructing meaning for a word can precede the verbal expression.

The study of other similar situations could give us information about other semiotic resources which might be involved in the meaning construction of a word, and might allow us to draw more general conclusions on how the meaning is constructed in French and Estonian learning classroom.

The analysis did not address the other students' co-occurring non-verbal feedback to the participants who were explaining the word. The study of feedback could uncover how the other students are engaged in the construction of the meaning for a vocabulary item and how they express understanding or confusion. Hopefully, future research into language learning will be able to shed light on the aspect of feedback and also extend the research into other areas of foreign language learning.

Acknowledgements

I would like to thank the participants in the seminars who agreed to be videotaped. This research is supported by the Estonian Science Foundation grant ETF8008 and the European Union through the European Regional Development Fund (Centre of Excellence in Cultural Theory).

References

Drissi, Samira. 2011. *Apprendre à enseigner par visioconférence : Étude d'interactions pédagogiques entre futurs enseignants et apprenants de FLE*. École Normale Supérieure de Lyon, Université de Lyon, Lyon. (Doctoral dissertation)

Goodwin, Charles. 1995. Co-constructing Meaning in Conversations with an Aphasic Man. *Research on Language and Social Interaction*, 28 (3), 233–260. `http://www.sscnet.ucla.edu/clic/cgoodwin/95co_aphasic.pdf` (accessed in August 2014)

Gullberg, Marianne. 1998. *Gesture as Communication Strategy in Second Language Discourse: A Study of Learners of French and Swedish*. Lund University Press, Lund.

Hall, Joan K. 2009. Interaction as Method and Result of Language Learning. *Language Teaching*, 43, 1–14.

Jefferson, Gail. 2004. Glossary of Transcript Symbols with an Introduction. In G. H. Lerner (Ed.) *Conversation Analysis: Studies from the First Generation*, John Benjamins, Philadelphia. 13–23.

Jokinen, Kristiina P., Tenjes, Silvi, and Ingrid Rummo. 2013. Embodied Interaction and Semiotic Categorization: Communicative Gestures of a Girl with Patau Syndrome. In C. Paradis, J. Hudson and U. Magnusson. (Eds.) *Conceptual Spaces and the Construal of Spatial Meaning*. Oxford University Press. 74–97.

Kendon, Adam. 1988. How Gestures Can Become Like Words. In F. Poyatos (Ed.) *In Cross-cultural Perspectives in Nonverbal Communication*. C. J. Hogrefe, Lewiston, New York, 131–141.

Kendon, Adam. 2004. *Gesture: Visible Action as Utterance*. Cambridge University Press, Cambridge.

Kress, Gunther, Carey Jewitt, Jon Ogborn and Charalampos Tsatsarelis. 2001a. *Multimodal Teaching and Learning: The Rethorics of the Science Classroom*. In C. Candlin and S. Sarangi (Ed.), BookEns Ltd, Royston, Herts, London New York.

Kress, Gunther and Theo van Leeuwen. 2001b. *Multimodal Discourse: The Modes and Media of Contemporary Communication*. Arnold, Hodder Headline Group, London, Oxford University Press Inc, New York.

Lazaraton, Anne. 2004. Gesture and Speech in the Vocabulary Explanations of One ESL Teacher: A Microanalytic Inquiry. In A. Z. Guiora, K. Bardovi-Harlig, Z. Dörnyei (Eds.). *Language Learning: A Journal of Research in Language Studies*, 54 (1), 79–117.

Lim Fei, Victor. 2011. *A Systemic Functional Multimodal Discourse Analysis Approach to Pedagogic Discourse*. National University of Singapore, Singapore. (Doctoral dissertation).

McNeill, David. 2000. Introduction. David McNeill (Ed.), *Language and Gesture*. Cambridge University Press, Cambridge. 1–10.

Mondada, Lorenza and Simona P. Doehler. 2004. Second Language Acquisition as Situated Practice: Task Accomplishment in the French Second Language Classroom. *The Modern Language Journal*, 88, 501–518.

O'Halloran, Kay L. 2011. Multimodal Discourse Analysis. In K. Hyland and B. Paltridge (Eds.) *Companion to Discourse*. Continuum, London and New York.

Rummo, Ingrid and Silvi Tenjes. 2011. AJA mõistestamine Patau sündroomiga subjekti suhtluses. *Eesti Rakenduslingvistika Ühingu aastaraamat*, 7, 231–247.

Shepherd, Michael A. 2010. A Discourse Analysis of Teacher-Student Classroom Interactions. University of Southern California, Southern California. (Doctoral dissertation). `http://digitallibrary.usc.edu/assetserver/controller/item/etd-Shepherd-3879.pdf` (accessed in October 2011)

Appendix A. Transcript symbols.

(...)	pausing
(( ))	hand gestures and other body movements (mimic, gaze, movements with different body parts)
T	teacher
A, B, C, D, E	students
l	left
r	right

On the Attribution of Affective-Epistemic States to Communicative Behavior in Different Modes of Recording

Stefano Lanzini
SCCIIL (SSKKII)
Gothenburg University
lanzhbk@hotmail.it

Jens Allwood
SCCIIL (SSKKII)
Gothenburg University
jens.allwood@gu.se

Abstract

Face-to-face communication is multimodal with varying contributions from all sensory modalities; see e.g. Kopp (2013), Kendon (1980) and Allwood (1979). This paper reports a study of respondents interpreting vocal and gestural verbal and non-verbal, behaviour. 10 clips from 5 different short video + audio recordings of two persons meeting for the first time were used as stimulus in a perception/classification study. The respondents were divided in 3 different groups. The first group watched only the video part of the clips without any sound. The second group listened to the audio track without video. The third group was exposed to both the audio and video tracks of the clip. In order to collect the data, we used a crowdsourcing questionnaire. The study reports on how respondents classified clips containing 4 different types of behaviour (looking up, looking down, nodding and laughing) that were found to be frequent in a previous study (Lanzini 2013) according to which Affective Epistemic State (AES) the behaviours were perceived as expressing.

We grouped the linguistic terms for the affective epistemic states that the respondents used into 27 different semantic fields. In this paper we will focus on the 7 most common fields, i.e. the fields of Thinking, Nervousness, Happiness, Assertiveness, Embarrassment, Indifference and Interest. The aim of the study is to increase understanding of how exposure to video and/or audio modalities affect the interpretation of vocal and gestural verbal and non-verbal behaviour, when they are displayed uni-modally and multi-modally.

Keywords: Affective Epistemic States, Multimodality, Gesture, Speech, Verbal, Non-verbal Communication, vocal, auditory

1 Introduction

This paper explores the relative role of auditory and visual information for the attribution of affective-epistemic states to 4 different types of behaviour (“looking up”, “looking down”, “nodding” and “laughing”) occurring in video clips taken from short video + audio recordings of two persons meeting for the first time.

By the term “Affective Epistemic State” we refer to internal human states that involve emotion, other aspects of cognition or perception (Allwood, 2012), e.g. Happiness, Sadness, Relaxation, Nervousness, alternatively described by Schroder (2011) “states which involve both knowledge and feeling” (Schroder, 2011).

We are considering both verbal and non-verbal behaviours expressed by vocal and gestural means, since many affective-epistemic and feedback functions are expressed simultaneously with all of these means, cf. Allwood, & Cerrato (2003) and Boholm (2011). Specifically, we are interested in investigating to what extent only visual, only auditory or both visual and auditory behaviour are involved.

K. Jokinen and M. Vels. 2015. Proceedings of The 2nd European and the 5th Nordic Symposium on Multimodal Communication.

2 Method

In this study we used 10 clips from 5 different recordings of pairs of 1:st language speakers of Swedish who are meeting for the first time, as stimulus in a crowd sourcing questionnaire study. The language used in the meetings is Swedish. The questionnaire was made with Google Drive and we employed random recruitment of respondents via social media. The duration of the clips varied from 7 sec to 20 sec, with an average length of 12.36 sec

There were 93 respondents, from different cultures. After having been exposed to the clips, they answered the questionnaire, in electronic form, available on the internet. We presented the subjects with recorded situations in three different conditions: video with audio (Video + Audio (30 persons)), video without audio (Video-only (35 persons)) and audio alone (Audio-only (28 persons)). The participants had to make an interpretation of which AES was expressed in a particular clip, in a particular presentation condition.

Each participant was exposed to 10 clips all in the same mode of presentation. The AESs had to be selected from a fixed list of options that were suggested by respondents in a previous study with Swedish stimulus material (Lanzini 2013). The AESs were given in English and were the following: Happiness, Sadness, Relaxation, Nervousness, Disinterestedness, Interest, Pride, Shyness, Confidence, Surprise, Sarcasticness, Aggressiveness, Thoughtfulness, Excitement, Unsureness and Playfulness

In addition, the participants could suggest other terms that according to them better described the AES they perceived. The participants also had to give a motivation for their answers.

3 Data

3.1 AESs grouped in Semantic fields

There are many words denoting different affective epistemic states in most languages. Some of the terms denote states that are closely related like “anger” and “wrath”. In our original study, we used free choice and consequently got very many different response terms for affective epistemic states. In order to make the data set more manageable we, in this study, grouped the terms used in the responses into semantic fields. A semantic field is a list of linguistic terms that share semantic characteristics. Below we present a list of the most frequent semantic fields made up of the linguistic terms for the AESs that we had obtained in a previous study (Lanzini 2013). For each clip, respondents were asked to write a term for only one Affective Epistemic State. The semantic fields were created after the data had been collected, and they were created on intuitive grounds by the researchers in order to group together AES terms with a similar semantic meaning. The following 7 semantic fields will be discussed below.

- **Thinking:** Thinking, Remembering, Reflective, Thoughtful, Giving Explanations
- **Nervousness**: Nervous, Uneasy, Unsure, Insecure, Uncomfortable, Hesitant, Reluctant, Uncertain, Unconfident
- **Happiness**: Happy, Good Mood, Amused, Joyous, Happy and Calm, Glad
- **Assertiveness**: Assertive, Sure, Proud, Confident, High Self Esteem, Assured Persistent, Insistent
- **Embarrassment**: Embarrassed, Self-conscious, Timid, Intimidated, Ashamed, Humbled, Shy, Reserved, Modest, Submissive
- **Indifference**: Indifferent, Apathetic, Lazy, Neutral, Evasive, Not Concentrated, Disinterested, Bored, not Interested
- **Interest**: Interested, Surprised, Participating, Engaged, Curious, Concerned, Hopeful, Motivated, Willing

3.2 Gestural behaviour

We will now present the most common interpretations of the following four gestural behaviours; “looking up” (2 clips from 2 videos), “looking down” (3 clips from 3 videos), “nodding” (3 clips from 3 videos) and “laughing” (2 clips from 2 videos). The 4 behaviours and their descriptive labels (“looking up” etc.) were chosen on the basis of being the most selected behavioural descriptive labels and, thus likely to be associated with easily perceived behaviour, in the previous study (Lanzini 2013). Every recorded behaviour was presented in the three presentation conditions (only audio, only video, audio + video) introduced above. The word "whole" means that the whole body including feet is presented on the video, while in other cases only the upper part of the body is presented. The yellow fields indicate the AES attribution with the highest proportion of respondents for a particular clip in a particular condition of presentation. All AESs that turned out to be the most popular in any of the three conditions of presentation for any recording of the chosen 4 types of behaviour are included. Capital letters are used when referring to a semantic field, e.g. "Nervousness". All tables below show the most frequent semantic fields used by respondents. The percentages are generated by dividing the number of responses using a particular semantic field with the number of respondents for a particular condition of presentation. There were 28 respondents in the audio condition, 30 respondents in the video+audio condition and 35 respondents in the video condition.

3.3 Looking-up (2 clips)

Only Video (35 persons)	Nervousness	Thinking
Clip(4) looking up	26%	49%
Clip (5) whole body	20%	60%

Video+Audio (30 persons)	Nervousness	Thinking
Clip(4) looking up	43%	20%
Clip (5) whole body	27%	50%

Only Audio (28 persons)	Nervousness	Thinking
Clip(4) looking up	36%	14%
Clip (5) whole body	21%	21%

Table 1. Percentage of respondents for each condition of presentation using the 2 most common AES interpretations of "looking up" in the 3 conditions of presentation

In table 1, the two most common AES interpretations of "looking-up" behaviour are Nervousness and Thinking. The Thinking field interpretation is most popular when respondents have access to only video without sound, while the second most popular; Nervousness, is most frequent when they have access to both audio and video. The data also shows that an audio presentation does not strongly evoke a Thinking interpretation while it does evoke an interpretation of Nervousness. In the audio-only presentation condition, the speech in clip (4) was mostly perceived as a sign of Nervousness (36%), while in clip (5) ”whole body” a smaller number of respondents (21%) perceived the speech as a sign of Thinking, which was the same percentage of respondents that interpreted it as an expression of Nervousness. According to respondents in the multimodal condition, the two clips of "looking up" are perceived differently mostly because of the verbal vocal behaviour which sounded more nervous in the audio-only presentation. In clip (4) the combination of speech and body movements increased the percentage of respondents that perceived Nervousness in comparison to both unimodal audio and unimodal video. In audio-only, Nervousness got a higher percentage of perceptions (36%) than Thinking (14%).

In contrast, in clip (5) "whole body", Thinking as an interpretation of "looking up", increases both in unimodal video and multimodal condition. This can be related to the fact that for clip (5), Nervousness and Thinking got the same number of interpretations (21%) in audio-only mode. Thus, Nervousness is most commonly attributed with multimodal data, less so with audio-only and least with video-only. So the combination of nervous speech and nervous body movement increases the perception of Nervousness.

Thoughtfulness and Thinking are most commonly attributed with video-only, less with multimodal data and least with audio-only. So the attribution of Thinking AESs decreases when the gestural behaviour of looking up is presented together with speech. It decreases a lot, so that if in audio-only, the speech is perceived as a sign of Nervousness.

3.4 Looking-down (3 clips)

Only Video (35 persons)	Nervousness	Embarrassment	Indifference
Clip (3) looking down	40%	11%	11%
Clip (7) looking down	31%	60%	0%
Clip (2) whole body	23%	31%	26%

Video+Audio (30 persons)	Nervousness	Embarrassment	Indifference
Clip (3) looking down	40%	3%	13%
Clip (7) looking down	43%	33%	0%
Clip (2) whole body	33%	13%	10%

Only Audio (28 persons)	Nervousness	Embarrassment	Indifference
Clip (3) looking down	25%	7%	50%
Clip (7) looking down	36%	14%	4%
Clip (2) whole body	29%	25%	4%

Table 2. Percentage of respondents for each condition of presentation using The 3 most common AES interpretations of "looking down", in 3 presentation conditions, and 3 clips.

In table 2, "Looking-down" is most strongly related to the 3 semantic fields of Nervousness, Embarrassment and Indifference. If we compare the multimodal mode with the unimodal conditions, we see that when the three clips (3), (7) and (2) "whole body" were presented with speech and gesture together, for clip (7) and clip (2) "whole body", the attribution of Nervousness increased, like it did in relation to "looking up". Clip (3) got the same number of attributions of Nervousness in the video-only and multimodal mode. Nervousness seems clearly noticeable in both speech and gesture, with speech cues possibly slightly more important.

If we consider the unimodal video condition, for "looking-down", the semantic field of Embarrassment has a higher number of attributions than it has for unimodal audio and multimodal audio+video, in all three clips.

In conclusion, it seems that speech has a negative effect on the attribution of Embarrassment-related AESs. This observation is supported by the fact that for these three clips, the semantic field of Nervousness got a much higher number of attributions, in the audio-only condition, than the field of Embarrassment.

It is also interesting to note that in the audio mode, 50% of respondents of clip (3), interpreted the speech as a sign of Indifference. The attribution of Embarrassment decreased in all three clips when presentation of body movement was combined with presentation of speech or given only in speech while it clearly increased the attribution of Nervousness. The attribution of Indifference shows a more varied picture, being most frequent for clip 3 when presented in audio-only.

3.5 Nodding (3 clips)

Only Video (35 persons)	Nervousness	Assertiveness	Interest
Clip (5) nodding	6%	14%	26%
Clip (6) nodding	31%	6%	29%
Clip (4) whole body	11%	11%	29%

Video+Audio (30 persons)	Nervousness	Assertiveness	Interest
Clip (5) nodding	7%	43%	10%
Clip (6) nodding	23%	0%	43%
Clip (4) whole body	3%	7%	50%

Only Audio (28 persons)	Nervousness	Assertiveness	Interest
Clip (5) nodding	0%	39%	32%
Clip (6)	4%	4%	54%
Clip (4) whole body	7%	7%	46%

Table 3. Percentage of respondents for each condition of presentation using The 3 most common AES interpretations of "nodding", in 3 presentation modes, and 3 clips.

Table 3 shows us that “Nodding” is most strongly related to the semantic field of Interest. For all clips this effect is strongest in the audio-only condition and lowest in the video-only condition. For clip (5), the Interest attribution is most infrequent in the multimodal condition. Thus, for Interest attributions, the vocal behaviour produced while people are nodding has equal or more influence than the nodding itself. Probably the video unimodal condition provides too little information for respondents to clearly attribute the AESs of Interest.

For the semantic field of Assertiveness, the case is less clear. For clip (5) speech plays an important role and this attribution decreases in the video-only presentation. However, the case is less clear for clip (6) and (4).

“Nodding” is also related Nervousness but here the relation to the video mode is stronger than in the case of “looking-up” and “looking-down

3.6 Laughing (2 clips)

Only Video (35 persons)	Nervousness	Happiness	Assertiveness	Embarrassment
Clip (2) laughing	23%	23%	6%	14%
Clip (7) whole body	29%	11%	3%	23%

Video+Audio (30 persons)	Nervousness	Happiness	Assertiveness	Embarrassment
Clip (2) laughing	23%	7%	3%	33%
Clip (7) whole body	23%	17%	7%	13%

Only Audio (28 persons)	Nervousness	Happiness	Assertive-ness	Embarrassment
Clip (2) laughing	32%	11%	18%	7%
Clip (7) whole body	21%	7%	21%	7%

Table 4. Percentage of respondents for each condition of presentation using The 4 most common AES interpretations of “laughing”, in 3 presentation modes, and 2 clips.

Laughing involves both gestural and vocal behaviours. In table 4, we see that when laughing is presented multimodally with both sound and visible behaviour, in clips (2) and (7), it is mostly interpreted as a sign of Nervousness and/or Embarrassment. However, the two clips are quite different and the video participants laughed in very different ways.

If we consider the semantic field of Assertiveness we can note that respondents more frequently attributed AESs of this type when the laughter was presented in audio-only condition (clip (2), 18% and clip (7), 21%). The attributions of Assertiveness decrease in both unimodal video condition and in multimodal condition. So it seems that the properties providing Assertiveness in speech loose their effect when combined with gesture. In contrast, the semantic fields of Happiness and Embarrassment got a higher number of attributions in the video mode than in the audio mode, indicating that for these types of AES, visual cues seem to carry more influence than auditive cues.

4 Summary and discussion

The main conclusion concerning the four behaviours we have studied (looking up, looking down, nodding and laughing) is that no easy generalizations are available. What type of affective-epistemic state the behaviours are seen as expressing depends on the particular person expressing the AES and which sensory modality it is presented in. If we consider the four types of behaviour, some of the main results are the following with regard to mode of presentation:

(i) Looking-up

The most frequent semantic field to be associated with this behaviour is the field of Thinking and thoughtfulness. The association is strongest with the visible behaviour of “looking-up” and much weaker with the speech accompanying the visible behaviour. Perhaps this reflects that for Thinking the visual cue of looking-up is the strongest. For the second most common AES field, Nervousness, the opposite holds. Nervousness is most frequently associated with the speech accompanying “looking-up” behaviour, if respondents also hear the speech accompanying the bodily movement or do not see the bodily behaviour.

(ii) Looking-down

"Looking-down" is most strongly related to the semantic field of Nervousness followed by Embarrassment and Indifference. As is the case for "looking-up", Nervousness is most frequently attributed when both speech and gesture are available.

The semantic field of Embarrassment has a higher number of attributions for unimodal video than it has for unimodal audio and multimodal audio+video, in all three clips. Thus Embarrassment like Thinking seems to have a strong visual side.

50% of the respondents to clip (3), interpreted the speech accompanying the "looking-down" sequence as a sign of Indifference, when only presented with the audio condition. When presented in video-only or multimodally this decreased the attribution of Indifference, indicating that for this clip the important cue for Indifference was auditive rather than visual.

(iii) Nodding

The most common semantic field attributed to "nodding" is Interest, followed by Assertiveness and Nervousness ". The connection is strongest in the audio-only and multimodal presentation condition and slightly weaker in the video-only condition indicating that audio cues play an important role in making nodding an expression of Interest.

For Assertiveness, the case is less clear. Only one clip (5) shows a clear pattern of speech playing an important role, similar to what is the case for Interest, occurring in the audio-only presentation and in the multimodal presentation with a decrease in the video-only presentation.

Somewhat surprisingly for nodding, the relation of Nervousness to the video mode is stronger than in the case of "looking-up" and "looking-down" where the auditive cues were more important.

(iv) Laughing

The most common attribution to "laughter" was Nervousness followed by Happiness, Assertiveness and Embarrassment, all about equally common. Nervousness was attributed to laughter to roughly the same degree in all conditions of presentation. For Happiness and Embarrassment visual cues seemed slightly more important than auditive while for Assertiveness the opposite seemed to the case, making auditive cues the most important.

References

Allwood, J. 1979."Ickeverbal kommunikation - en översikt" in Stedje and af Trampe (Ed.) *Tvåspråkighet*. Stockholm, Akademilitteratur. Also in *Invandrare och Minoriteter* nr 3, 1979, pp. 16-24.

Allwood, J. and Cerrato, L. 2003 A Study of Gestural Feedback Expressions. *First Nordic Symposium on Multimodal Communication*. Paggio P. Jokinen, K. Jönsson, A. (eds). Copenhagen, 23-24, September 2003, pp. 7-22.

Boholm M. and Lindblad G. 2011. Head movements and prosody in multimodal feedback. *NEALT Proceedings Series: 3rd Nordic Symposium on Multimodal Communication,* 15, p. 25-32.

Chindamo, Massimo, Allwood, Jens & Ahlsén, Elisabeth. 2012. Some suggestions for the study of stance in communication. *Proceedings of IEEE SocialCom Amsterdam 2012*, 3-5.

Kopp Stefan. 2013. Giving interaction a hand: deep models of co-speech gesture in multimodal systems. *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 245-246.

Kendon,Adam. 1980. Gesticulation and speech: two aspects of the process of utterance. In M.R.Key (ed), *The Relationship of Verbal and Nonverbal Communication*, pp. 207-227. The Hague: Mouton and Co.

Lanzini, Stefano. 2013. *How do different modes contribute to the interpretation of affective epistemic states*. Published master's thesis for master's degree. University Gothenburg, Division of Communication and Cognition, Department of Applied IT.

Schroder Marc, Bevacqua Elisabetta, Cowie Roddy, Eyben Florian, Gunes Hatice, Heylen Dirk, ter Maat Mark, McKeown Gary, Pammi Sathish, Pantic Maja, Pelachaud Catherine, Schuller Bjorn, de Sevin Etienne, Valstar Michel, and Wollmer Martin. 2011. Building Autonomous Sensitive Artificial Listeners. *IEEE Trans. Affective Computing*. 9. (1). p. 1

Multimodal Communicative Feedback in Swedish

Gustaf Lindblad
SCCIIL (SSKKII)
University of Gothenburg
Gothenburg, Sweden
gustaf.lindblad@gu.se

Jens Allwood
SCCIIL (SSKKII)
University of Gothenburg
Gothenburg, Sweden
jens.allwood@gu.se

Abstract

This study investigates multimodal communicative feedback among speakers of Swedish. We find that the most common way of providing feedback in Swedish is by a multimodal combination of a gestural verbal and a vocal-verbal basic feedback unit, or by just a feedback word or a verbal head gesture on its own. The most common verbal head gestures are nods, and the most common vocal-verbal feedback is just one of four short words. We also find that while nods are primarily used for giving feedback, all other head gestures are more typically used for non-feedback purposes.

1 Introduction

In this paper, it is our intention to describe multimodal communicative feedback in Swedish. Our aim is to present a fairly general overview of Swedish multimodal feedback. The present paper thus continues the work presented in Cerrato 2002) and 2007). The present examination is based on a different data set than previous work. On the basis of this, a second aim is to substantiate or call in to question previous results. The paper will focus on the most common types of communicative feedback, trying to see the typical and broader patterns. Because of the combinatorial properties of multimodal communication, an in depth description would be too extensive, if it were to handle all possible combinations and aspects.

Based on Allwood et al. (1992) and Allwood, Kopp et al. 2007), we define communicative feedback as unobtrusive vocal and gestural communicative contributions that “inform an interlocutor about the ability and willingness to (i) continue the interaction, to (ii) perceive, and (iii) understand what is communicated, and (iv) in other ways attitudinally and emotionally react to this” (Boholm and Lindblad, 2011).

We define gestures as all non-vocal bodily movements that are used for communication. This includes non-voluntary movements that are nevertheless interpreted by the second party as giving information about the message or states of the first party. This inclusive definition is motivated by the fact that it is difficult to draw a definitive line between volitional and non-volitional communicative behavior.

2 Method

The data consists of ANVIL annotations (Kipp, 2001) of six dyadic first acquaintance interactions of Swedish people. In total 11 different persons participate in the interactions (one person participates in two), four of which are female-male interaction, one female-female and one male-male. Each interaction lasts approximately eight minutes (the total length of the six interactions is 48 minutes, 5 seconds), and was filmed using three different camera angles (see Figure 1).

The annotations were transcribed using the Gothenburg Transcription Standard, GTS, (Nivre, 2004), transcriptions imported into ANVIL using Praat (Boersma, 2001), and annotated using the MUMIN coding scheme (Allwood, Cerrato et al., 2007). Regrettably it is not possible to present intercoder reliability, as the data set used in this article has not been double coded. However, transcriptions and ANVIL annotations alike were checked by at least one person other than the annotator to make

K. Jokinen and M. Vels. 2015. Proceedings of The 2nd European and the 5th Nordic Symposium on Multimodal Communication.

sure that they complied with the specifications. We therefore have a fairly strong confidence in the reliability of this data. It should also be noted, that inter-coder reliability is a somewhat blunt measure of data usefulness, as it does not measure the most valuable characteristic, which is validity.

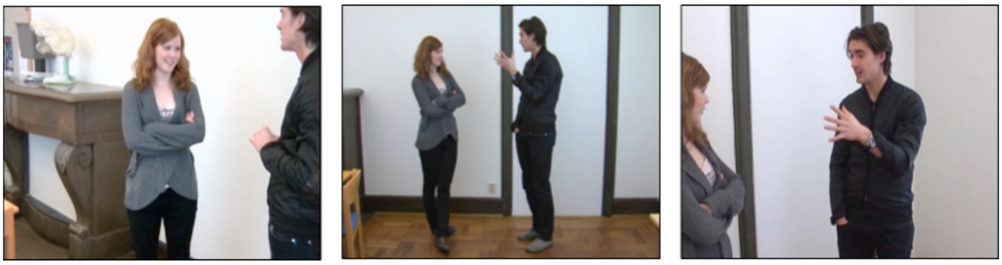


Figure 1. Examples of what the three camera angels captured during one conversation.

The MUMIN coding scheme provides guidelines for classification of bodily behaviour into discrete units in our annotations. We will not describe all the different possibilities here, but because head movements are the most commonly used gestures to provide feedback in conversation, a short description of these varieties is called for.

In accordance with the MUMIN coding scheme we differentiate four different types of nods based on two dimensions of expression: the direction of the initial movement of the nod, and whether it is a single or a repeated nod. This yields the four basic types: down-nod single (ds), down-nod repeated (dr), up-nod single (us) and up-nod repeated (ur). Previous research (e.g. Boholm and Allwood, 2010; Boholm and Lindblad, 2011) has supported this classification, as these different types show different patterns of production. Apart from nods we classify head movements into seven further categories: shake, side turn, tilt, waggle, head forward, head backward and other. The 'shake' refers to the repeated turning of the head from side to side around the longitudinal axis common in most European cultures, 'side turn' refers to just turning the head non-repeatedly. 'Tilt' is a sideways (left or right) slanting of the head away from the longitudinal axis of the body, 'waggle' refers to a rapidly repeated 'tilt'. 'Head forward' and 'head backward' are somewhat similar to nods, but features a rapid initial movement and subsequent slower normalization of the head position, whereas nods are characterized by a more oscillating movement. The 'other' category is used for all other conceivable movements of the head that are not captured by the specified categories.

Every distinct bodily gesture was coded as its own feature (element) in ANVIL, and coded as either feedback or non-feedback. In some cases it is not immediately clear where one gesture ends and the next one begins, but as a general rule we would separate a continuous bodily movement into two or more elements if the movement had salient different parts described by the MUMIN coding scheme. This was primarily an issue with regards to hand gestures, whereas facial expression, head movements and other bodily movements generally had a more pronounced beginning and end.

Vocal verbal contributions were annotated as their own units according to the GTS, with one exception, which is contributions beginning with feedback and then continuing with non-feedback. In these cases, the feedback part and the rest of the contribution was coded as separate units.

3 Results

Out of 4993 annotated features (elements) in our data set, 1486 were coded as providing communicative feedback. Of these, 1406 included either vocal-verbal or verbal head gestures. This means that there were only 80 feedback features using facial, hand or other bodily gestures. Because there are so few of each kind, these are excluded from the further analysis in the present paper.

Gesture category	n.	Multimodal
Body posture	15	14
Facial expression	53	50
Hand gesture	12	10

Table 1. Non-vocal, non-head gesture feedback.

Of the 1406 remaining feedback features 912 are annotated as being multimodal (456 vocal-verbal, 456 verbal head gestures), which means that there are 950 feedback units (1406 - 456 = 950) in the data set. This means that, on average, there is feedback every 3 seconds in these recordings (((48 * 60 + 5) seconds) / (950 feedback units) = 3.04 seconds/feedback unit), illuminating the ubiquity of this phenomenon in conversation.

3.1 Multimodal and unimodal overview

The most common way to give feedback is by means of a multimodal combination of vocal-verbal plus verbal head movement, 456 out of 950 instances (48%). Second most common is a unimodal vocal-verbal feedback, 331 of 950 (35%), and third a unimodal verbal head movement, 163 of 950 (17%). Overall, we see that multimodal and unimodal feedback are equally common, but from the perspective of the respective modalities you can also say that both vocal-verbal feedback and gestural verbal feedback is more often produced as a multimodal unit than as a unimodal unit, with 456 out of 787 (58%) of vocal-verbal feedback and 456 out of 619 (74%) of feedback head movements being produced in a multimodal unit. The ratios are close to identical with what Boholm and Lindblad 2011) found in a different but comparable data set, indicating that these patterns are stable in this kind of casual conversation.

	This study		Boholm & Lindblad 2011)	
	n.	**%**	**n.**	**%**
Multimodal	456	48,0%	413	48,9%
Unimodal vocal-verbal	331	34,8%	290	34,4%
Unimodal head movement	163	17,2%	141	16,7%
Total	950	100,0%	844	100,0%

Table 2. Comparison of overall multimodal and unimodal feedback in this study to a study by Boholm and Lindblad 2011).

3.2 Head gestures

There were 1297 head gestures annotated in our data set, of which 621 were annotated as feedback and 676 as non-feedback head gestures. Since there were only two instances of the ‘waggle’ head gesture used for feedback, this type has been left out from further analysis as a feedback gesture in this paper. Table 3 presents all occurrences of all head gesture types.

Head gesture	dr	ds	ur	us	back	forward	shake	side turn	tilt	waggle	other
Total	242	127	103	135	89	109	48	179	167	31	67
Non-feedback	68	44	17	26	50	76	33	163	129	29	41
Feedback	174	83	86	109	39	33	15	16	38	2	26
% feedback	72%	65%	83%	81%	44%	30%	31%	9%	23%	6%	39%
Multimodal fb	116	63	63	97	36	10	12	11	32	0	16
Unimodal fb	58	20	23	12	3	23	3	5	6	2	10
% multimodal	67%	76%	73%	89%	92%	30%	80%	69%	84%	0%	62%

Table 3. Occurrences of the different types of head gestures.
(dr = down repeated, ds = down single, ur = up repeated, us = up single)

Something that immediately stands out is that all types of nods are much more frequently used for giving feedback, whereas all other head gestures are more frequently used for non-feedback gestures. This is shown more clearly in Figure 2. We also note that this is most pronounced for up-nods, that seem to be used predominantly for giving feedback, as well as for ‘side turn’, ‘tilt’ and ‘waggle’ which are mainly used for non-feedback gesturing.

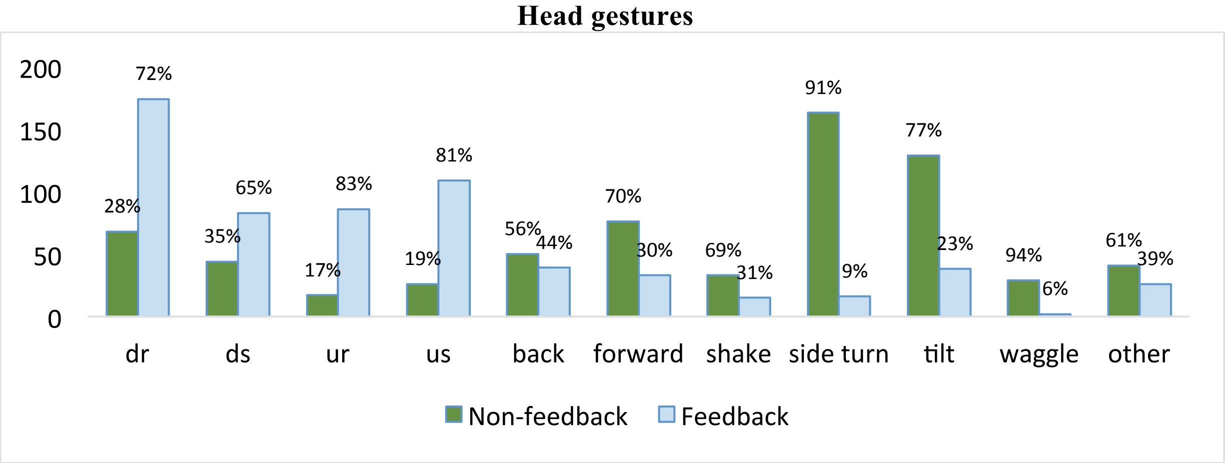


Figure 2. Comparison of non-feedback to feedback head gestures.

Nods are the most common head gestures used in Swedish to express feedback, with 452 out of 619 instances of verbal feedback head gestures in our data (73%) being nods. By contrast, headshakes are the least common of our basic types of head movements, with only 15 instances 2%). This is especially interesting considering that nods and shakes are often regarded as basic head gestures expressing ‘yes’ and ‘no’ respectively. For comparison, different basic varieties of ‘yes’ (‘ja’) account for 292 out of 787 instances (37%) of vocal-verbal feedback, and basic varieties of ‘no’ (‘nej’) for 56 (7%). Considering the multimodal combinations, we find only one instance of a headshake coupled with a vocal-verbal ‘yes’, whereas seven are coupled with a single ‘no’, three are coupled with a short phrase beginning with the word ‘no’, three are unimodal, and one is coupled with a feedback cluster containing the word ‘no’.

Most feedback head gestures are multimodal (74%), but broken down into the different types, we find that there are differences. The single up-nod and the head backwards gestures are the most likely gestures to be produced multimodally (around 90% of the time), which is interesting as these gestures are quite similar in their initial phase with an upward-backward movement of the head. The head forward gesture is the only gesture that is produced unimodally most of the time.

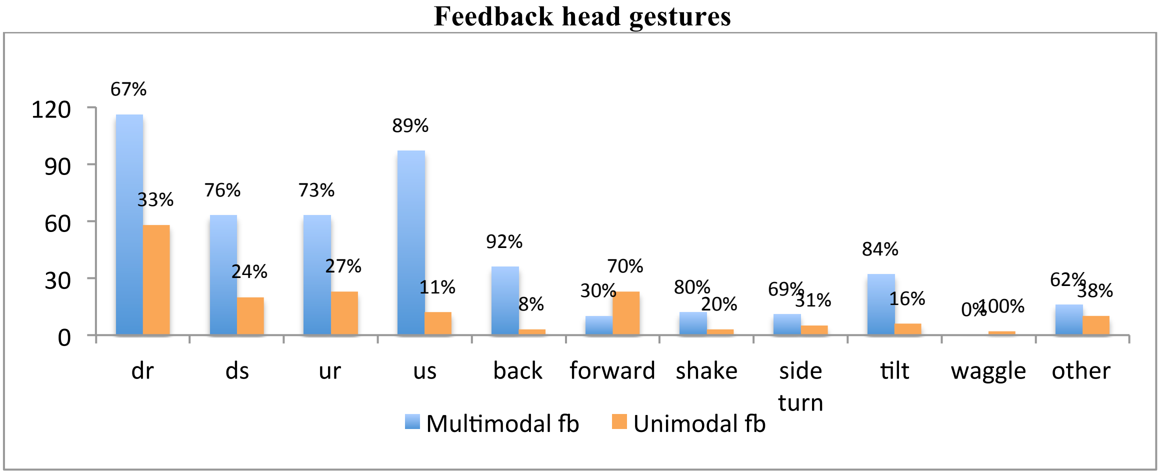


Figure 3. Occurrences of feedback head gestures.

3.3 Vocal-verbal

There were 1570 vocal-verbal contributions (or utterances) in our annotated material, with 787 annotated as containing communicative feedback, which leaves 783 as non-feedback. This means that half of all utterances in our data are feedback, which does not mean that half of what is being said is feedback, as the average duration of feedback utterances is 0.49 seconds (st.dev. 0.37) and the average duration of non-feedback utterances is 2.97 seconds (st.dev. 3.67). It should be noted that in cases where a feedback expression heads up a longer contribution, only the initial feedback part is being used for our calculations. The possible different forms of vocal-verbal feedback are many, but in reality the majority of the feedback utterances fall into a more limited set of categories. In our data 458 of the 787 feedback utterances are one of four basic Swedish feedback words: 'ja', 'm', 'okej' and 'nej'. These words can be produced in some different varieties, for instance with reduction of the 'j' phoneme in 'ja', 'okej' and 'nej'. For the sake of brevity we will disregard these differences and focus only on the basic word types in this paper, though we acknowledge that these differences can be of significance.

There are also 119 cases of what we call feedback clusters or feedback phrases, which are two or more of the basic feedback words produced together in rapid succession. It is very common to repeat the same word (e.g. 'ja ja ja'), but also combinations of two or more different words occurs (e.g. 'ja okej'). In total, this means that 577 out of 787 feedback utterances (73%) consist of one or more of the four most common feedback words in Swedish.

There are 20 cases of what we call 'other repetition', which is when a person gives feedback by repeating a word or utterance that the interlocutor has just said (e.g. A: "I will come tomorrow" B: "Tomorrow", where B's utterance would count as other repetition feedback). Basic feedback words are excluded from this category as not to be counted twice. But it should be noted that also these words can be other-repeated, which reinforces their feedback function.

Of the remaining 190 feedback contributions, no one type has an occurrence of 20 times or more, and most only occur once. Many of them consist of a basic feedback word and a few other words, e.g. 'ja det är det' ('yes it is'), 'ja visst' ('yes sure') or 'nä jag förstår' ('no I see').

Feedback type	**ja**	**m**	**okej**	**nej**	**cluster**	**other repetition**	**all others**	**TOTAL**
Total	238	139	38	43	119	20	190	787
Unimodal	106	65	14	22	27	7	90	331
Multimodal	132	74	24	21	92	13	100	456
% Multimodal	55%	53%	63%	49%	77%	65%	53%	58%

Table 4. The most common types of vocal-verbal feedback and their multimodal frequencies.

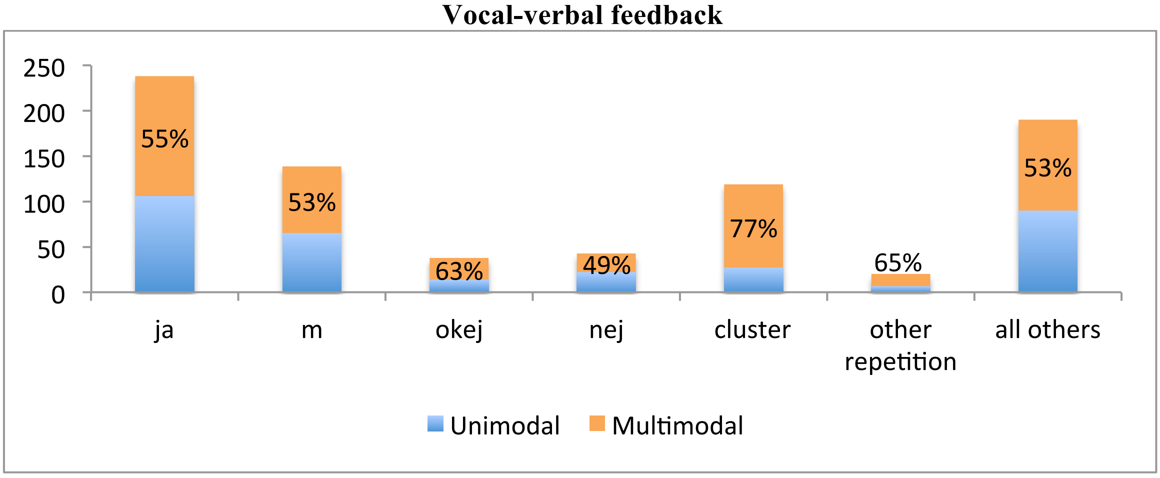


Figure 4. The most common types of vocal-verbal feedback and their multimodal distribution.

It is clear that the most common feedback word in Swedish is 'ja' followed by 'm', whereas 'okej' and 'nej' are much less common although still fairly frequent. This pattern has previously been shown in several studies (e.g. Allwood 2000), Boholm and Lindblad, 2011; Navarretta et al., 2012), and seems to be fairly stable. We also note that the basic feedback words are produced multimodally with head gestures about 50% of the time, with the exception of 'okej' that has a tendency to be coproduced with head gestures more often. Other repetition is also more likely to be co-produced with a head gesture, and feedback clusters even more so.

3.4 Multimodal: vocal-verbal and head gesture

When we consider the combinations of vocal-verbal and head gesture feedback, we find that some combinations seem to be more common than others. It is difficult to make a table that reflects all the interplay between the types, as their frequencies are so varied. Some trends are more easily discernable though. In table 5 we have shaded the cells darker for higher numbers, comparing on the horizontal axis, from the perspective of vocal-verbal feedback. Table 6 is shaded vertically, from the perspective of the verbal head gestures. Each perspective tells a somewhat different story, but we also see several cells where there seems to be some agreement between the perspectives.

Feedback type	dr	ds	ur	us	backward	forward	shake	side turn	tilt	other
ja	40	28	8	27	6	2	1	0	15	5
m	28	11	21	10	3	0	0	1	0	0
okej	1	1	4	12	4	1	0	0	1	0
nej	0	3	2	1	3	0	7	2	2	1
feedback cluster	26	5	18	18	8	1	1	3	6	6
other repetition	3	1	3	2	1	1	0	1	0	1
all others	18	14	7	27	11	5	3	4	8	3

Table 5. Multimodal combinations of vocal-verbal and head gesture feedback, shaded horizontally.

Feedback type	dr	ds	ur	us	backward	forward	shake	side turn	tilt	other
ja	40	28	8	27	6	2	1	0	15	5
m	28	11	21	10	3	0	0	1	0	0
okej	1	1	4	12	4	1	0	0	1	0
nej	0	3	2	1	3	0	7	2	2	1
feedback cluster	26	5	18	18	8	1	1	3	6	6
other repetition	3	1	3	2	1	1	0	1	0	1
all others	18	14	7	27	11	5	3	4	8	3

Table 6. Multimodal combinations of vocal-verbal and head gesture feedback, shaded vertically.

There seems to be a strong coupling of nods and all positive feedback words. Repeated down-nods are most strongly connected with ‘ja’ and repeated up-nods that are mostly coupled with ‘m’ and feedback clusters. Similarly to what Boholm and Lindblad 2011) found, we see that ‘m’ has a correlation with repeated nods. Boholm and Allwood 2010) found a correlation between ‘okej’ and single up-nods, a result that is repeated here. Head shakes and ‘no’ have a strong coupling, as discussed earlier. We also notice that feedback clusters seem to favor repeated head nods somewhat, and it would be interesting to see whether this is correlated to word repetition within these clusters. In the previously cited study by Boholm and Allwood 2010), no such relation was found, but since that study relied on a fairly small data sat, further investigation would still be interesting. Repeated up-nods show the interesting pattern of being somewhat disassociated from ‘ja’ but closely associated with ‘m’ and clusters, raising the question of whether these clusters have ‘m’ in them, or if there is something else going on.

4 Discussion

Even if many of the subtleties of the use of feedback are still unknown, there are some patterns in Swedish communicative feedback that we have noticed re-emerging (e.g. Boholm and Allwood, 2010; Boholm and Lindblad, 2011; Navarretta et al., 2012). Nods are the most common head gestures for feedback, and among them the repeated down nod is the most common, with the single up-nod being the second most common in Swedish feedback. These two nod types show an interesting dissimilarity, in that single up-nods are almost always multimodal, whereas repeated down-nods are the type of nod most often produced unimodally. One reason for this, we hypothesise, could be that the single up-nod is more often used for emphasis or uptake, while the repeated down-nod is more typically used for giving silent agreement. Single up-nods are sometimes used to signal that the information is new or surprising. It is likely that other aspects of the head gestures, such as intensity, are important for their functions in this regard. In order to investigate these kinds of issues, more in-depth qualitative analysis is needed.

Feedback clusters need to be broken down into their components to see if they show any patterns depending on their parts, such as if repeated nods are correlated to repetition of words, if there are ordering effects or dominant words. We also need to look closer at the big lump of 'others' and we acknowledge that more statistical analysis is needed to substantiate our findings. A very interesting challenge is to look into individual variation in this regard.

It is our intention to increase our sample size, as it is somewhat small. However, we are encouraged by the fact that many of our findings replicate what has been found in other comparable studies. We suspect that there might be more order in this chaos than first meets the eye, and this warrants further investigation.

References

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 Martin et al. (eds) Multimodal Corpora for Modelling Human Multimodal Behaviour.

Allwood J., Kopp S., Grammer K., Ahlsén E., Oberzaucher E. and Koppensteiner M. 2007. The analysis of embodied communicative feedback in multimodal corpora: a prerequisite for behavior simulation. Language Resources and Evaluation, 41(3-4), 255-272.

Allwood J., Nivre J., Ahlsén E. 1992. On the semantics and pragmatics of linguistic feedback. Journal of Semantics, 9(1), 1-26.

Allwood J. (ed.). 2000 Talspråksfrekvenser (Spoken Language frequencies). Gothenburg Papers in Theoretical Linguistics S21. University of Gothenburg. ISSN 0281-2847.

Boersma P. 2001. Praat, a system for doing phonetics by computer. Glot International 5:9/10, 41-345.

Boholm M. and Allwood J. 2010. Repeated head movements, their function and relation to speech. In Kipp et al. (eds.) Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality. LREC 2010.

Boholm M. and Lindblad G. 2011. Head movements and prosody in multimodal feedback. NEALT Proceedings Series: 3rd Nordic Symposium on Multimodal Communication, pp. 25-32.

Cerrato, L. 2002. Some characteristics of feedback expressions in Swedish, Proceedings of Fonetik, TMH-QPSR, vol. 44, n.1, 2002, pages: 101-104.

Cerrato L. 2007. Investigating Communicative Feedback Phenomena across Languages and Modalities. University dissertation from Stockholm : KTH, TRITA-CSC-A 2007:3 ISSN-1653-5723 ISRN-KTH/CSC/A--07/03—SE ISBN 978-91-7178-632-6 Format (including language).

Kipp M. 2001. Anvil - A Generic Annotation Tool for Multimodal Dialogue. Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech), pp. 1367-1370.

Navarretta C., Ahlsén E., Allwood J., Jokinen K., and Paggio P. 2012. Feedback in Nordic First-Encounters: a Comparative Study. Proceedings of the Language Resources and Evaluation Conference 2012, 494-2499.

Nivre J. 2004. Göteborg Transcription Standard. (GTS) V. 6.4. Department of Linguistics, University of Gothenburg.

Multimodal Human-Horse Interaction in Therapy and Leisure Riding

Nataliya Berbyuk Lindström **Jens Allwood** **Margareta Håkanson** **Anna Lundberg**

SCCIIL Interdisciplinary Center
Department of Applied IT
University of Gothenburg, Sweden

Department of Veterinarian Medicine and Domestic Animal Science
Swedish University of Agricultural Sciences

berlinds@chalmers.se jens@ling.gu.se Margareta.hakanson@comhem.se anna.lundberg@slu.se

Abstract

Horseback-riding in general, and equine-assisted therapy in particular, are widely used for leisure and rehabilitation purposes. However, few scientific studies on human-horse interaction are available. The aim of this article is to provide a description and analysis of multimodal human-horse interaction in riding sessions. Video and audio-recordings of riding sessions, interviews with the riders and observations were done in a small riding school in Western Sweden. A combination of linguistic and ethological methods is used for data analysis. The recordings are transcribed, and the sequences when human-horse interactions occur are analysed using activity-based communication analysis and ethograms. The following typical sub-activities of riding session are distinguished and considered: "greeting horse", "care of horse before riding", "tacking", "mounting horse", "waiting for co-riders," "riding lesson", "dismounting horse", "care of horse after riding" and "saying goodbye to horse." The analysis shows that the riders use vocal verbal, visual and tactile signals when they communicate with the horses. The riders tend to communicate more verbally while caring before the ride compared to after the riding lesson. The horses' reactions are complex, comprising tactile (e.g. touch with the muzzle), visual (e.g. lifting legs, moving in the box/stable, ear and head movements, movements of the tail, etc.) as well as auditory ones, (e.g. snorting).

1 Introduction

The relationship between humans and horses has a long history. Historically, horses were kept for meat and used for transportation (Brown and Anthony, 1998; Anthony and Brown, 2000; Levine, 2005), while today they are used for leisure, sport, as working companions in rural areas and for equine-assisted interventions (for therapy and learning).

Contacts between humans and horses are not unproblematic; accidents occur among both professional riders and laymen. Research shows that the occurrence of accidents depends more on the frequency and amount of interactions between humans and horses than on the level of riding competency (Hauseberger et al 2008). Thus, a better understanding of the human-horse relationship in general, and their interaction in particular is needed to enhance safety and quality in human-horse contacts.

At this moment, little research is available on human-horse interaction in general, and even less

K. Jokinen and M. Vels. 2015. Proceedings of The 2nd European and the 5th Nordic Symposium on Multimodal Communication.

with a focus on human communication is particular. In this article, we provide a step toward a better understanding by an analysis of aspects of the multimodal human interaction with horses in riding sessions used for both therapy and leisure riding.

2 Background

A few studies, though no linguistic studies, on human-horse interaction are available. The majority of them are in the field of ethology and focus on horse behaviour (Hausberger et al, 2008) as well as in the health sciences (Keaveney, 2008), biological sciences, agriculture (Birke et al, 2011), and sport sciences (Münz et al, 2014) to mention a few.

Human-horse interaction is complex. The reactions of horses to humans are mostly the result of interplay between the temperament of the horses, the temperament and skills of the humans and the experience of the horses acquired with humans. This means that such factors as the personality of horses and humans, the horses' positive/negative experiences in interaction with humans, e.g. being mistreated, are important.

Human-horse interaction is multimodal, which means that at least two of the sensory modalities (vision, hearing, touch, smell and taste) are involved. Visual and tactile communication is central. A study performed on emotional cues shows that when people have negative feelings towards animals, while stroking a horse, they induce an increase of heart rate in the horse in the first few minutes. "Neutral" or "positive" persons do not have such an influence (Hama et al., 1996). Chamove et al. (2002), who performed a study on the effect of human attitude on horse behaviour, suggested that human attitude correlates with the horse behaviour when led through a predefined course. Below we will now consider some studies reporting on communication with horses using different sensory modalities.

Seaman et al. (2002) did not find an influence of the direction of gaze of the human on the reactions of horses; there was no difference between a person approaching with or without visual contact. Similarly, Verril et al (2008) studied whether having direct eye contact or avoiding eye contact with horses influenced how easily they were captured in the pastures. No differences between these behaviours were found.

Testing cross-modal recognition of horses, Sankey et al (2010) conclude that a horse's recognition of humans is based on a multimodal combination of vocal and visual identification, suggesting that horses have a "concept of person".

Concerning touch, Hausberger et al. (2008) point out that gentling (being patted or stroked) is not necessarily rewarding for animals, but that instead positive reinforcement using food has been shown to be connected with a positive association for many animals including horses. These findings raise questions concerning whether patting horses is relevant and whether it can even be contra productive when intending to give positive feedback.

Hausberger et al (2008) suggest that by training people to observe the body postures of horses, be attentive to their signals and attitudes as well as to avoid anthropomorphic interpretation of behaviour, horse-human interactions will be safer and accidents due to misunderstandings can be avoided.

Continuing to the auditory communication of horses, familiar horses in the herd were paired with the "wrong" sound when the horse was out of sight. The other horses reacted when a mismatch was at hand. The researchers' conclusion was that horses recognize known individuals by auditory, visual and olfactory information (Proops 2010).

Also, human communicative behaviour towards horses in equine assisted intervention has been studied with a focus on whether horses respond to the non-vocal expressions such as body posture, movements and orientation of humans. The understanding of the communicative aspects of human non-vocal behaviour was based on psychological/psychoanalytical theories (Zink 2008). Garcia (2010) mentions the ability of horses to respond to human behaviour as the core reason for the therapeutic use of horses in treating humans with mental disorders. In order to understand the communication between horses and humans it is essential to expand our understanding of the signalling mechanisms of horses, she claims. Working with horses enhances the ability of humans to use their body as a sensing mechanism which appears to go beyond learning to grasp body language, toward developing more intuitive levels of awareness and environmental scanning skills. Garcia asks for more research addressing "how humans and horses communicate and how they learn with/ from each other" (ibid).

With the exception of the studies mentioned above, we have found no research, which describes and analyses a multimodal combination of verbal and non-verbal tactile, visual and auditory communication in human – horse interaction from the human perspective. This study is a contribution to filling this gap. In addition, we try to make a contribution to developing methods for studying human interaction with horses.

3 Method

The data has been collected within an interdisciplinary research project, "Ethnicity and human-horse interaction", financed by Region West Sweden (Västra Götalandsregionen) in 2010-2014.The study focuses on communication between humans and horses during a riding session.

All data were collected in a small riding school situated in a rural area close to Gothenburg, Sweden. The riding school aims at riders with and without functional disabilities and provides riding sessions for both treatment and leisure. The context of communication between humans and horses has been stable over time as most of the staff at the riding school has been permanent for several years, as have most of the riders and therapists involved in the activity. A riding session as a social activity in the school comprises the following sub-activities: "greeting horse", "care of horse before riding", "tacking", "mounting horse", "waiting for co-riders," "riding lesson", "dismounting horse", "care of horse after riding" and "saying goodbye to horse."

3.1 Data collection and participants

The data comprises video and audio-recordings of riding sessions, interviews with the riders and observations of other interactions between humans and horses.

Seven (7) participants, all female, age range 11-20 y.o.a, one with paralysis, 5 native Swedes and two 2nd generation immigrants (Danish and Russian) were video- and audio recorded during their riding sessions. Total recording time is 120 minutes.

Seven audio recordings (480 min) cover a total time spent with the horse, while video recordings cover all activities but the sub-activity "riding lesson", due to technical difficulties in getting video recordings here. Instead, the riders were observed. Their horses were 8-24 years old, 3 geldings and 4 mares, and the riders had previous experience from riding these horses.

Interviews with three of the informants regarding their views on their communication with the horses were carried out.

Forty hours of observations in the school were documented using field notes.‒

3.2 Data analysis

In the project, we use a combination of linguistic and ethological methods for data collection and analysis. Activity-based communication analysis (Allwood 2000, Berbyuk-Lindström 2008) was used to analyze the communication in a riding session as a social activity.

The video and audio recordings were transcribed using Gothenburg Transcription Standard. The transcription conventions are presented below:

Symbol	Explanation
R, H, A	participants (e.g. rider, horse, assistant)
[ ]	overlap brackets; numbers used to indicate the overlapped parts
( )	transcriber's uncertain interpretation of what is being said, e.g. (jobbi)
/, //, ///	a short, intermediate and a long pause respectively
+	incomplete word, a pause within word
CAPITALS	contrastive stress
:	lengthening
<>, @ <>	comments about non-verbal behaviour, comment on standard orthography, intonation, other actions, clarifications

Table 1. Transcription conventions

Visual, tactile and auditory (often verbal) communication between the riders and horses is analysed in relation to the sub-activity in which the communication occurs. The MUMIN coding scheme (Allwood, J. et al. (2007)) is used to analyse the data. Adapted ethograms from McGreevy (2012), see appendix, are used for analysing the behaviour of horses. For the horses, the neck, head and ear position, tail movements, posture, feet position and movements towards and from the humans as well as sounds were coded. Direction of gaze and eye contact and the responses from the horse to tactile and auditory signals were also recorded. Horse movements were not always easy to observe. However, all codings were checked by another coder. Unfortunately, time has not allowed for a more formal reliability control.

From the interviews, the comments of the riders concerning their communication with the horses are analysed and can be combined with the analysis of the recordings. However, in this paper, our analysis is based on an analysis of the video recordings, we hope to return to the interviews in a later paper.

Notes from the observations are used to provide additional insights into the interactions and the settings of the riding sessions.

4 Ethical considerations

The study has been approved by the ethics committee of Västra Götaland, regarding both human and animal welfare. All riders in the study volunteered to participate. They were given information, in written and oral form, and signed a consent form. For those under fifteen years of age, consent by parents has been given as well. The horses were recorded in a riding school, in their normal environment. No invasive methods have been used. When possible, the researchers were not present during recordings in order not to disturb the activities.

5 Results

First, an overview of the sub-activities of a riding session is provided, followed by general comments about communication based on the interviews. Next, an analysis of communication in each sub-activity is provided. Finally, a general summary that presents an overall picture of communication in the riding sessions is provided.

5.1 Analysis of the riding session as a social activity

The **purpose** of the riders to come to the school is riding itself as well as contact and caring for the horses (depending on the physical ability of the riders). In leisure riding, the riders' **goal** is primarily to improve riding skills and to have fun. In the case of therapy, enhancing physical, emotional and social well being is central. Leisure riding and therapy goals often overlap.

The people present during riding sessions include riders, groom/assistants, riding instructors/therapists, fellow riders, and accompanying persons. The horse is selected and placed in the box/tie stall in advance. The name of the rider and the paired horse is on a list on the wall. The rider is expected to check which horse to take; in the case of a disabled rider, the riding instructors/therapists do this.

A typical riding session starts with the riders coming to the stables and greeting the horses they will ride (sub-activity "greeting horse"). In case of leisure riding, they start by caring for the horse, which includes brushing and cleaning the hoofs ("care of horse before riding"), which is followed by the subactivity "tacking".

When ready, the riders leave the stable and go to the mounting block, where they mount the horses with/without help from the staff ("mounting horse"). In therapy, only some of the disabled riders can (partially) carry out the care; if they are not able to do this, they go directly to the mounting block and get help with mounting. When all riders are mounted ("waiting for the co-riders to mount"), the instructor gives a signal and all involved leave for the riding lesson, which is conducted in the arena or in the nearest wood to the track. The riding lasts for 30 - 45 min (sub-activity "riding lesson").

After the lesson, all return to the yard and wait on horsebacks till dismounting is allowed ("waiting for the co-riders to dismount"). Dismounted (with/without assistance) and on ground, they (or assistants) loose the girth and the stirrups are placed in a secure position ("dismounting horse"). The rider approaches the horse and pats it, sometimes also gives it a carrot or piece of bread as a thank you and farewell sign. On a signal from the instructor, the horse is led from the yard into the stable by the riders or assistants. The disabled riders usually thank the horse and say goodbye to everyone before the horse is led to the stable. The healthy (and some disabled) riders usually put a halter around the horses' neck, remove the bridle and the saddle. The halter is properly put on the head and the horses' body is brushed, the feet are cleaned, the bit is rinsed and the tack is removed ("care of horse after riding"). After saying goodbye the horse is left in the box/tie stall and the riders leave the premises (sub activity "saying goodbye to horse").

5.2 Communication in riding session

General comments about communication from the interviews and observations

In the interviews, the riders commented that their relationships are different with different horses; in some cases the riders experience that their personalities match ("we are soul mates" who "understand each other"), in other they don't. Our observations indicate that riders who were fond of their horses tended to communicate more with them compared to other riders.

In general, positive feelings by the riders are experienced when coming to the stables ("a wave of joy and happiness"), and the horses are believed to be sensitive to such riders when they come to the stable. Further, the riders' personalities and the nature of a disability influence how much the riders can communicate with the horses. The staff in the riding school often encourages the riders to communicate with the horses, e.g. asking the rider to thank the horses for the riding lesson.

In the stable, many people are present and there is hardly any privacy between rider and horse. The most private time with the horse is during caring before and after riding, and this is the time when most communication is reported and observed to occur. Below we will now exemplify and discuss communication with the horses from the point of view of the rider. We will describe what communicative acts occur both functionally and behaviourally. We will also exemplify how the joint activities of riders and humans influence the kind of communication that occurs. In future work we hope to examine the same interactions from the communicative point of view of the horse.

Greeting a horse – a combination of tactile and auditory signals

A common strategy when greeting the horse is using one or more clicking sounds to attract horse's attention, which is followed by "hej" (hello)/"heej" (heello), often in combination with the horses' names or nicknames, e.g. "Carat, fröken" (Carat, miss), "gumman" (baby), "stumpan" (honey, baby), "fröken skiten" (miss muddy). Vocal (often verbal) greetings are often combined with patting the horse's neck, kissing the horse (often on the muzzle, throat and neck) and stroking the mane. In addition, some riders enquire how their horses feel, e.g. "älskling, hur är det?" (honey, how are you?). Other riders are observed (or report) to just pat their horses on the back and hindquarters. In the example below, the greeting activity is presented. In other words, there are considerable degrees of freedom for how to accomplish the greeting and no really narrow script can be observed.

The horse stands and rests on a back leg with his head turned towards the door of the box	
$H7:	Changes rest leg
The rider enters the box and stands in front of the horse looking at him	
$R7:	< Hej hästen > *< Hello horse >*
$H7:	H7 stands still
$R7:	< Ariel > *< Ariel >*
@ < horse's name >	
$H7:	stands still
$R7:	The rider goes out of the box

Example 1. R7 - Rider 7, H7– Horse 7

In the example above, the rider greets the horse prior to entering the box/tie stall, saying "hello horse" and mentioning the name of the horse. Seemingly, the horse doesn't react to the greeting.

Care of horse before riding

The care of a horse consists of brushing and cleaning the hoofs. Communicative acts observed in this sub-activity are the following ones:

- **giving orders to horses**, e.g. moving to another side of the box/tie stall to give space for the rider to brush, to stand still, to lift the hoofs, etc.
- **expressing approval/disapproval to horses and gratefulness**, e.g. when the horses stand still and allow the riders to brush them, follow their orders or not
- **riders commenting their actions to the horses**, e.g. starting/finishing brushing or cleaning the hoofs
- **riders commenting on the horses appearance and giving compliments**, e.g. being dirty
- **more or less unconscious patting/stroking the horse**

Instructions to the horse: To manage brushing and cleaning the hoofs, the riders need to move around the horse in the box/tie stall. They need to instruct the horse, to move to one or another side of the box. A common way of doing this is to come from behind using clicking sounds to draw horse's attention in combination with pushing the horse with the force of the whole body, with a hand or with finger(s) on the croup, hindquarters, buttock or thigh:

R4 finished brushing the left side and the tail of H4. Coming from behind, she wants to start brushing the right side of H4. She asks H4 to move to the left, to give her space.	
$R4	< Clicking sound twice >
< coming from behind, right hand stretching towards H4's back >	
$H4	Neck up, ears up, head turned to the left
$R4	< clicking sound twice >
@ < slightly pushing the H4's buttock, approaching to H4's head from the right >	
$H4	Moving left
@R4:	< clicking sound >
< pushing H4's shoulder with her bent right hand finder >	
$H4	Moving to the left and backwards
$R4	Slightly patting H4's nose and start brushing H4's head

Example 2. R4 - Rider 4, H4 – Horse 4

In the example above, the horse reacts and moves to the left which gives space to the rider to brush the horse.

Other examples include asking the horse to stand still e.g. "ska du vara snäll idag" (will you be nice today), "Carat – ta det försiktigt – bra" (Carat be careful good).

Expressing approval/disapproval and gratefulness: The riders sometimes express approval, e.g. ("bra Carat bra det var inte farligt" (good Carat it was not dangerous), "bra" (good) in combination with stroking and patting the horses one or more times, often on the relevant part of the body, e.g. legs if the rider asked the horse to lift them. Disapproval is expressed by using "nej" (no) one or more times in a raised voice, often in combination with the nickname, e.g. "nej Carat" (no Carat) when the horse disobeys. Pushing the horse from oneself can also be observed. Dissatisfaction with the horse's behaviour is expressed, e.g. "O jä…va du skvätter ner dig fröken" (oh damn you splash yourself miss). The riders also express gratefulness if the horse is obedient (e.g. "SÅ – tack så mycket – nästan" (SO thank you so much almost). They also ask if the horse likes cleaning, e.g. "Är det skönt" (is it nice). Patting the horse is common.

Commenting their actions: The riders tend to comment on their actions. Clicking sounds are often used to draw attention to new actions, followed by a question to the horse, e.g. "Kan vi tränsa nu?" (can we bridle now), "Ska vi ta hovarna fröken? Ska vi ta hovarna" (Shall we take the hooves miss? Shall we take the hooves), "O nu skall vi borsta dig" (oh now we will brush you). Finally, when they finish the action, they pat the horse to signal this and that the horse has behaved well.

The riders also **comment on the appearance of the horses and give compliments**, e.g. “Du e söt”, “du e fin”, “å va du e gosig”, “du är så duktig” (you are so sweet, you are so nice, oh you are so cuddly, you are so clever) in combination with smacking sounds.

More or less unconscious patting/stroking the horse is common in this sub-activity. While brushing, the riders use one hand, another lying on the horse’s side. Often the riders more or less unconsciously pat and stroke the horse.

Tacking

While tacking (taking off the halter from the horses head and putting the halter around the horses’ neck), putting on the bridle and the saddle, less auditory communication is observed. It is worth mentioning that while tacking, the horse’s head is close, which results in the riders patting, stroking and kissing the horses’ face, poll, mane and nose.

Mounting the horse and waiting for co-riders to mount

While mounting, the assistants often help the riders. Here, both riders and other people tend to communicate with the horses. The riders ask their horses to come closer (“kom” (come)), in combination with smacking, clicking and whistling. They show **disapproval** when the horse doesn’t stand still, which makes it difficult to mount (“Ja men DU!” (well but YOU)) and **disappointment** (Ahh FRÖKEN (oh MISS)).

Some auditory communication with the horse is observed while waiting. Stroking the horse’s mane and neck, is common. Asking him/her to stand also occurs (“Duktig fröken, duktigt, Vi ska inte gå ut än fröken,””Nu skall vi snart gå fröken”, “Ta det lugnt gumman” (clever miss, clever, we will not go yet, now we will soon go, take it easy)). Both louder voice and whispering, e.g. “sluta” (stop whispering) and stop (“Proo”) are used.

Riding lesson

The sub-activity the “riding lesson” for riders is divided into transport to the riding hall, the lesson itself (warm up time, riding tasks and cool down time) and transport back to the yard. During the transport and warm up/cool down, rider and horse mostly communicate in a tactile manner. The pace is walk. During the riding tasks, the riders follow the directives from the riding instructor and are focused on accomplishing the tasks in trot and in canter. The physical level of activity is high. Since the riders are supposed to make their horses follow what the instructor says, clicking sounds, pressing the legs to horses’ sides and body balancing are used as means to do this.

When the lesson is finished, the riders are often physically tired. Some riders are humming and patting/stroking the horses to thank them for the ride. Others are quiet. The horses with the riders walk to the mounting area one after another.

Dismounting the horse

Dismounting starts when all riders are safely positioned in the yard and their horses are standing still. The riders hug or pat the horse on the shoulder or neck, then dismount with or without help. Especially when riders have disabilities the instructors ask the riders to thank the horses and pat them. Often the horses get carrots or apples. When the rider stands on the ground, the horse turns its head towards the rider and sniffs. The human loosens the girth and place the stirrups on the top of the leathers, first on one side, then on the other walking around the horse’s front. When the equipment is adjusted, the horse is lead into the stable.

Care of the horse after riding

After riding, care of the horse is done primarily by those riders, who can manage it physically. Being led into the stall/box, the horse is being untacked, halter is put on and tied to a lead-rein by an instructor or the rider. The horse’s back and stomach (where the saddle was put) are brushed. If the horse is sweaty, it is cooled down with water from a sponge. The legs are controlled for wounds and the hoofs are checked for stones and dirt to be removed.

When the horse is checked out the lead-rein is hooked of and the horse is left to rest in its tie stall/box. Not much auditory communication can be observed during this sub-activity. Often riders are tired. Similar patterns, as in brushing before the riding lesson, can be observed in care of the horse after riding.

In the example below, the communication between the horse and the rider can be observed:

$R7:	begins brushing right hind leg
$H7:	the horse moves a bit
$R7:	< Ähh sluta > < *Ahh stop* >
@ < irritated >	
$R7:	proceeds brushing right hind leg
$H7:	H7 lifts the right hind leg three times
$R7:	The rider stands up and sighs. Brushes bit on the back and then goes over to the other side and brush where, towards the back hind leg.
$H7:	H7 takes a step to the right
$R7:	< Näe > < *neeh* >
@ < irritated >	
$R7:	The rider stands up, walks up to the head of the horse and starts brushing the left side

Example 3. R7 - Rider 7, H7 – Horse 7

In example 3 above, the horse shows disobedience and doesn't follow the rider's requests which results in irritation.

Saying goodbye to the horse

The rider says goodbye by using a combination of vocal verbal and tactile signals by a pat on head, shoulder or bottom. Usually, riders are in a hurry to go home, they talk to their co-riders and can be observed to show less interest in their horses.

6 Discussion and conclusions

Our study is qualitative and it is based on a limited amount of data. Thus, conclusions have to tentative. In general, the study shows that communication between humans and horses is complex and has different features in different stages of the riding session in terms of the communicative means used, their functions and intensity. Human-horse communication is multimodal, comprising both bodily and verbal communication. The riders tend to use verbal greetings, without or in combination with the names (nicknames) of the horses to greet their horses, sometimes even asking how the horses feel, which resembles human-human greetings. A different way of greeting is tactile by patting the horse on the back and hindquarters, which resembles humans patting an interlocutor's shoulder. A possible reason for choosing patting rather than talking can be that the riders approach the horses from behind, and patting the horse might be more natural as a better attention-getting strategy as many people are present in the stables, it can be quite noisy, and the horse might not pay attention to the greeting (in the way a human would in the same situation).

Most of the auditory and tactile communication between riders and horses occurs while taking care of horses before riding which includes brushing the horse and cleaning the hoofs. A possible reason for this can be that in this sub-activity the rider and the horse are on the same level (both standing on the floor), which makes this kind of communication possible. In addition, it is the most private part of the riding session, as no other people are usually involved. Further, the fact that the riders often get different horses for each session might necessitate a need and desire for creating a relationship with the horse. Another factor is that the riders are anticipating the riding session and are willing to make contacts with horses.

While caring for horses before riding, the riders are carrying out certain tasks, such as brushing the horses' body and cleaning the hoofs, which necessitates using instructions to the horses to move to give space, to lift legs or to stand still. Clicking sounds, patting and pushing the horse are used, often resulting in the horse doing what the riders want (Example 2) or not (Example 3). It can be observed that the clicking sounds are used to attract attention, while touch (pushing and patting) give a more specific and a stronger signal to horses, e.g. patting on the leg is a signal to lift the leg, etc. In the data (Example 2), we can observe the reactions of a horse to the rider's actions and, compared to Example

1, a reaction from the horse can be observed. Possibly, the horses react more strongly to tactile contact than they do to only vocal cues.

An interesting type of behaviour is that the riders are commenting their actions to the horses and asking for permission to brush or pick up hooves. Similar patterns can be observed in e.g. medical consultations and child-parent interactions, when physicians/parents comment their actions to patients/children in order for them to be calm, prepared to and informed about what is going on. It probably reflects a somewhat superior and caring attitude to the horses from the riders' side, which is even present in the riders' expressing approval/disapproval and commenting on appearance/giving compliments. Smacking sounds are used, which is a way to show affection.

More or less unconscious patting/stroking of the horse is also common in this sub-activity. While brushing, the riders use one hand, another lying on the horse's side. Often the riders more or less unconsciously pat and stroke the horse. It is unclear if the riders are attempting to calm the horse down or just are automatically leaning on the horse.

Proximity to the horse's head is closest while tacking, which makes it possible for the riders to touch the head and, as can be observed from the data, to pat, stroke and kiss the horses' face, poll, mane and nose.

Presence of other people seems to influence the frequency and intensity of communication between the riders and their horses. Communication can be observed while mounting and dismounting the horse and during the riding itself. While mounting/dismounting, instructions given by the riders to the horses to stand still/come close can be observed. While waiting for the lesson, the riders mainly try to calm the horses down. Verbal cues and stroking the horse's mane and neck are used, which resembles a human way of calming another human being. Interesting that both whispering to horses and exclamations are used, which probably reflects a more or less caring attitude to horses and a need for more or less strict orders. After the lesson, the physically able riders get down with/without help and often just silently take their horses to the stable to brush them. It is interesting why so little auditory communication is observed after riding, probably due to the riders being tired and even a feeling of an accomplished task, which does not necessitate any further cooperation.

Little verbal communication is observed during riding, as the instructor provides the orders the riders should follow. Audio-recordings show that clicking sounds are used, probably because they can be heard by the horses, also pressing the legs can be observed. The horse is mainly seen as a tool for carrying out the tasks. In general, functional requirements concerning both the communicative acts used by the riders and the way in which the joint activities between horses and riders influence the communication leave considerable degrees of freedom for how to accomplish these functions. This means that riders show a fair amount of variation in how they communicate with their horses. No real conventional "scripts" seem to have developed, instead local circumstances play a large role for how a given function is communicated mostly using a combination of vocal sounds (often verbal) and tactile signals. As far as we have been able to see vision is less important. We have not really been able to investigate the role of smell and taste both of which no doubt also play a role.

Acknowledgements

The authors wish to thank the riders and the staff in the riding school for participation in the project and Region West Sweden (Västra Götalandsregionen) for financing of the project

References

Anthony, D.W., Brown, D.R. 2000. Eneolithic horse exploitation in the Eurasian steppes: diet, ritual and riding. *Antiquity 74, 75–86.*

Allwood, J. 2000. An activity based approach to pragmatics. In: H. Bunt & B. Black (Eds.), *Abduction, belief and context in dialogue: Studies in computational pragmatics* (pp. 47-80). Amsterdam: John Benjamins.

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

Berbyuk Lindström, N. 2008. *Intercultural communication in health care. Non-Swedish physicians in Sweden*. Gothenburg: Department of Linguistics, University of Gothenburg.

Birke, L., Hockenhull, J., Creighton, E., Pinno, L., Mee, J., Mill, D. 2011. Horses' responses to variation in human approach. *Applied Animal Behaviour Science, 134*, 56–63.

Brown, D.R., Anthony, D.W. 1998. Bit wear, horseback riding, and the Botai site in Kazakstan. *Journal of Archaeological Science, 25, 331–347.*

Chamove, A. S., Crawley-Hartrick, O. J. E., Stafford, K. J. 2002. Horse reactions to human attitudes and behaviour. *Anthrozoos, 15*, 323-331.

Garcia D. 2010. Of Equines and Humans: Toward a New Ecology. *Ecopsychology,, 2(2),* 85-89.

Hama, H., Yogo, M., Matsuyama, Y., 1996. Effects of stroking horses on both humans' and horses' heart rate responses. *Japanese. Psychology Research, 38*, 66–73.

Hausberger, M., Roche, H., Henry, S., Visser, E. K. 2008. A review of the human–horse relationship. *Applied Animal Behaviour Science*, *109,* 1-24.

Keaveney, S. M. 2008. Equines and their human companions. Journal of Business Research 61, 444-454.

Levine, M.A. 2005. Domestication and early history of the horse. In: Mills, D.M., McDonnell, S.M. (Eds.), *The Domestic Horse: The Origins, Development, and Management of Its Behaviour.* Cambridge University Press, Cambridge, pp. 5–22.

McDonnel, S. 2003. *A practical field guide to horse behavior.* Hong Kong: Eclipse Press, Hong Kong.

McGreevy, P. 2004. *Equine Behavior: A Guide for Veterinarians and Equine Scientists*. China: Saunders.

McGreevy, P. 2012. *Equine Behavior: A Guide for Veterinarians and Equine Scientists*. China: Saunders.

Münz, A., Eckardt, F., Witte, K. 2014. Horse–rider interaction in dressage riding. *Human Movement Science, 33*, 227-237.

Proops, L., McComb, K. 2010. Attributing attention: the use of human-given cues by domestic horses (Equus caballus). *Animal Cognition, 13*, 197-205.

Sankey C., Henry S., Go´ recka-Bruzda, A., Richard-Yris, M.-A., Hausberger, M. 2010. The Way to a Man's Heart Is through His Stomach: What about Horses? *PLoS ONE 5(11),* e15446

Seaman, S., Davidson, H., Waran, N. 2002. How reliable is temperament assessment in the domestic horse (Equus caballus)? *Applied Animal Behavior Science, 78*, 175–191.

Verrill, S., McDonnell, S. 2008. Equal outcomes with and without human-to-horse eye contact when catching horses and ponies in an open pasture. *Journal of Equine Veterinary Science, 28*, 309-312.

Young, T., Creighton, E., Smith, T. and Hosie C. 2012. A novel scale of behavioural indicators of stress for use with domestic horses. *Applied Animal Behavior Science, 140,* 33-43.

Zink R. 2008. Do horses co-communicate? Inbetween the worlds of horse and human. Insights to results of behavioral research of horses expressive behavior. Proceedings from Mench unf Pferd in Dialog. EU-konferenz. Wien 2008.

Appendix

Table 2. Description of different horse behaviours registered from the video recordings. Adapted from McDonnel (2003), McGreevy (2004) and Young *et al*. (2012).

Name	Description
Alert	Horse stands with neck elevated and eye level elevated above height of withers
Relaxed	Horse stands with head and eye level lowered under the wither, often with one back hoof lifted and weight distributed among only three legs
Yawn	Deep long inhalation with mouth widely open and jaws either directly opposed or moved from side to side
Approaches	Horse takes one or more steps towards human
Backs of	Horse takes one or more steps away from human
Attention toward	Head and/or ears directing towards human
Attention away	Head or/and ears directed away from human
Push	Pressing of the head, neck, shoulder chest, body or rump against the human, equipment or interior design
Tail swish	Tail is flicked to one side and/or the other of the quarters
Ears flat	Ears pressed caudally against head and neck
Bite threat	Neck is stretched and ears are pinned back, the jaws are opened and closed rapidly as the head swings towards, without biting, the target
Kick threat	One or both hind legs are lifted slightly off the ground without subsequent backward extension
Bite	Ears pinned, lips retracted and jaws are opened and rapidly closed with teeth grasping clothes or skin/flesh of human
Kick	One or both hind legs lift off the ground and rapidly extend backwards
Eating from ground	Horse is foraging with the head lowered at the ground
Eating from hay net	Horse is foraging from the hay net
Eating from hand	Horse takes food item with the muzzle from the human hand, chews and swallows
Drinking	Horse immerse lips in the water bowl
Exploratory	Lick, sniff or touch with muzzle or tongue
Pawing	Strikes a vertical or horizontal surface, or the air, with a front leg
Defecation/urination	Elimination of faeces and urine
Lifting hoof	Horse lifts hoof of the ground as the human touches that leg with the hand and/or shoulder
Follows	Horse walks forwards in a four beat gait after or beside the human when led in the reins

Finding Appropriate Interaction Strategies for Proactive Dialogue Systems—An Open Quest

Florian Nothdurft, Stefan Ultes, Wolfgang Minker
Institute of Communications Engineering
University of Ulm
Ulm, Germany
{florian.nothdurft,stefan.ultes,wolfgang.minker}@uni-ulm.de

Abstract

In this paper we elucidate the challenges of proactiveness in dialogue systems and how these influence the effectiveness of turn-taking behaviour in multimodal as well as in unimodal dialogue systems. Effective turn-taking is essential for a natural and qualitatively high human-computer interaction. Especially in spoken dialogue systems, analysing whether the dialogue system should or could take the floor, seems to be an important process in the overall perceived quality of the interaction. Additionally, as technical systems get increasingly complex and evolve in the direction of intelligent assistants rather than simple problem solvers, proactive system behaviour may influence the perception of the ongoing dialogue between human and computer. Autonomously made decisions or triggered system actions may surprise or even disturb the user, which may result in a reduced transparency of the technical system. Therefore, the decision if, when and how to take the floor in a proactive system yields additional challenges. We discuss each layer of decision-making and explain how multimodal cognitive systems can help to control this decision-making in a valuable fashion.

1 Introduction

For spoken human-machine dialogues, the system decision of when to talk poses an important question. While this is usually an easy task for humans, a technical system is not yet able to analyse the complexity and nuances of a conversation. Hence, turn-taking strategies have been developed. State-of-the art interactive voice response (IVR) systems and spoken dialogue systems (SDS) usually use a predefined threshold to decide whether the user is willing to yield the floor. This simplistic approach leads to an unsatisfying and confusing user-experience, for example, because the user is interrupted by the system's re-prompting while thinking and trying to understand what the system expects (Ward et al., 2005; Raux et al., 2006). They also state that sometimes system time-outs are too long, leading to unusual and as awkward experienced waiting periods. Then both phenomena combine, this may lead to parallel attempts to take the floor. Hence, most recent research focuses on a more human-like approach to manage turn-taking behaviour in an SDS, for example by using automatically extractable features to inform efficient end-of-turn detection, and use this amongst other factors to train a turn-taking decision model based on decision theory (i.e., using statistical models), leading to significantly better results than fixed-threshold approaches (e.g., (Raux and Eskenazi, 2012)).

However, technical systems have evolved since the past decade from simple task-solving systems to technical companion systems (Honold et al., 2014) which solve tasks of increasing complexity cooperatively with the user. Hence, as the capabilities of such systems increase, it seems natural that technical systems will take over some of the responsibilities from the user and become an assistive system and life companion. To achieve this, these systems must also be able initiate interaction and not only react to the user. This will also lead to a more complex problem of turn-taking. While for conventional systems, only the question of when to take the floor is of interest, proactive agents also have to decide how to act

K. Jokinen and M. Vels. 2015. Proceedings of The 2nd European and the 5th Nordic Symposium on Multimodal Communication.

and whether to act at all. Here, multimodal systems have a significant advantage over unimodal systems as they are able to exploit more cues about the interaction to make their decisions. This strategy also reflects human behaviour. It has notably been shown that human turn-taking not only depends on a various number of language cues but also on non-verbal cues like gesture or gaze (cf. (Duncan, 1972; Sacks et al., 1974; Gravano and Hirschberg, 2011)).

Hence, in this contribution, we describe and analyse the challenges of turn-taking for proactive agents in multimodal interaction and identify those key issues which have to be solved along the way to foster a healthy and sound human-computer interaction. In the next section we will elucidate the concept of proactive system behaviour followed by a description of our use-case at hand in Section 3. Section 4 will then discuss the resulting challenges for each layer of decision-making to give guidance for a future solution processes.

2 Proactive Behaviour

Proactivity in technical systems is an autonomous, anticipatory system-initiated behaviour, with the purpose to act in advance of a future situation, rather than only reacting to it. Therefore, for our research, we consider proactive behaviour as induced by implicit information and not by any kind of direct or explicit user interaction or user-made adaptation criteria. This means, for example, that user defined temperature values for a room, and the automatic adaptation to this preference when entering this room, do not count as proactive behaviour. Contrary to that the implicit sensing, e.g. by measuring body temperature using infrared sensors, that the user is feeling cold and the system's reaction to that by increasing the room temperature may be considered proactive. Respectively, the change of the user-interface modality to the user's characteristics may not be regarded as a proactive but an adaptive system behaviour. Therefore, only implicit reasons for proactive behaviour recognized by a cognitive system (Figure 1)—sensing a user's affective state for example—and the subsequent system actions may fulfil the requirements of proactive behaviour.

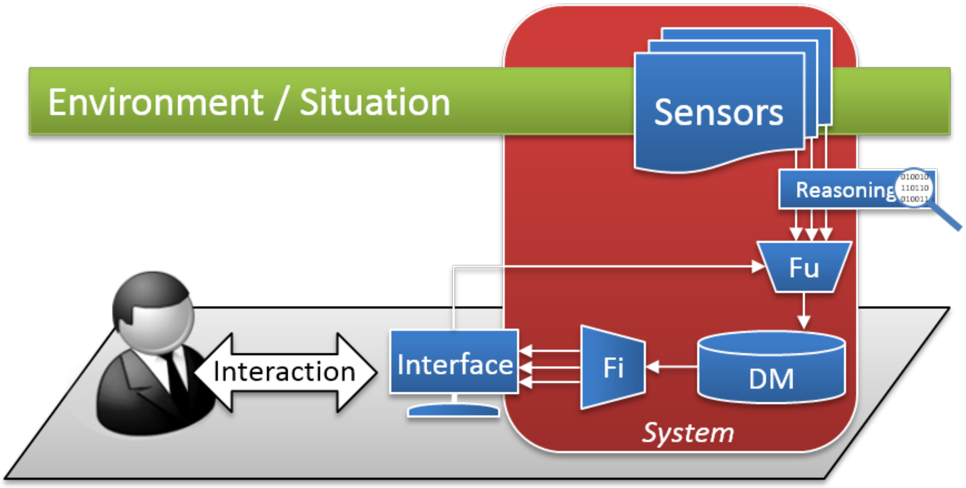


Figure 1: A typical architecture of a cognitive system. Only reasoned information coming from implicit interaction information (e.g., observations of the sensors) trigger proactive behaviour. *Fi* stands for Fission, which controls the modality arbitration. *DM* for Dialogue Management, which controls the flow of the dialogue and *Fu* for Fusion, which merges all input modalities to one consistent semantic representation.

3 Application

Proactive behaviour may occur in many different settings. In this work, we focus on proactive behaviour in a mixed-initiative system combining planning with dialogue. The research field of automated planning (e.g., (Biundo et al., 2011)) and scheduling deals with the development of methods and techniques to automatically and autonomously create solutions, mostly action sequences, which will help a user or an autonomous system to achieve a predefined goal. The user proposes a goal to achieve and thereafter the system tries to come up with a solution. Such an autonomous process usually involves the risk of an unsatisfying or confusing user-experience. The user has no saying during the planning of the solution, and the proposed solution might not be the best in his mind.

Therefore, the application at hand rendering proactivity in dialogue systems is a cooperative planning system, which involves the user in the decision-making during the planning process (see Figure 2). Here, the interactive planning process is manifested in a fitness scenario. The user is guided through the process of selecting appropriate fitness exercises, to arrange an effective but also individual training plan. The automated planning will vary between four different variants: a fully-autonomous process, adding notifications to the user about the system's decisions, asking the user to confirm decisions, or leaving the decision completely to the user. For the latter, the users may decide about several options at times when the process is interrupted because of internal (e.g., planning heuristics) or external (e.g., affective user state) reasons. Hence, proactive behaviour is both the system-initiated integration of the user into the planning process due to planning heuristics and the proactive system reaction to implicit information like user behaviour observed by sensors. Therefore, this includes also proactive behaviour, which is triggered by the user's reaction to previous proactive behaviour. For example, proactive behaviour induced during planning may surprise the user and therefore lead again to proactive system behaviour.

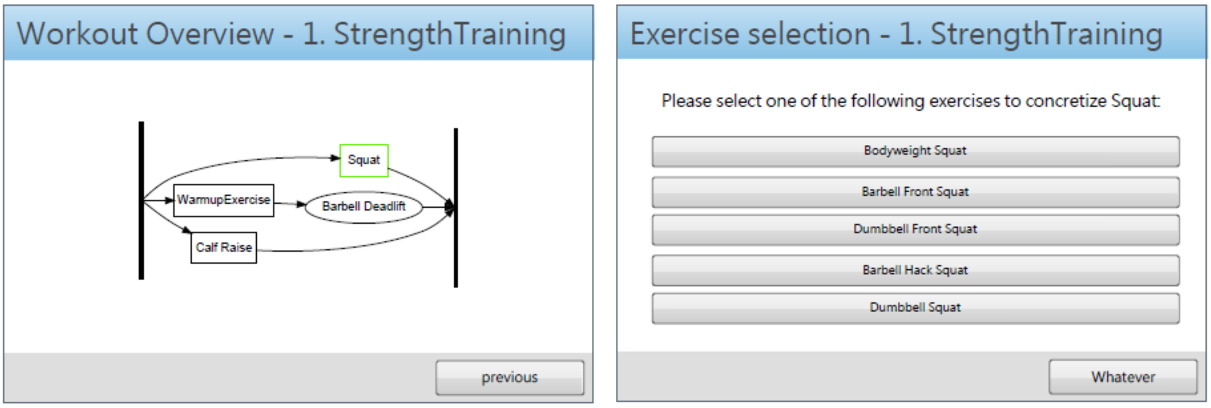

Please select one of the following exercises to concretize Squat:

Figure 2: A screenshot of the interface of the prototypical mixed-initiative planning system, whose scenario is the interactive generation of an individual fitness training. The left image shows an overview of the current plan. The plan step *Squat* is still to be decomposed. In this case, the user can make the decision by selecting an appropriate refinement, the selection of a fitness exercise, shown on the right.

Taking a closer look at the user interaction in such a use-case, we encounter several questions regarding proactivity and turn-taking: *If* proactive behaviour is useful or necessary, *when* proactive behaviour should be integrated in the ongoing interaction (which is the one most related to classic turn-taking), and *how* this proactive behaviour should look like. In the next section, these questions will be discussed more thoroughly with regard to our application scenario.

4 Challenges for Proactive Systems

The three key questions for a proactive system during HCI are if, how, and when a proactive system behaviour is needed. Although those key questions will be discussed individually, we will see that they are nevertheless related to each other.

If

Proactive behaviour is per definition anticipatory, with the idea to react to a future situation. Hence, proactive behaviour involves system actions apart from the expected task-oriented dialogue between human and computer. In our scenario this is either the user-integration into the planning process or anticipatory system-actions dealing with an affective user state. Whether proactive behaviour is needed depends on several factors:

- How important is the proactive behaviour for the successful continuation of the dialogue, i.e., is it critical and required for short-term goals, but risks the cooperativity for interaction in the long run, or only beneficial in a longer perspective, to induce proactivity?
- Does the current user situation allow for additional system behaviour, e.g., additional system prompts?
- What is the classification probability for the cause of the proactive behaviour?

These main dimensions of whether proactive behaviour is adequate span the decision space depicted in Figure 3. If all three dimensions show significant values, proactive behaviour should be induced. It is situation-adequate and triggered with a high probability based on a proactivity area within the decision space. Notwithstanding having the proactivity area usually originating with all axes at maximum value, its size and shape is highly dependent on the task at hand: while for non-critical tasks, the area may be quite big, critical tasks may have higher requirements to trigger proactive behaviour. Hence, a careful balance of all three dimensions is necessary.

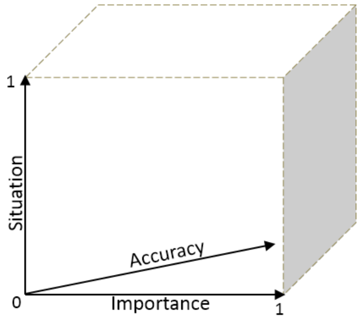


Figure 3: Decision Space: The *Situation* axis depicts if the external (e.g., environment) and internal user situations (i.e., user model) are adequate for proactive behaviour, the *Importance* axis whether the proactive behaviour addresses major or minor flaws in the interaction, and the *Accuracy* axis the recognition hypothesis classification results.

Proactive behaviour itself, however, may have its own pitfalls. Apart from the usual reactive system behaviour where the reaction is anticipated by the user, autonomous decision-making by a proactive system involves the risk of creating incomprehensible and unexpected situations for the user. In the decision space, this maps to the dimension of *Situation*. Those situations usually occur due to incongruent models of the system: during interaction, the user builds a mental model of the system and its underlying processes determining system actions and output. If this perceived *mental model* and the actual system model do not match, the situation will be perceived as inconsistent - the user will not understand it.

In the present application scenario autonomous behaviour by the planner may lead to such situations. For example, the system's automatic preselection between a set of available options may cause user's confusion. The proactive system behaviour—adapting to the user's history of interaction—was not expected by the user, and might therefore be incomprehensible. These unexpected or incomprehensible situations have shown to reduce the user's trust in the system (Muir, 1992) and may ultimately result in reduced frequency or complexity of use (Nothdurft and Minker, 2014). The recognition of improper mental models appears not to be an easy task and requires the recognition of the "symptoms rather than the disease". This means that affective user states like *confusion* have to be recognized and compared to, e.g., the dialogue history to infer whether the mental models were incongruent.

The recognition of emotions and affective user states is one of the much studied research questions at the moment. Apart from basic emotion recognition, the most used affective user states to be recognized via vision-based, audio-based, and audio-visual recognition are *interest*, *frustration*, *boredom*, and *confusion* (Zeng et al., 2009). In a meta-analysis on unimodal and multimodal affect detection, D'Mello and Kory (2012) stated that multimodal recognition accuracies yield performance improvements compared to unimodal affect recognition accuracies. However, in a naturalistic or semi natural (induced) context the improvements are minimal compared to classifiers trained on acted data. They found that contemporary affect detection mostly concentrates on bimodal or trimodal approaches. The most commonly used modalities are acoustic-prosodic cues and facial expressions (77% of all classifiers), followed by gestures, body movement and postures (30% of all classifiers). In general, recognition results based on non-acted data lead to accuracies ranging from 55% to 89%, with an average of 66%. Although there is promising work on this topic, spontaneous affective behaviour analysis in real settings, also commonly called "in the wild", still got a long way to go.

How

The next step in proactive system behaviour is the decision on how the system behaviour should be rendered, i.e., what kind of intervention is the most adequate. If we take a look at the prototypical application at hand, the interactive planner, several open questions arise. For example, even if the planning system decides that the user should be integrated in the next decision, still the question remains at which level the integration should be done (e.g., implicit vs. explicit, pruned vs. original). On the one hand an implicit confirmation of a system-preselected option may be possible where the user is only notified about the decision. On the other hand an explicit selection from a list of choices where the user's choice is unconstrained. For the former, the user also has the option to discard the system choice. Though these issues are related to proactive behaviour in our application, it is part of previous research. The most prominent work dealing with pruning (i.e., removing options) when presenting alternatives as lists was conducted by Sears and Shneiderman (Sears and Shneiderman, 1994). They stated that lists pruned to frequent selection options were faster and subjectively preferred to alphabetic lists.

In our work, the focus lies on how the systems' behaviour should be shaped when recognizing incomprehensible situations. As mentioned before, this may occur due to non-matching models. The user's mental model is a perceived representation of the reality, in this case of the system and it's underlying processes. However, the mental and the actual system model do not necessarily align, which may cause incomprehensibility. In (Nothdurft et al., 2014), we showed that incomprehensible proactive behaviour indeed will significantly reduce the user's perceived understandability and reliability of the system. This was done by training the user on a specific system and then confronting the participants with proactive, not yet experienced system behaviour, where the system did change the user's decision. In order to find out, *how* those situations should be handled by a technical system we took a closer look at human-human interaction. Here, misunderstandings or incomprehension are taken care of by providing explanations. In general, explanations are given to clarify, to change , or to impart knowledge. In these situations, the implicit idea consists of aligning the mental models and to establish a common ground between the participating parties.

Following that, we conclude that a technical system should attempt to clarify its actual model to the user in incomprehensible HCI situations. This means, that explanations should be given, to align the perceived mental model to the actual system model. However, there exists a variety of explanations which pursue different goals (Sørmo and Cassens, 2004):

Conceptualisation usually has the goal to address the user's declarative knowledge (e.g. describing things).

Learning addresses procedural knowledge in the sense, that for example tutorials are provided in order to learn how to do new things.

Justifications are the most obvious goal an explanation can pursue. The main idea of this goal is to provide support for and increase confidence in given system advices or actions.

Transparency increases the user's understanding in how the system works and reasons. This can help the user to change his perception of the system from a black-box to a system the user can comprehend. Thereby, the user may build a better mental model of the system and its underlying reasoning processes.

Relevance explains why the task at hand is relevant to the user. In contrast to the previous two goals that focus on the solution, relevance tries to justify the system-pursued strategy.

Though all of these goals are important, justification and transparency explanations are the most promising ones for incomprehensible situations in HCI. Therefore, we conducted a study testing whether those two explanation goals differ in their effects between each other and to providing no explanations as well. Our hypothesis was that though both explanation goals will help remedy negative effects, transparency explanations will be more helpful. Indeed, we could show that when providing transparency explanations in incomprehensible situations the *perceived understandability*, which measures the ability to build a correct system model using questionnaires, diminished on average only by 0.4 when providing *transparency* explanations (no explanation vs. transparency t(34)=-3.557 p<0.001), and on average by 0.5 with *justifications* (no explanation vs. justifications t(36)=-2.023 p<0.045), compared to 1.2 on a Likert scale with a range from 1 to 5 when providing no explanation at all (see (Nothdurft et al., 2014) for more details).

This showed that providing explanations can help to build a better model, or at least to maintain a model by reducing the impairment, and by that reducing the negative risks of incomprehensible situations. The first part of our hypothesis could be confirmed, whereas the second part is still unclear. Currently we are not yet perfectly sure whether the not-significant difference between transparency and justification explanations was due to improper explanation design or whether those two indeed do not differ in their effects. However, in our opinion the former is more likely, because the complexity of transparency explanations was reduced in our experiment. This means, that in other systems consisting of more complex system processes, the difference between justification and transparency explanations will increase in terms of understanding and building a coherent mental model.

Regarding to our application scenario at hand, this means that incomprehensible situations have to be addressed by providing explanations about the system processes leading to the current system behaviour. For example, the automatic preselection of an action by the system could be motivated using the dialogue history. For example, by providing the explanation that the proactive system behaviour (i.e., the preselection the options) results from recognized user preferences using previous episodes of interaction.

These experimental results show that it seems to be worthwhile to use explanations to cope with incomprehensible situations in HCI. For the decision on *how* proactive behaviour should be shaped, we can state that explaining system processes or providing justifications help to deal with incomprehensible situations. Even if we can decide whether and how the proactive intervention should be shaped, we still need to determine an adequate point of time in the ongoing HCI to provide the proactive behaviour.

When

The problem of *when* to initiate proactive behaviour for HCI means that appropriate turn-taking points in the ongoing interaction need to be found. Those must guarantee sound and effective proactive behaviour. This issue is mostly related to the classic turn-taking problem, which deals with organizing and structuring the conversation by deciding on the system side whether or not to take the floor. Classic ideas in Spoken Dialogue Systems include using pause durations, discourse structure, semantics, or prosodic information and timing features to detect appropriate turn-taking points (Raux and Eskenazi, 2012). While recognizing turn-taking cues in human-human interaction via multimodal signals has been covered in recent research (see (Mondada, 2007) for an overview), the use of multiple modalities to control turn-taking in HCI is only recently emerging as a hot topic. For multimodal systems, this includes analysing the user's verbal and non-verbal signals (e.g., gaze, gestures, body movement) to generate and display well-timed and natural multimodal system behaviour, including feedback and turn-taking signals. While turn-taking itself is already a difficult problem, proactive behaviour includes even more challenges regarding appropriate turn-taking points. For instance, behaving proactively in a given situation might even be so important that it has to be initiated despite inappropriate discourse structure or semantics. Therefore, this issue can be related to the *Importance* axis of the *Proactivity Space* shown in Figure 3. When proactive behaviour is of utmost interest, inappropriate turn-taking has to be tolerated.

5 Conclusion

Future dialogue systems will have to solve increasingly complex tasks cooperatively with the user. As the task complexity as well as the capabilities of such systems increase, it seems natural that these systems will take over some of the responsibilities and help the user achieve the task by proactive system behaviour. Though this might relieve the user by reducing work and cognitive load, it nevertheless involves the risk of incomprehensible HCI situations. In this paper, we elucidated the challenges of turn-taking in proactive system behaviour and how multimodal approaches can help with this issue in the three different decision making layers *if*, *how*, and *when*. The described Decision Space is constructed by the dimensions $Importance$, $Accuracy$ and $Situation$, which are the most important ones to decide *if* proactive behaviour is necessary. In terms of *how* to intervene, providing explanations to foster the building of correct mental models was described in detail. The most promising explanations to foster coherent mental and actual system models seem to be transparency explanations. $When$ to initiate proactive behaviour is mostly related to the classic turn-taking problem. Here recent statistical approaches did lead to a more human-like turn-taking and increased user-experience. However, conclusively we can state that finding appropriate turn-taking strategies for proactive dialogue systems is still an open quest, involving many challenging as well as interesting research questions.

Acknowledgements

This work was supported by the Transregional Collaborative Research Centre SFB/TRR 62 “Companion-Technology for Cognitive Technical Systems” which is funded by the German Research Foundation (DFG).

References

Susanne Biundo, Pascal Bercher, Thomas Geier, Felix Mller, and Bernd Schattenberg. 2011. Advanced user assistance based on ai planning. *Cognitive Systems Research*, 12(34):219 – 236. ¡ce:title¿Special Issue on Complex Cognition¡/ce:title¿.

Sidney D’Mello and Jacqueline Kory. 2012. Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 31–38. ACM.

Starkey Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23:283–292.

Agustín Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Comput. Speech Lang.*, 25(3):601–634, July.

Frank Honold, Pascal Bercher, Felix Richter, Florian Nothdurft, Thomas Geier, Roland Barth, Thilo Hoernle, Felix Schüssel, Stephan Reuter, Matthias Rau, Gregor Bertrand, Bastian Seegebarth, Peter Kurzok, Bernd Schattenberg, Wolfgang Minker, Michael Weber, and Susanne Biundo. 2014. Companion-technology: Towards user- and situation-adaptive functionality of technical systems. In *10th International Conference on Intelligent Environments (IE 2014)*, pages 378–381. IEEE. SFB-TRR-62,Planning,KnowledgeModeling.

Lorenza Mondada. 2007. Multimodal resources for turn-taking pointing and the emergence of possible next speakers. *Discourse Studies*, 9(2):194–225.

B M Muir. 1992. Trust in automation: Part i. theoretical issues in the study of trust and human intervention in automated systems. In *Ergonomics*, pages 1905–1922.

Florian Nothdurft and Wolfgang Minker. 2014. Justification and transparency explanations in dialogue systems to maintain human-computer trust. In *Proceedings of the 4th International Workshop On Spoken Dialogue Systems (IWSDS)*. Springer, January.

Florian Nothdurft, Felix Richter, and Wolfgang Minker. 2014. Probabilistic human-computer trust handling. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 51–59, Philadelphia, PA, U.S.A., June. Association for Computational Linguistics.

Antoine Raux and Maxine Eskenazi. 2012. Optimizing the turn-taking behavior of task-oriented spoken dialog systems. *ACM Trans. Speech Lang. Process.*, 9(1):1:1–1:23, May.

Antoine Raux, Dan Bohus, Brian Langner, Alan W Black, and Maxine Eskenazi. 2006. Doing research on a deployed spoken dialogue system: One year of lets go! experience. In *in Proc. INTERSPEECH, 2006*, pages 65–68.

H. Sacks, E.A. Schegloff, and G. Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4, Part 1):696–735, December.

Andrew Sears and Ben Shneiderman. 1994. Split menus: effectively using selection frequency to organize menus. *ACM Transaction Computer-Human Interaction*, 1:27–51.

F. Sørmo and J. Cassens. 2004. Explanation goals in case-based reasoning. In *Proceedings of the 7th European Conference on Case-Based Reasoning*, pages 165–174.

Nigel G. Ward, Anais G. Rivera, Karen Ward, and David G. Novick. 2005. Root causes of lost time and user stress in a simple dialog system. In *INTERSPEECH*, pages 1565–1568. ISCA.

Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58.

Experiments With Hand-tracking Algorithms in Video Conversations

Pihel Saatmann
Institute of Computer Science
University of Tartu
Estonia
pihels@gmail.com

Kristiina Jokinen
Institute of Computer Science
University of Tartu
Estonia
kristiina.jokinen@ut.ee

Abstract

This paper describes a simple colour-based object tracking plugin for the video annotation tool ANVIL. The tracker can be used to automatically annotate hand gestures or the movements of any object that is distinguishable from its background. The tracker records velocity duration and total travel distance of hand gestures and can be configured to display gesture direction. Results of the tracker are compared to manually created annotations for hand gestures. Data recorded by the tracker is not accurate enough to provide a complete alternative to manual annotation, but could rather be used as a basis for determining where hand gestures can be detected. Thus using the tracker in combination with a human annotator could significantly speed up the annotation process.

1 Introduction

Annotation of video data is an important prerequisite for human communication studies, but doing this manually is time and resource consuming. Analysis requires annotation by trained annotators, and although annotation tools such as Anvil (Kipp 2001) are available, the annotation process is slow and prone to inconsistencies and misconceptions. For instance, considering multimodal video annotation, one problematic issue is the segmentation of gestures and their temporal length: hand gestures can be considered to be the movement of hand(s) only, or they can also include the period after movement when the position of hands can be identified as static. Such differences can be eliminated by clear and unambiguous instructions to the annotators in advance, but also by automatic hand tracking algorithms where the detected hand movements provide a common set of elements to be analysed further by the annotators. Automatic gesture recognition also contributes to objective viewing of gesture elements and to more systematic treatment of data in general. It is thus useful to implement object tracking and gesture recognition algorithms in order to automatically annotate hand gestures, and given the large amounts of video data being collected in various projects, there is an urgent need for such advanced tools. The tracking tool described in this paper is intended to speed up the process of manual annotation, and help in the analysis by performing automatic annotations.

In this paper we describe a hand tracker plugin for Anvil. The goal is to create a technical solution that visually identifies hand movement on video files and tags this with descriptive and quantitative information. The plugin works in a similar manner as the plugin for face tracking (Jongejan 2012), but the hand tracker interface has controls for minimum saturation threshold and how many frames to skip on each iteration. Moreover, hand tracking is a more difficult task than face tracking, since the shape of the hand is not constant or as easily recognizable as that of the face, and hand movements are also rapid and irregular.

The paper is structured as follows. Section 2 gives an overview of the object detection and tracking algorithm CAMShift which is used in the paper. Section 3 describes the plugin for the Anvil while Section 4 gives an overview of the evaluation of the implementation. Section 5 provides discussion on the results, and Section 6 outlines possible solutions and future improvements, as well as draws conclusions and describes future work.

K. Jokinen and M. Vels. 2015. Proceedings of The 2nd European and the 5th Nordic Symposium on Multimodal Communication.

2 CAMShift Algorithm

Hand gesture recognition is a complex task which consists of several subtasks. As in object tracking in general, hand tracking algorithm must first detect the hand, then track the hand in each consecutive frame of a video, and finally estimate the trajectory and identify the complex movements as a gesture. Yilmaz et al. (2006) provide an overview of object tracking algorithms, and point out that an efficient algorithm also has to run in real-time and overcome problems such as image noise, poor or changing lighting, complex object shapes, irregular object motion, occlusion, and distractors present in the video. The objects of interest are represented by their shapes and appearances, and appropriate features are selected to distinguish them in the feature space. The four common features are colour, edges, optical flow and texture, and additional information about the object's orientation and shape can also be provided. Tracking algorithms often use a combination of various features (Han et al. 2009).

In this work we use the CAMShift (Continuously Adaptive Mean-Shift) algorithm (Bradski, 1998). CAMShift is a simple colour-based object tracking algorithm and it uses the HSV (Hue Saturation Value) colour system, which separates the colour (hue) of a pixel from the concentration of the colour (saturation) and brightness (value). For tracking an object only the hue data is used. Brightness and saturation values can be used to filter out noise caused by grey (low saturation) or white (high brightness) pixels which hue values cannot be accurately extracted.

CAMShift is based on the mean-shift algorithm, which works by finding the local maximum of a probability density function and iteratively moving a predefined search window until its centre is located over the maximum. In object tracking this means that if given an initial search window the algorithm will find the point in an image where the pixels matching the tracked object's hue value have the highest density and move the search window so it is centered over the point of maximum density. The process is repeated for each consecutive frame in a video recording and can be used to find the location of the tracked object in each frame. CAMShift was chosen for the implementation because it is fast, computationally efficient, and does not require prior training or a feature database.

3 Hand-tracking Annotation Plugin

The CAMShift algorithm is implemented as a plugin for the ANVIL annotation software (Kipp, 2001). ANVIL is a freeware video annotation tool that allows the user to create multi-layered colour-coded annotations (http://www.anvil-software.org/). The hand-tracker plugin produces a new track, camshift, in the Anvil specification file, where the detected gesture elements are marked.

The plugin is able to track hands and other coloured objects in a video and detect movements. The initial detection of the hand is left to the user by having them select a rectangular area that includes at least a part of the tracked object. The colour data for the selected area is used as a template for finding the tracked object in each consecutive video frame.

The tracker is able to track a single target at a time. If there is more than one object that matches the selected colour template in the video, it is possible that the tracker may switch targets at some point when another object of similar hue happens to be near the tracked object in the video. It is currently not possible to change the colour template for the tracked object after the tracker is started. In order to track a different coloured object the plugin must be restarted first so that a new area for tracking can be selected.

3.1 User Interface

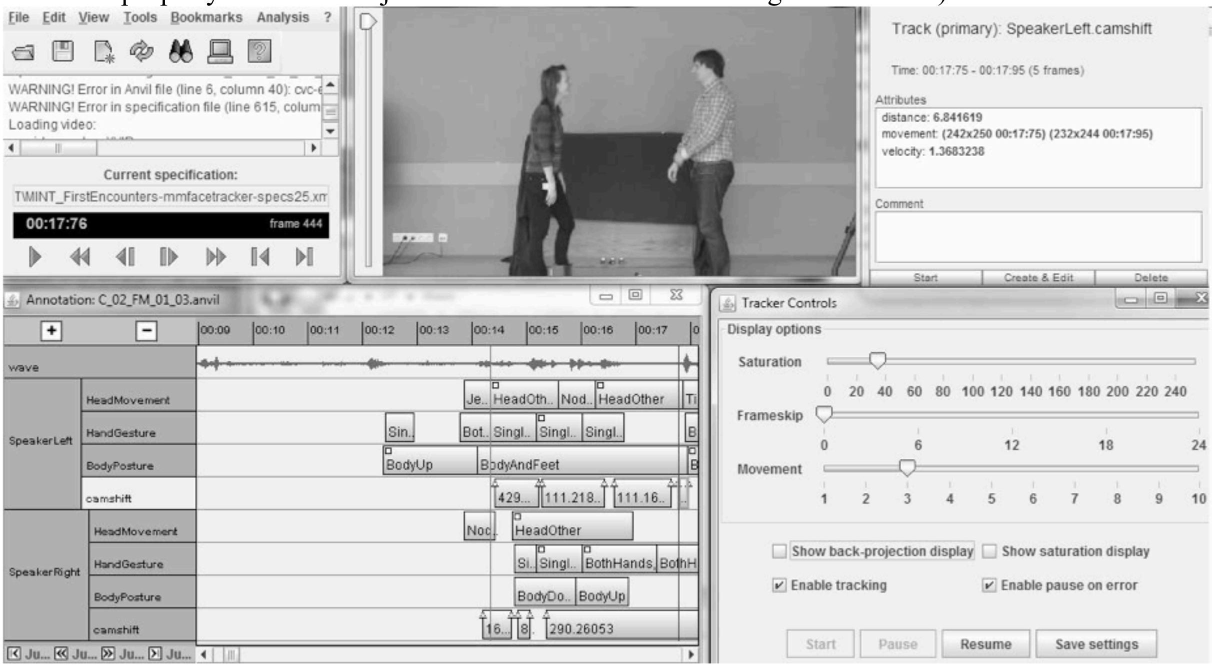
The user interface to the Anvil and the tracker is given in Figure 1 (next page). Besides the Anvil-related windows of the annotation board, the video file, and the control and element information windows, the interface contains a specific control board for the plugin (Figure 1 bottom right). This includes controls for settings that can improve tracking accuracy and efficiency:

- **Saturation –** threshold for ignoring pixels with low saturation values.
- **Frameskip** – number of frames skipped after each processed frame.
- **Movement** – threshold for detecting movement in pixels.

Saturation can be used to filter out video noise. The possible range of saturation values is 0-255, usually filtering out pixels below the default value of 35 should be enough to improve tracking precision. With *Frameskip* the user can control the trade-off between computational cost and accuracy: higher values will lower computational costs, but may also lower tracking accuracy. *Movement* is used to control the length of the gesture. In each frame movement is measured by comparing the current location of the object to its location in the previous frame. If the difference between the two points is above the threshold in three consecutive frames, then a new annotation element is created and the movement is considered to be continued. If the difference is below the threshold in three consecutive frames, then no new annotation element is created, the previous annotation element will be completed and the movement is considered to have ended. This means that with higher values small or very slow movements will be ignored.

The best values for saturation and movement thresholds depend on the video quality and content. The user can change the values at any time during or before tracking. In order to achieve better accuracy different these values can be tried out depending on the user's needs. For example if the priority is to detect all movements in the video then the movement threshold should be lower, however if the user wishes to ignore small movements then the value can be set higher.

The user can also enable pausing the tracker when errors occur (for example when the tracked object becomes occluded and the search window is lost). The user can re-select the tracked object in the main video screen at any time, including when the tracker is paused. It is also possible to select a new object for tracking, however the colour data from the initial template will still be used (meaning the tracker will work properly if the new object is of similar colour to the original selection).



WARNING! Error in Anvil file (line 6, column 40): cvc-
WARNING! Error in specification file (line 615, colum
Loading video:

Current specification:
T\MINT_FirstEncounters-mmfacetracker-specs25.xm

00:17:76 frame 444

Track (primary): SpeakerLeft.camshift

Time: 00:17:75 - 00:17:95 (5 frames)

Attributes
distance: 6.841619
movement: (242x250 00:17:75) (232x244 00:17:95)
velocity: 1.3683238

Comment

Annotation: C_02_FM_01_03.anvil

Tracker Controls

Display options
Saturation
Frameskip
Movement

□ Show back-projection display □ Show saturation display
☑ Enable tracking ☑ Enable pause on error

Figure 1. ANVIL and the tracker's user interface.

3.2 Anvil annotation elements

The tracked object's movements are automatically annotated by writing a new element containing data about the movement to a specified annotation track camshift in ANVIL (Figure 1 bottom left). In our experiments, the specification file contains multimodal behaviour tracks as specified in the NOMCO annotation scheme (Navarretta et al. 2012). In particular, it contains manually annotated gestures with which the detected gesture elements can be compared.

The recorded detailed information for each element can be viewed by clicking on a specific element on the annotation track. The details window (Figure 1 top right) shows information about the start and end point of the movement in the video, total movement distance and average velocity. The start and end point of the movement are also used to display a vector in the main video window between the two points.

3.3 Initial element selection

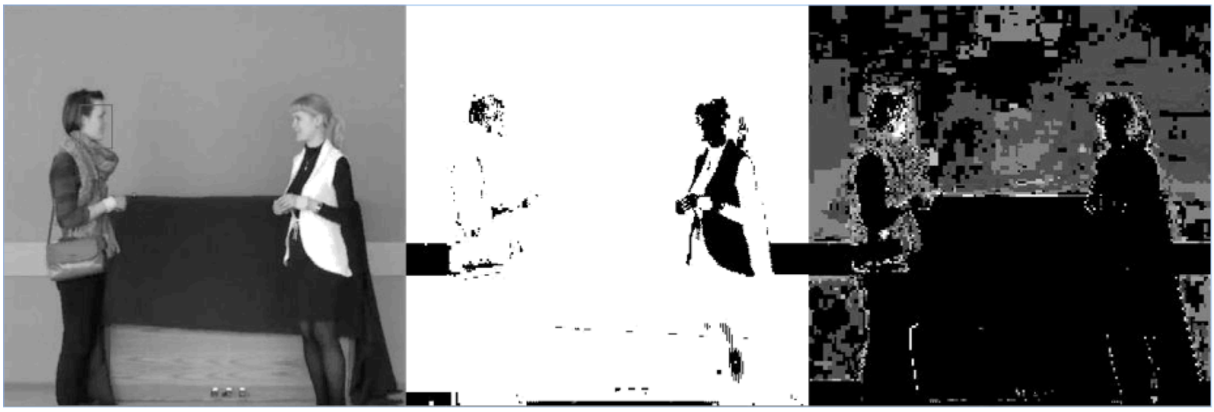


Figure 2. Normal video frame, saturation mask and probability map.

In order to select the most suitable area of an object for tracking, the user can enable the back projection display which shows the colour probability map of the current video frame. The colour probability map shows the most probable locations for the tracked object in the video frame, higher probability is represented by lighter pixel values. It is especially important to initially select as large an area of the object for tracking as possible but not include any parts of the background in the selection.

The saturation display shows pixels in the current video frame that have a value below the saturation threshold and are thus ignored when calculating the probability map. The saturation and probability representation of a video frame can be seen in Figure 2.

4 Evaluation

The plugin has been tested on real dialogue data recorded as part of the ETF (Estonian Science Foundation) project MINT (Multimodal INTeraction, Jokinen and Tenjes, 2012). The dialogue recordings feature two participants who are facing each other and are filmed from a side view as can be seen in Figure 2. The plugin was tested by tracking both bare hands and the yellow bracelets worn by some of the participants. The tracker has been evaluated by comparing annotations created by the tracker and manual annotations (see Section 4.2).

4.1 Tracking Bare Hands and Coloured Objects

The simple colour-based object tracking plugin proved to be problematic when tracking bare hands. The tracker would jump to other hands or faces in the frame if the regions overlapped (for example if the participants shook hands or crossed their arms) or get lost if the hand was completely occluded (for example if the person put their hands in their pockets). Movement was not accurately detected and very often small movements were detected when there actually was none. Detection accuracy was also affected when the participants had bare forearms as the tracker would then try to track the entire length of the bare arm instead of just the hand area.

However, when tracking the yellow bracelets, the tracker was working as expected. The tracking window didn't move to the wrong object and was lost only if the bracelet was completely occluded. Movement was detected relatively accurately. Tracking the bracelets was more accurate because the bracelets have a hue that is easily distinguishable from the background. In most cases it was the only object in the video of that particular colour. Even if more than one bracelet was present in the video it did not come into direct contact with other bracelets and thus the tracker would always stay on the right target.

4.2 Comparison to Manual Annotations

The tracker was applied to four randomly seclected video files that had been manually annotated for hand gestures following the NOMCO annotation scheme. The tracker was set to track the participants' yellow bracelets. In cases where one manually annotated movement was detected by the tracker as multiple movements, it was counted as one true positive detection. All single false positive annotation elements created by the tracker were counted as separate detections even if they were consecutive. In a more detailed comparison, the separated elements which have a distance below a threshold could be regarded as belonging to the same gesture, thus mimicking separation of the detected true positives. When the tracked object was lost then the tracker was manually reset to the correct position.

The average precision, recall and F-measure values for the tracker was calculated using the formulas below. True positive values are marked as TP, false positive as FP and false negative as FN. Table 1 shows a comparison of the tracker's movement detection ability to manual annotations.

Precision or positive predictive value represents the fraction of correct detections over all movements that were detected:

$P = TP \div (TP + FP) \approx 0{,}306$

Recall represents the fraction of correct detections over the number of movements that should have been detected:

$R = TP \div (TP + FN) \approx 0{,}982$

F-score is a weighted average of precision and recall and calculated as their harmonic mean:

$F1 = 2 * P * R \div (P + R) \approx 0{,}455$

Speaker	TP	FP	FN	Precision	Recall	F-score
1	22	68	0	0,244	1	0,393
2	9	43	0	0,173	1	0,295
3	17	24	0	0,415	1	0,586
4	60	84	4	0,417	0,938	0,577
5	63	96	3	0,396	0,955	0,560
6	25	107	0	0,189	1	0,318
Average				**0,306**	**0,982**	**0,455**

Table 1. Summary of the tracker's performance.

5 Discussion

The results show that the tracker is able to detect hand movements in video files, however the accuracy of gesture detection is low due to a large amount of false positive detections. As can be seen from the evaluation data the tracker has high recall values, but low precision due to a large number of false positive detections. The tracker also often detects multiple smaller movements where there is actually one long movement as seen in Figure 3. This is due to the tracked object's velocity dropping to very low values during some parts of the movement.

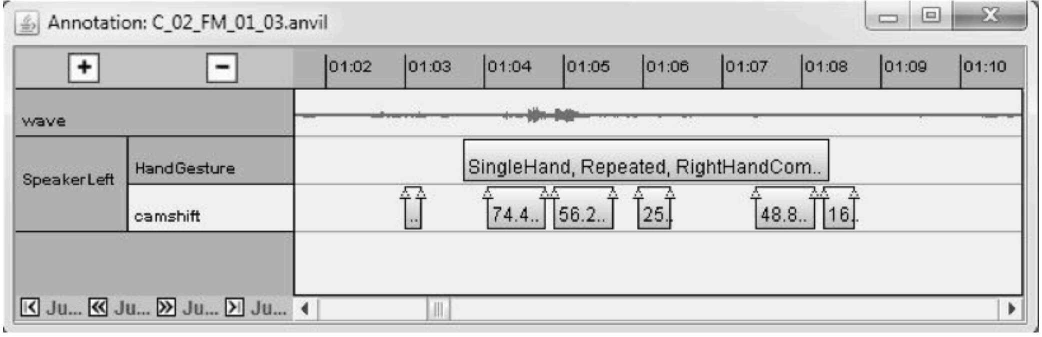


Figure 3. Comparision of manual (top track) and tracker (bottom track) annotation.

One of the main problems encountered during testing was that the tracker detects movement when there is none. Increasing the movement detection threshold would solve this problem, but then actual movements that are below the threshold (for example very slow movements) would also not be detected. However, the problem is alleviated by the fact that the manual annotations contained only data about hand movements that were deemed communicatively significant. Very small movements and those caused by the whole body moving were not annotated. Therefore in reality the number of false positives can be lower than represented by the evaluation results. In fact, this can be interpreted as a novel means to support more objective manual annotation as the tool detects gestures which the human annotators may not have even noticed. The annotators are thus provided with the same set of objectively recognised movements which they have to annotate and agree on their interpretation.

6 Conclusion

This paper discusses automatic hand tracking in video files and compares the automatic recognition results with human annotations. The tracker is good at detecting movements, but sometimes finds movement when there was none. Changing the detection threshold can improve the results, but is a trade-off since it would also prevent small movements being detected. In order to minimise false positive detections, the movement detection algorithm would have to be improved so that it would ignore very small movements yet be able to recognise long and slow movements as a single gesture. On the other hand, this can also be used as a basis for objective gesture segmentation and a starting point for the interpretation of communicative function of the detected hand movements.

The tracking and movement detection precision depend on the quality of the video being used and on the user specified settings. The tracker works better on objects that are easily distinguishable from their background by colour. Problems were identified when the hue of the hands was similar to the background colour. For fully automated annotations, possibilities to ignore distractors such as other hands present in the video and static background objects would have to be researched. Pre-processing the video material in order to improve quality could also be considered.

At current state the plugin cannot be used as a fully independent annotating tool as the number of false positive results is too high, and the tracker requires some human interference when the tracked object is lost. However, although the tool does not currently provide a complete alternative to manual annotation it can be used for creating a rough basis for annotations. Since evaluation showed that the tracker was able to detect all movements, the basis can be used to determine where in a video hand gestures can be detected. The user would be required to manually remove false positive detections.

We will continue to test the algorithm with different video files, and to evaluate the robustness of the algorithm. We will also test the quality of the results when the video quality goes down. Future work will also see detailed comparison of the linguistic-pragmatic annotations with the recognized gesture signals. The algorithm can be applied to human-robot interaction in order to study the robot's understanding of human gestures. Experiments on these lines are described in Han et al. (2012).

References

Gary R. Bradski. 1998. Computer video face tracking for use in a perceptual user interface. Intel Techno-logy Journal. Q2, pp.705-740.

Bing Han, Christopher Paulson, Taoran Lu, Dapeng Wu and Jian Li. 2009. Tracking of Multiple Objects under Partial Occlusion. Automatic target Recognition XIX, 7335. Available at: http://www.wu.ece.ufl.edu/mypapers/trackingSPIE09.pdf

Jing Guang Han, Nick Campbell, Kristiina Jokinen and Graham Wilcock. 2012. Investigating the use of non-verbal cues in human-robot interaction with a Nao robot. Proceedings of 3rd IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2012), Kosice, 679-683.

Bart Jongejan. 2012. Automatic annotation of face velocity and acceleration in Anvil. Proceedings of the Language Resources and Evaluation Conference (LREC-2012). Istanbul, Turkey.

Kristiina Jokinen and Silvi Tenjes. 2012. Investigating Engagement - intercultural and technological aspects of the collection, analysis, and use of the Estonian Multiparty Conversational video data. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), 23-25 May 2012, Istanbul, Turkey

Kristiina Jokinen and Graham Wilcock. 2014. Automatic and manual annotations in first encounter dialogues. Proceedings of the 6th International Conference: Human Language Technologies – The Baltic Perspective. Kaunas, Lithuania.

Costanza Navarretta, Elisabet Ahlsén, Jens Allwood, Kristiina Jokinen and Patrizia Paggio. 2012. Feedback in Nordic First-Encounters: a Comparative Study. Proceedings of the Language Resources and Evaluation Conference (LREC-2012). Istanbul, Turkey.

Michael Kipp. 2001. Anvil - A Generic Annotation Tool for Multimodal Dialogue. Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech), pp. 1367-1370.

Alper Yilmaz, Omar Javed, and Mubarak Shah. 2006. Object tracking: A survey. ACM Computing Surveys. 38 (4), Article 13. Available at: http://crcv.ucf.edu/papers/Object%20Tracking.pdf

Nodding in Estonian First Encounters

Dage Särg
Institute of Estonian and General Linguistics
University of Tartu
Estonia
dage@ut.ee

Kristiina Jokinen
Department of Computer Science
University of Tartu
Estonia
kjokinen@ut.ee

Abstract

This paper discusses nodding as one of the most significant means of feedback signalling in human conversations. It focuses on nodding in Estonian first encounter conversations, and compares nodding with similar feedback behaviour in Finnish, Swedish and Danish. Different types of nods (up-nods and down-nods, one-way, single and repeated nods) are discussed in terms of frequency, durations and variations between individuals and genders.

1 Introduction

Feedback has a central role in human communication, enabling the speakers to understand each other, to build a shared context, and to connect with each other emotionally. In natural conversations, participants have different goals that they aim to achieve more or less consciously (e.g. exchange information, perform a task, learn to know each other, keep communicative channel open) and feedback is a means of monitoring and facilitating the progress of these goals, by acknowledging the partner's contributions and showing interest and willingness to continue the conversation. The conversational model by Clark and Schaefer (1989) introduces feedback as part of the grounding process whereby the participants present information and accept it through a continuing process of presentation-acceptance cycle. In dialogue management, grounding – establishing common ground – has been modelled via specific grounding actions which the agent can plan as part of the dialogue progressing (cf. original idea in Traum 1999), or it has been assumed to take place as a side-effect of the agent planning how to react in the dialogue situation, using conversational principles and rationality considerations (cf. Jokinen 1996).

Feedback can be signalled in several ways, using various verbal and non-verbal means. It can consist of short linguistic morphemes like "ok", "fine", or backchannelling vocalisations like "hmm", "uhm", or some form of multimodal gesturing, facial expressions, head movements, body posture, etc. Often more than one modality is used simultaneously to give feedback, e.g. feedback combines both verbal and non-verbal forms such as head nodding while saying "yeah". Much research exists about the various aspects of feedback, its timing, form and function in the interaction in general (Allwood, Nivre, Ahlsén 1990; Allwood et al. 2008; Toivio and Jokinen 2012). Recently feedback has been studied especially in the context of multimodal communication, based on the holistic view of communication, as well as in different cultural and technological contexts. For instance, Navarretta et al. (2012) compared nodding in the three Nordic countries, while Nunn and Tamura (2003) studied nodding behaviour of Japanese and foreign students in intercultural communication context.

The aim of this paper is to study nodding as a feedback signal in Estonian first encounter dialogues, and also to compare and contrast the Estonian data with the results reported in Navarretta et al. (2012) on Danish, Swedish and Finnish data. Since the activity type is the same in all data (first encounter dialogues), it is easier to compare nodding behaviour in the four languages. Moreover, it is interesting to study nodding in the neighbouring countries which share geographical closeness and are linguistically related such as Finnish and Estonian. We assume that feedback is expressed in a language specific manner and interpreted through learnt cultural contexts, so it is not a straightforward conversational

K. Jokinen and M. Vels. 2015. Proceedings of The 2nd European and the 5th Nordic Symposium on Multimodal Communication.

action, but forms complicated behaviour patterns even in closely related cultures like the Nordic countries. Thus comparison of different feedback strategies in different cultural contexts and among different speakers does not only contribute to the research on feedback functions in different communicative contexts, but also to a deeper understanding of human interaction in general.

The rest of the paper is structured as follows. Section 2 gives a brief review of nodding in general. Section 3 describes the data used for the present study, and Section 4 discusses the annotation of the data. Section 5 presents the results of nodding frequency received on the Estonian data, Section 6 brings out the interpersonal variations in nodding, and Section 7 regards the duration of different types of nods in Estonian. Finally, Section 8 compares and contrasts the Estonian results with the results received by Navarretta et al. (2012).

2 Review of nodding

As already mentioned, feedback is crucial in smooth communication. Nodding can be regarded as one of the conversational signals universally inherent to humans although it should also be emphasised that the interpretation of head nods is culture-specific. Even though in many cultures (including Estonian), nodding serves as a sign of agreement, this is not universal. For instance, Andonova and Taylor (2012) describe that the similar nodding behaviour is interpreted as negative agreement in Bulgaria. However, nodding is one of the main ways to give feedback. For instance, Navarretta et al. (2012) found that nodding is the most common communicative signal among the various head gestures, while Knapp and Hall (1997) regard head nodding as the "primary non-verbal signals" in back-channelling. In automatic agents and social robotics, it is also common to enable the agent to provide feedback to the user by nodding (e.g. Jokinen and Wilcock, 2013).

Nodding is an intuitively clear means of communication but like many other gestures, nodding can serve different purposes. As already discussed, nodding is the main means to provide feedback to the partner concerning if the communication is successful or not. It can also be used to manage turns in conversations, e.g. to indicate the next speaker explicitly, and it can be used in sequencing turns, i.e. as a speech act, or as a part of a dialogue (Allwood et al. 2008; Poggi et al., 2010).

Nodding features moving one's head vertically up and/or down, either once or repeatedly. Therefore, nods can also be divided into different types based on the direction that they start with (up-nods with an upward movement and down-nods with a downward movement) and the number of movements performed (single nods with maximum one movement in both directions, repeated nods with more). In MUMIN coding scheme, four types of nods are distinguished: single up-nod, single down-nod, repeated up-nod and repeated down-nod (Navarretta et al. 2012).

The different nodding types can also function differently in the interaction. For instance, Toivio and Jokinen (2012) observed that different types of nods are used in different ways in communication. Their research on Finnish nodding has confirmed their hypothesis: they found that up-nods usually mark the information that is new to the listener, while down-nods signal that the information is already known. Poggi et al. (2010) have explored this topic more thoroughly, describing in detail what kind of nods are used for which purpose in political TV debates in Swiss French. They noticed that backchanneling is indicated by repeated nods in their data.

Considering interactions between human users and future communicating machines in general, such as social robots and embodied conversational agents, it is important to understand multimodal feedback in order to support natural communication. Cassell (2001) discusses how humans use all the opportunities offered by their bodies in conversations, and accordingly, she proposes that when human-computer conversations are designed, the opportunities that come from human bodies have to be taken into account as well, in order to make the conversation feel like a real conversation. This view accords with the notion of affordance discussed in Jokinen (2009), as a property of natural language interactive systems (the notion is an extension of affordance in product design, introduced by Norman (1999)). If an interface or a communication offered by a computer agent is affordable, this means that the agent suggests to the user the natural intuitive ways to communicate by its own behaviour. The agent can also adjust itself to the user's needs and expectations, and provide useful feedback to the user concerning shared tasks and their completion. The automated agent can also adapt to different user strategies and to different emotional situations (e.g. Cassell 2001; Csabo et al. 2012; Beck et al. 2010).

3 Data

The data used for this paper was collected in the framework of the project MINT (Multimodal Interaction – intercultural and technological aspects of video data collection, analysis, and use), described in Jokinen and (2012). The videos were collected according to the same principles as the NOMCO files (Paggio et al. 2010) and featured encounters with two people who meet for the first time. The purpose of the conversation was to become acquainted with each other and the videos are thus unique situations. The snapshot of one of the videos is in Figure 1.

There are 12 videos that were analysed, and among them there are 4 female-male conversations, 4 female-female conversations and 4 male-male conversations. All the participants are either university students or they already had a university degree.

The lengths of the videos vary from slightly under 5 minutes to almost 7 minutes, and the total duration of the analysed video files is 68 minutes 45 seconds. As for the nods, altogether there are 1342 nods, and the average duration of a nod is 1.29 seconds (i.e. almost 29 minutes of nodding behaviour). The average nodding frequency is 0.16 nods/second (including single and repeated nods).

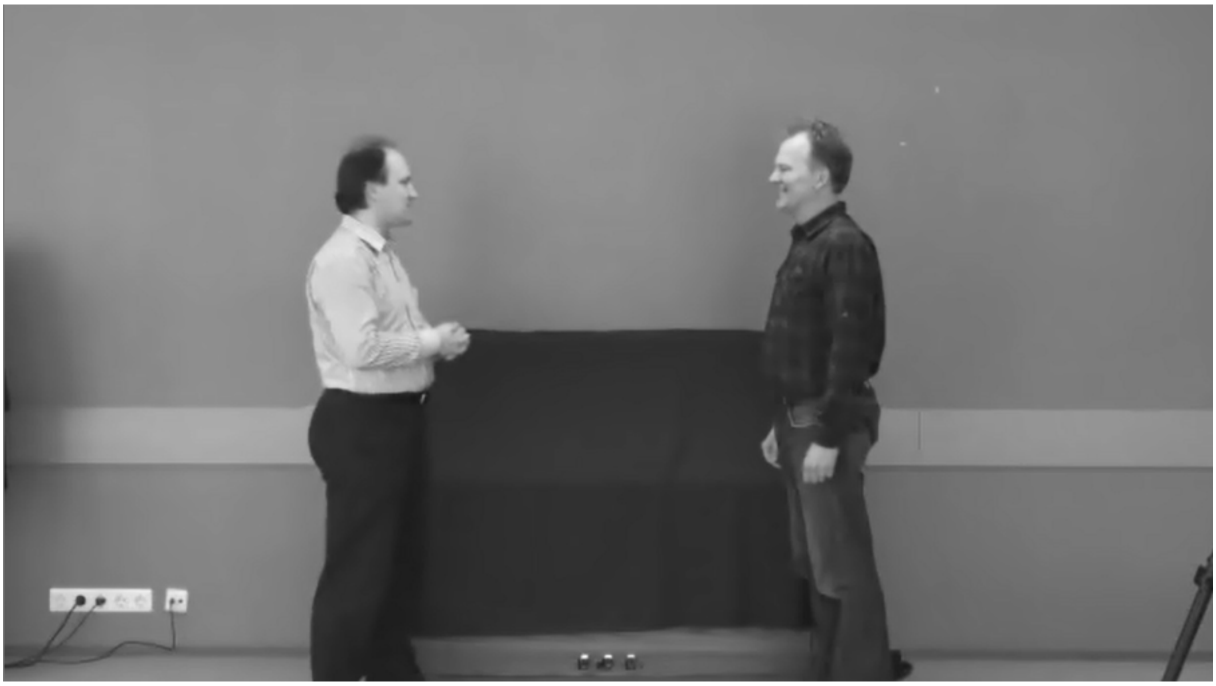


Figure 1. A snapshot of an Estonian first-encounter video.

4 Annotation

The ELAN software (version 4.6.1) was used to annotate the videos, and the MUMIN annotation scheme (Allwood et al. 2008) was used as a guideline for annotations. The annotation thus followed the one in Navarretta et al. (2012), and the same types of basic nods were distinguished as follows:

1. single up-nod – the person moves their head first upwards and then downwards
2. single down-nod – the person moves their head first downwards and then back upwards
3. repeated up-nod – the person moves their head upwards and downwards several times, starting with an upwards movement
4. repeated down-nod – the person moves their head downwards and upwards several times, starting a downwards movement

In addition, two more single nods were distinguished, since they were present in the videos but did not fit in the description of single up-nod or down-nod. These one-way nods were distinguished from the other single nods and annotated differently:

5. one-way up-nod – the person moves their head upwards but not back downwards
6. one-way down-nod – the person moves their head downwards but not back upwards

In cases 5 and 6, the person of course moves their head up/down but this happens some time later and cannot be considered as part of the up-down nodding feedback; rather it is related to some other function in conversation such as focussing attention back to the speaker.

The annotation was done by the first author and checked by the second author. It must be emphasised that in some cases it is hard to distinguish whether the repeated nod actually begins with a downward or an upward movement, even when watching the video in slow motion. In these cases, the first intuition of the movement direction was used.

5 Frequency of Nods and Gender Variation

The total length of the 12 files was 4125 seconds (doubled to 8250 for calculations as there are two speakers in each file) and there were 1342 annotations altogether. Table 1 shows the nod counts, percentage and time-wise frequencies of the different nod types averaged over time.

Considering nod counts, down-nods (almost 71%) are twice as common as up-nods (29 %) and the most numerous nodding type is a single nod: almost half of all the occurrences (49 %, 661 occurrences of all the occurrences) are single nods, although repeated nods are fairly close (41%, 550 occurrences). Within the down-nods, the difference between single down-nods (48 %, occurred 450 times) and repeated down-nods (45 %, 429 times) is small, but within the up-nods, single up-nods (53 %, 211 occurrences) are almost twice as common as repeated up-nods (31 %, 121 times). One-way nods are the least common ones (10%, 131 occurrences of all occurrences), but there is a significant difference if we compare one-way nods relative to down-nods and up-nods: one-way up-nods are about twice as common among up-nods (16% of type) as one-way down-nods among down-nods (7 % of type).

Considering the nod frequencies (number of nods per second), the tendencies are also clear: single nods are more frequent than repeated ones, down-nods are more than twice as frequent as up-nods, and single down-nods are the most frequent nod type. The difference between single and repeated down-nods is not significant, whereas the difference between single and repeated up-nods is; the single up-nods are almost twice as frequent as repeated up-nods. One-way nods are the least frequent, but compared with the count percentages above, that there is no difference between the frequency of one-way up-nods and one-way down-nods: we now compare the frequency of nods relative to the whole length of the dialogue, not relative to number of up-nods and down-nod types in the dialogue.

		PERCENTAGE		FREQUENCY (n/sec)		
TYPE	COUNT	of all	of type	TOTAL	FEMALE	MALE
ALL NODS	**1342**	**100 %**		**0.163**	**0.163**	**0.162**
One-way	131	10 %		0.016	0.016	0.016
Single	661	49 %		0.080	0.079	0.081
Repeated	550	41 %		0.067	0.068	0.065
DOWN-NODS	**947**	**71 %**	**100 %**	**0.115**	**0.107**	**0.123**
One-way	68	5 %	*7 %*	0.008	0.007	0.010
Single	450	34 %	48 %	0.055	*0.049*	*0.060*
Repeated	429	32 %	45 %	0.052	0.051	0.053
UP-NODS	**395**	**29 %**	**100 %**	**0.048**	**0.056**	**0.039**
One-way	63	5 %	*16 %*	0.008	0.009	0.006
Single	211	16 %	53 %	0.026	*0.030*	*0.021*
Repeated	121	9 %	31 %	0.015	*0.017*	*0.012*

Table 1. The frequency of different types of nods in Estonian (nods/sec)

Considering the differences between male and female nodding, it can be seen that, on average, the frequencies of female and male nodding are practically the same: the figures differ only in the third decimal place. Indeed, t-tests (two-tail t-tests with two-sample unequal variance) also support this conclusion: none of the differences between male and female nodding frequencies can be considered statistically significant as the p-value always exceeds 0.1.

However, it can be seen in Figure 2 that although men and women seem to use one-way and repeated nods almost at the same frequency, they differ in their use of single down-nods and single up-nods.

Male participants use more down-nods than women, while female participants use more up-nods than men (single and also repeated). Toivio and Jokinen (2012) noticed that in Finnish, down-nods are used to acknowledge information as part of the shared context, i.e. down-nods signal that the presented information is already known to the listener, while up-nods are used when information is new to the listener in the given context, i.e. up-nods contain an element of surprise. If the same hypothesis is applied to Estonian, it can be concluded that women express more frequently than men, that the information is new for them – it may be further assumed that this is because female participants try to be polite and show interest in their partner by indicating that what the partner says is new and interesting information to them, whereas men show politeness by acknowledging the partner's message and indicating that they share the information with the speaker

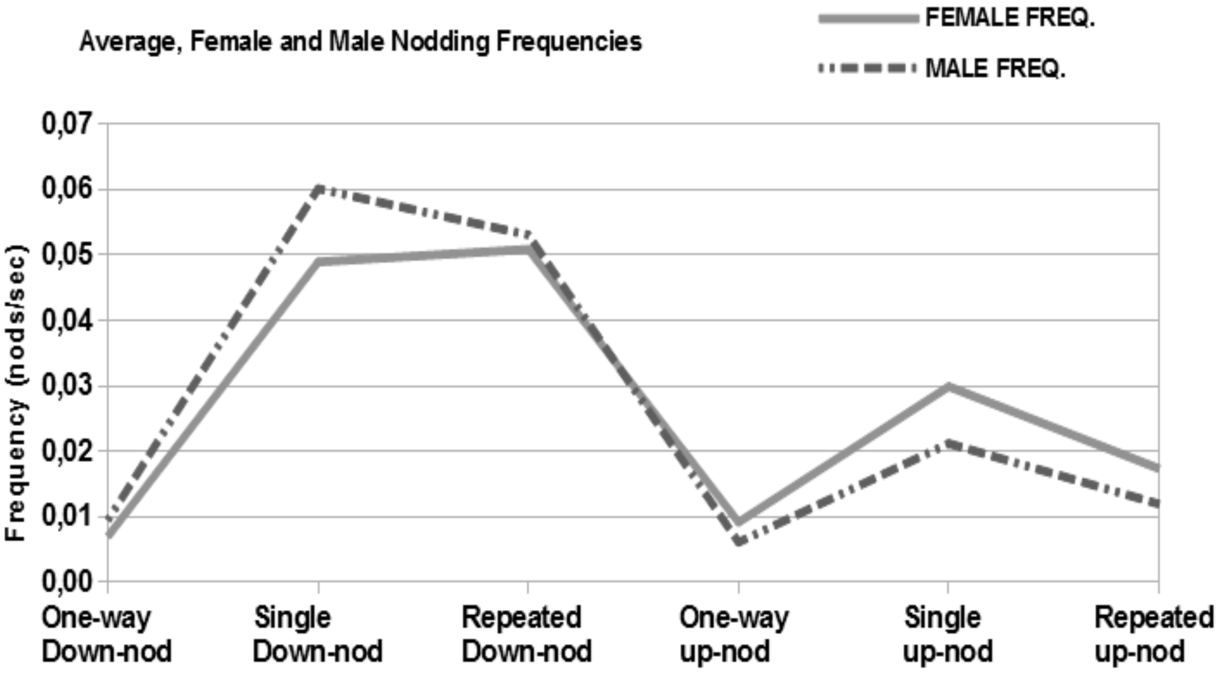


Figure 2: The frequency of different types of nods in Estonian (average, female and male)

6 Interpersonal Variations in Nodding

Navaretta et al. (2012) report on large variation in the nodding behaviour of different persons, and this is true of Estonian speakers, too. Some participants nod in a big and strong manner, so it is easy to recognize their nodding, while others move their head only slightly. Table 2 lists variation in the frequency of nodding by the analysed 18 persons, and large variation can be seen between individual speakers. Some speakers participated in two conversations, and there is also a small difference in their nodding behaviour depending on the partner.

Table 2 shows that the overall variability in nodding frequency is almost 4-fold: the smallest nodding frequency is 0.08 nods/second and the largest is 0.31 nods/second. It is interesting to note that the least frequent nodder (ID4 right) and the most frequent nodder (ID18 left) were both female participants engaged in a female-female conversation. The smallest nodding frequency among male participants is 0.11 nods/second (ID 12 right) and the largest is 0.26 nods/second (ID15 left). The standard deviation of the nodding frequency for male participants is 0.039 nods/second and for female participants 0.069 nods/second, so it can be said that the men's nodding frequency is more constant than that of women's. The nodding frequency in mix gender female-male conversations (IDs 2, 7, 11, 20) varies between 0.13-0.25 nods/second, and both the most and the least frequent nodders in these conversations are female (ID11 left and ID20 left, respectively).

Video ID	PARTICIPANT	GENDER	FREQ	Video ID	PARTICIPANT	GENDER	FREQ
2	left	F	0.23	7	left	M	0.16
	right	M $	0.17		right	F	0.18
9	left	M $	0.19	15	left	M	0.26
	right	M	0.19		right	M &	0.13
4	right	F	0.08	16	left	M &	0.15
	left	F #	0.14		right	M	0.13
5	left	F #	0.14	18	left	F	0.31
	right	F	0.09		right	F %	0.12
11	left	F	0.25	19	right	F %	0.14
	right	M ¤	0.16		left	F £	0.18
12	right	M ¤	0.11	20	left	F £	0.13
	left	M	0.13		right	M	0.17

Table 2. The participants' frequency of nodding (nods/sec) and their gender. A person participating in two conversations is indicated by the matching symbols after their gender ($, #, etc.)

Six people participated in two conversations (3 females and 3 males), and these are marked with a matching symbol in Table 2 (e.g. the participant on the left side in the videos 4 and 5 was the same person). We notice that the nodding frequency of these participants is not precisely the same in both conversations, but if compared with the overall variability in the nodding frequency between persons, the differences are quite small: the largest difference between an individual's nodding frequency is 0.05 nods/second (ID 11 left and ID 12 left, as well as ID19 left and ID20 left), while among different participants, this is 0.23 nods/second (ID 4 right and ID18 left).

Three of these six participants had partners of the same gender in both of their conversations (e.g. a woman in ID4 left and ID5 left participates in two female-female conversations). The differences in their nodding frequencies were negligible: 0.00–0.02 nods/second. The other three had one conversation with a female and one with a male partner, and their nodding frequencies in two conversations differed slightly: in one case the difference was 0.02 nods/seconds, but in two cases it was 0.05 nods/second. Thus we could anticipate that the gender of the partner has an influence on the person's nodding frequency, although the data is too scarce to draw any definite conclusions about which way the influence goes, and it is likely that the influence is more dependent on the personal characteristics of the partners than their gender.

Navarretta et al. (2012) also notices that in the Finnish data, strong nodding is typical while greeting. This seems to be typical in the Estonian data as well, since 12 out of 18 different participants nodded while greeting (3 participants did the greeting before reaching the filming spot, so there is no information whether they nodded or not). Of the persons who participated in two conversations, one participant nodded during both greeting and two nodded in one but not in the other one. All the persons who did not nod during greeting (or about whom there is no information) were male, so it can be said that in Estonian, at least women typically nod while greeting.

7 Durations of Nods

Given that there is a significant difference between single and repeated nods, it is also interesting to study their lengths. The average durations of the different types of nods in the Estonian data is given in Table 3. As was anticipated, one-way nods (just an upward or downward head movement) are the shortest, then come single nods, and repeated nods have the longest duration. The differences in durations are also confirmed by t-tests (two-tail t-tests with two-sample unequal variance) as p-values for both the durations of one-way nods versus single nods and for single nods versus repeated nods were <0.001. There are also slight differences in the average duration of up-nods and down-nods. However, two-tail t-tests with two-sample unequal variance were performed in order to find out whether the differences are statistically significant or not, and the only one that could be considered statistically significant is the difference in the duration of repeated up-nods versus repeated down-nods where up-nods are significantly longer than down-nods (p-value = 0.05; <0.1).

However, comparing the length of single and repeated nods, it can be noticed that single nods are about half of the length of the repeated ones, indicating that individual up-down movements in repeated nods are not necessarily much faster than the single nods. There may be a physiological reason for this due the same motor control mechanism for he head movement up-and-down. However, an interesting observation is that the one-way nods are significantly slower movements than single (or repeated) nods, indicating that they are intentionally different from the other nod types and also have a different function in conversations. It can be hypothesised that the slowness of one-way nods indicate thinking and pondering about the presented information, so that the reaction is not a straightforward acknowledgement or surprise feedback, but includes some hesitation.

	DURATION		
	TOTAL	UP-NODS	DOWN-NODS
ALL NODS	1.29	1.24	1.32
one-way	0.77	0.78	0.77
single	0.98	0.96	0.99
repeated	1.79	1.96	1.74

Table 3. The average duration of nods in the Estonian data (seconds)

8 Cultural Comparison

Table 4 presents frequency results on the Estonian data together with the results presented in Navarretta et al. (2012). As there is no equivalent to the one-way nod in Navarretta et al. (2012), we regarded one-way nods as single nods for the purpose of comparing Estonian results with the ones on Danish, Finnish and Swedish, and added the one-way nod frequencies to our single nods. Even though it is possible that one-way nods have a different interpretation from single nods (as discussed above), they nevertheless provide feedback to the partner and as sometime it may be difficult to distinguish them from single nods, we consider this combination reasonable. The Estonian nodding frequencies calculated in this way can be seen in column Estonian in Table 4, while the original 3-desimal numbers separating one-way nods and single nods are in their own separate column.

	Danish	Finnish	Swedish	Estonian	3-desim+one-way
Nod (Down-nod + Up-nod)	**0.17**	**0.16**	**0.14**	**0.16**	= 0.163
Single	0.08	0.12	0.05	0.10	= 0.080+0.016
Repeat	0.09	0.04	0.09	0.07	= 0.067
Down-nod	**0.14**	**0.11**	**0.07**	**0.12**	= 0.115
Down-nod single	**0.05**	**0.08**	**0.02**	**0.06**	= 0.055+0.008
Down-nod repeated	0.09	0.03	0.05	0.05	= 0.052
Up-nod	**0.03**	**0.05**	**0.07**	**0.05**	= 0.048
Up-nod single	0.03	0.04	0.03	0.03	= 0.026+0.008
Up-nod repeated	0.00	0.01	0.04	0.02	= 0.015

Table 4. Comparison of the frequency of different nod types (nods/sec). The figures of Danish, Finnish, and Swedish are from Navaretta et al. (2012).

The average frequency of nodding in Estonian seems to be the same as in Finnish (0.16 nods/ second). Moreover, the frequency of down-nods (0.12 nods/sec) and up-nods (0.05 nods/sec) are also the same as for Finnish. Yet another similarity with Finnish is the fact that there are more single nods than repeated nods while in Danish and Swedish the opposite is the case. However, the difference between the frequency of single and repeated nods is smaller in Estonian than that in Finnish. The differences are not huge but indicate the differences between neighbouring countries.

9 Conclusions and Future Work

In this paper we studied nodding in Estonian first encounter conversations in terms of frequency and durations of the nods. Interpersonal variations were considered as well. We observed that down-nods are more than twice as frequent as up-nods, and single nods are more frequent than repeated nods. In fact, single up-nods are almost twice as frequent as repeated up-nods, but the difference between single and repeated down-nods is not significant. As for the gender differences in nodding, they were not statistically significant. However, an interesting difference was observed regarding single up-nods and down-nods: female participants tended to use more single up-nods than the male participants, while the male participants used more single down-nods. We hypothesise that this may be related to different politeness strategies.

In this paper, we also distinguished a new type of nodding, that of one-way nod where the head is moved up or down slowly and not returned back to the original position but sometime later. We hypothesised that this kind of feedback may signal hesitation and pondering upon the presented information, and thus differ from the straightforward acknowledgement of proper nodding.

Finally, the frequency of nodding in Estonian was compared with the figures presented in Navarretta et al. (2012), concerning nodding frequencies in the similar Danish, Swedish and Finnish data. It turns out that Estonian nodding is comparable with nodding in the Nordic countries and mostly resembles the Finnish nodding. The Estonian nodding frequency was the same as in Finnish (0.16 nods/sec) and also the distribution of nods into up-nods and down-nods was the same. This is expected considering the close linguistic relationship between the two languages.

In the future, we aim to have more data, and use another annotator, so as to achieve more consolidated results and comparisons. With the larger data, it would be useful to study the one-way feedback signal further and compare its function in other languages, too. It would also be good to study the given-new status of the presented information with respect to different nodding types so as to confirm the hypothesis of the interpretation of down-nods and up-nods, and also shed light on one-way nods. It would also be interesting to study nodding in relation with vocal feedback in Estonian, as has already been done for Finnish (Toivio and Jokinen, 2012). When it comes to individual nodding, we have mainly concentrated on interpersonal variations, but it would be good to study also intrapersonal variations, i.e. how the nodding of a person depends on the partner with whom they converse with. In general, the work presented here provides a fruitful and interesting basis for future research which would enrich our understanding of Estonian communication and conversational interactions in general.

References

Allwood, Jens; Cerrato, Loredana.; Jokinen, Kristiina.; Navarretta, Costanza; Paggio, Patrizia 2008. *The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena.* Language Resources and Evaluation, Volume 41, Nr. 3-4, pp. 273-287

Allwood, Jens; Nivr,e Joachim; Ahlsén, Elisabeth 1990. *Speech Management on the Non-Written Life of Speech.* Nordic Journal of Linguistics, 13, no. 1

Andonova, Elena; Taylor, Holly A. 2012. *Nodding in dis/agreement: a tale of two cultures.* Cognitive Processing 13, Issue 1 supplement, pp. 79-82

Aryel Beck, Antoine Hiolle, Alexandre Mazel, and Lola Cañamero. 2010. *Interpretation of Emotional Body Language Displayed by Robots.* Proceesings of the 3rd International Workshop on Affective Interaction in Natural Environments (AFFINE'10), Firenze, Italy. pp.37-42

Cassell, Justine 2001. *Nudge Nudge Wink Wink: Elements of Face-to-face Conversation for Embodied Conversational Agents.* Embodied Conversational Agents. MIT Press Cambridge, MA, USA, pp. 1-27

Clark, Herbert H.; Schaefer, Edward F. 1989. *Contributing to Discourse.* Cognitive Science 13, 259-294

Mark Knapp and Judith Hall. 1997. *Nonverbal Communication in Human Interaction.* 4th Edition. Fort Worth: Harcourt Brace

Jokinen, Kristiina; Tenjes, Silvi 2012. *Investigating Engagement – Intercultural and Technological Aspects of the Collection, Analysis, and Use of Estonian Multiparty Conversational Video Data.* Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mar (Eds.). Proceedings of the Eight International Conference on Language Recourses and Evaluation (LREC'12) (2764 - 2769). Istanbul, Turkey: European Language Resources Association (ELRA)

Jokinen, Kristiina 2009. *Natural Language and Dialogue Interfaces.* In C. Stephanidis (Ed.): The Universal Access Handbook. Chapter 31. CRC Press Taylor & Francis Group, pp. 495-506

Jokinen, Kristiina 1996. *Cooperative Response Planning in CDM.* Proceedings of the 11th Twente Workshop on Language Technology: Dialogue Management in Natural Language Processing Systems. Twente, The Netherlands, pp. 159-168

Navarretta, Costanza; Ahlsén, Elisabeth; Allwood, Jens; Jokinen, Kristiina; Paggio, Patrizia 2012. *Feedback in Nordic First-Encounters: a Comparative Study.* Nicoletta Calzolari (Conference Chair); Khalid Choukri; Thierry Declerck; Mehmet Uğur Doğan (Eds.). Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pp. 2494-2499

Norman, Don 1999. *Affordance, Conventions, and Design.* Interactions, May, pp. 38-43

Roger Nunn and Maya Tamura. 2003. *Head Nodding in Intercultural Conversation.* Japan Journal of Multilingualism and Multiculturalism. Vol. 9, No. 1, pp. 69-86.

Poggi, Isabella; D'Errico, Francesca; Vincze, Laura 2010. *Types of Nods. The Polysemy of a Social Signal.* Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, pp. 17-23

Toivio, Emmi; Jokinen, Kristiina. 2012. *Multimodal Feedback Signaling in Finnish.* Human Language Technologies – The Baltic Perspective: Proceedings of the Fifth International Conference Baltic HLT 2012, eds. Arvi Tavast, Kadri Muischnek, Mare Koit, 247-255

Traum, David R. 1999. *Computational Models of Grounding in Collaborative Systems.* Working notes of AAAI Fall Symposium on Psychological Models of Communication, pp. 124-131

Recognition of Human Body Movements for Studying Engagement in Conversational Video Files

Martin Vels
Institute of Computer Science
University of Tartu
Estonia
martin.vels@ut.ee

Kristiina Jokinen
Institute of Computer Science
University of Tartu
Estonia
kristiina.jokinen@ut.ee

Abstract

This paper investigates object recognition techniques to automatically detect human behavior in video conversations. The ViBe background subtraction algorithm, together with standard image processing techniques is applied to conversational videos where two people meet for the first time, and the results show the usefulness of the technique in human communication analysis. By detecting the conversational participants and analyzing their conversational styles through the detected body movements, we can visualize, and draw conclusions concerning the participants' engagement in the communicative activity. The paper discusses these novel observations that show the synchrony and engagement in the participants' behavior.

1 Introduction

The main questions in interaction studies are how people engage themselves in the interaction, how does their focus of attention work (where it is directed to and how it changes over time), and how their body movement, gestures, and other multimodal communication signals help in interaction management (Goodwin, 1981; Kendon, 2004). Humans use several communicative systems ("modalities") and several physical "carriers" of the messages of these systems (vocal sounds and visible hand movements) (Chellappa et al., 1997). The various modalities are not used independently of each other in communication, however, but usually one modality depends on the other modalities, and the communicated information is interpreted in a holistic way, using the signals from all modality channels simultaneously (Jokinen and Wilcock, 2012). Thus, in order to understand human multimodal communication it is necessary to identify those signals that convey communicative meaning, and to understand their intended meanings in the given context. In this respect visual signals and especially human body movements play a central role, besides verbal communication, facial expressions, eye-gazing, breathing, etc. in order to express meanings, emotions, and attitudes, and to coordinate the interaction in general (see e.g. Argyle, 1975)

Concerning the recognition of body movements, relevant questions deal with how humans segment and assign meanings to objects and scenes in their visual field, and what are the appropriate techniques that would enable automatic segmentation and interpretation of visually presented information. Such work contributes to a better understanding of human visual information processing, and also of the requirements for developing intelligent systems and smooth user interactions with such systems. It also opens up new possibilities for building various applications that deal with personal digital devices. It is evident that interaction strategies are important regardless whether the communication takes place face-to-face, via audio-only, or via both video and audio. Advancement of technology in digital devices, like cameras, smartphones, and wearable computers like Google Glass, has made more holistic communication capabilities possible both for the users and the systems, and thus human movement detection is crucial to facilitate natural and flexible interaction.

However, detecting humans (Santhanam et al., 2012) and their gestures (Mitra and Acharya, 2012) from videos is a non-trivial task algorithmically. The quality of the detection depends on the quality of

K. Jokinen and M. Vels. 2015. Proceedings of The 2nd European and the 5th Nordic Symposium on Multimodal Communication.

the video, the illumination and background of the scene, the position of the human compared to the video camera (i.e., facing the camera, standing in profile), what persons in the scene are wearing (color, patterns), occlusion (e.g., hand is behind the body, thus not visible for other party), etc. On the other hand, manual annotation of video data is a common exercise which, however, requires a lot of resources. Work on automatic behaviour annotation using video data concerns e.g.gesture and gesture expressivity (Caridakis et al. 2006; Oikonomopoulos 2006), hand, head and body movements (Allwood et al. 2007), as well as synchrony and body posture analysis using depth cameras (Michelet et al. 2012; Baur et al. 2013). However, this work differs from our approach since we do not only seek to provide and compare annotations with respect to human gesturing, but to study how well the image processing techniques can be applied and modified in order to recognize human body movement in natural conversational interactions.

In this paper we use the video collection of Estonian First Encounter Dialogues from the MINT (Multimodal INTeraction) project (Jokinen and Tenjes, 2012). We have created a technical solution that visually identifies human body movement on video files and tags them with descriptive and quantitative information, thus reducing the work needed for annotating videos manually. We use the ViBe background subtraction algorithm, together with standard image processing techniques to automatically detect human body in the natural conversation dialogue data and we create a diagram representation of the whole video that expresses changes in the detected human body location in the given scene. We also compare the performance of the algorithm to the manually annotated data so as to explore the applicability of the algorithm in human communication studies such as gesture and posture movements, synchrony, and engagement.

The rest of the paper is structured as follows. Section 2 introduces the MINT dataset in more detail. Section 3 describes the algorithms used for human body detection. Section 4 presents the results of the experiments and discusses them with respect to manually annotated data. Finally, Section 5 provides discussion and interpretation of the results, draws conclusions and describes plans for future work.

2 MINT Dataset

The MINT (Multimodal INTeraction) dataset contains 23 videos of the Estonian First Encounters Dialogues. Each of these videos is approximately 5 minutes long and contains two people having a conversation. The videos were recorded by three SonyHDR-XR550V cameras and three external Sony ECM-HW2 wireless microphones. The full 1920x1080 HD quality was used for the recordings. The Sony Vegas Pro 11 software was used to cut, edit and merge, and sync raw video clips to create video files that combined all three cameras, and to export videos to SD format that could be read by further analysis of the files. The resolution of the compressed avi files is 640x360 pixels, with 25 fps. These dialogues were filmed with three cameras from different angles. The first video contains the interaction shown from the center (see Figure 1), the second video shows half frontal view of the person on the left (see Figure 2), and the third video contains half frontal view of the person on the right (see Figure 3).

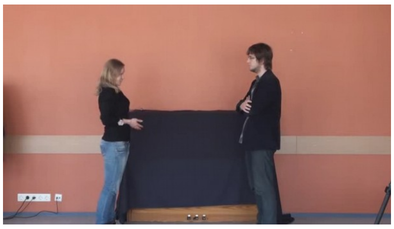


Figure 1: Center view of the scene.

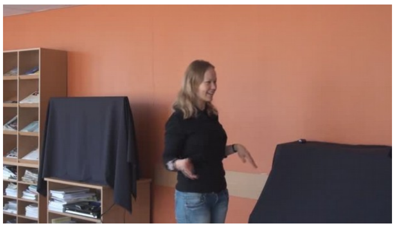


Figure 2: Left view of the scene.

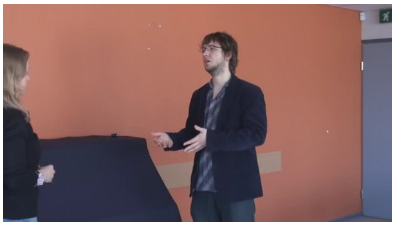


Figure 3: Right view of the scene.

There were 23 different people participating in these scenes, 11 female and 12 male. Each person participated in 2 videos, with different conversation partners in both cases.

3 Person Detection

We analyzed all the center-view videos using a background subtraction technique to detect persons in these videos. After the detection of persons we generated a diagram for each conversation which shows how persons moved back and forth (horizontal movement in the scene) during the conversation. This gave us a quick overview of the activity of the persons in the video. Unfortunately the other possible views (left and right) were not suitable for detecting horizontal human movements properly as the people in these scenes were filmed from a certain angle and not facing camera directly.

3.1 Background subtraction

To detect persons in video we used the ViBe (visual background extractor) algorithm (Barrich and Droogenbroeck, 2011) as background subtraction technique to remove all the image content except for the moving object, which in our case are the moving persons. Background subtraction is a widely used technique for image segmentation and there are various ways how this can be done. We used the ViBe algorithm because it is one of the fastest and most accurate background subtraction algorithms currently available. It was capable of processing videos with resolution of 640x360 pixels and 25 fps in real time.

The basic idea behind background subtraction is trivial. We have the static background frame and the current video frame. We compare these two frames pixel by pixel, remove all the pixels that are the same (background) on both frames, and retain the ones that are different (foreground).

Usually the clean background frame needs to be modeled. This means that we have to build a model of the background, containing more than only one frame and compare our current frame to the whole model to achieve the segmentation of the moving objects. There are several ways to build the background model. Often the background model is built using several sequential frames of the video, which means that the initialization of the model can take a long time (several seconds) and also keeping the model takes a lot of memory. If the frame rate of the video is high (30+ fps), then using this technique means that a lot of memory and computational power are needed. E.g., one gray-scale frame with size 640x480 pixels requires 300kB. Each second of this video is 9MB. Also, even if we build our model from such a video, we still need to analyze about 5-10 seconds of the video (150-300 frames) until we get some idea of moving parts in the video to build a decent model for our background.

The ViBe algorithm (Barrich and Droogenbroeck, 2011) that we were using for segmentation introduced a novel idea where the background model is initialized from a single frame using 8-neighborhood of each of the frame pixel and randomly choosing 20 instances of these neighbor-pixels to build a background model.

The next problem with the background model is to keep it up-to-date when time passes and the scene changes. One of the most widely used techniques is to remove the oldest samples of pixel values from the background model and adding new values from the newer scenes. This method, even if it seems the most natural, may still not be the best solution. Namely, the fact that some of the pixels in the background model are old does not automatically mean that these pixels are no longer correct background representatives. Thus the ViBe algorithm that we are using, had a different approach — it replaces the values in the background model randomly. This technique gave this algorithm a better response speed in case of changing scenes but at the same time not removing correct background pixels from the model just because of the age.

Segmentation itself was simple, comparing each pixel of the current frame to according pixel in our background model and when certain amount of model representatives were close enough, we considered the pixel to be background, otherwise, it was foreground. Euclidean distance was used to determine the closeness of the pixel value to according background model values. It is possible to work in RGB or gray-scale images, the only algorithmic difference is that in case of gray-scale, we only compare one value of the pixel (intensity in the range of 0..255), but in case of RGB-video, we need to compare all the three color-components of the pixel (red, green and blue) to determine the distance between two pixel values. As the results of the segmentation were similar in both gray-scale and RGB-images, we used the gray-scale version because of less computational complexity involved.

3.2 Erosion and Dilation

Finally, when the frame was segmented, we had a black-and-white image (see Figure 4), where zero means background and one means foreground. Now, as there are usually some kind of illumination changes in frames which result in noise in our segmented image, we used two image processing techniques to get rid of the noise. Namely, we used erosion (Gonzales and Woods, 2010) (see Figure 5), which helped us remove all the single pixels. Next, we employed a dilation operation (Gonzales and Woods, 2010) (see Figure 6) operation, which helped us make the interesting objects larger and remove some of the small gaps between close parts of the objects. Erosion and dilation are the morphological operations. Dilation adds pixels to the boundaries of objects in images and erosion removes pixels of object boundaries. The amount of pixels added or removed is determined by the structuring element used during these morphological operations. In our case we used 3x3 square structuring element.

3.3 Object detection

After the segmented image was eroded and dilated it was possible to use a contour detection algorithm (Suzuki and Abe, 1985) available in OpenCV library. This algorithm returns a set of contour areas, which are hierarchically arranged. We used the top-level hierarchy and found the top left and bottom right coordinates of the contours, which had certain size area. This way we were able to find the positions of the moving persons in the frame that were larger than a certain predetermined threshold (see Figure 7).

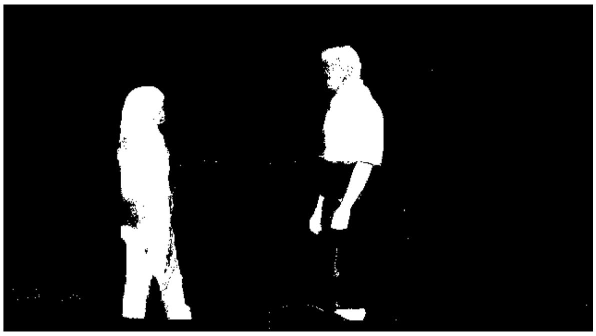


Figure 4: Segmented image.

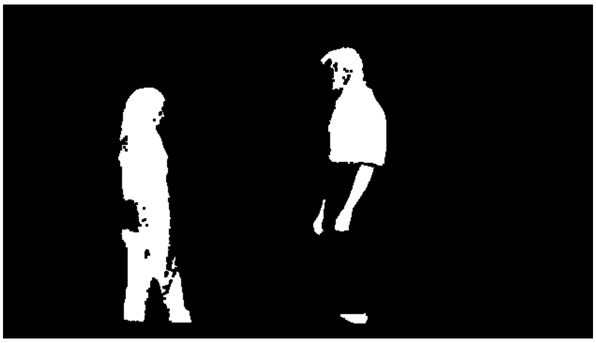


Figure 5: Eroded image.

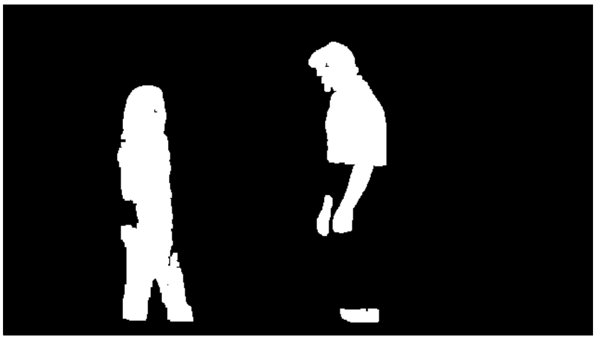


Figure 6: Dilated image.

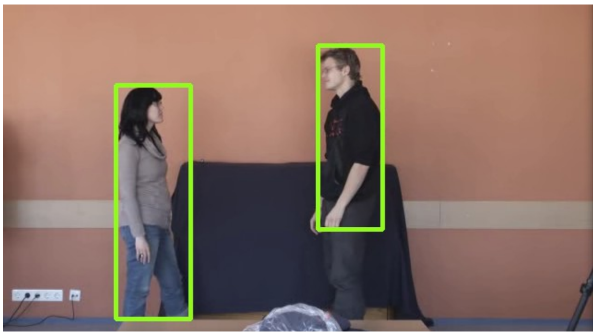


Figure 7: Image with moving persons detected.

4 Engagement in video files

4.1 Hand and body movement

We used left and right coordinates of the surrounding boxes around the persons (see Figure 7) found by the algorithm to draw a diagram that summarized the whole video (see Figure 8). We abbreviate the front and back coordinates of the surrounding box around the person on the left as LFC (left front coordinate) and LBC (left back coordinate). Similarly, the front and back coordinates of the surrounding box around the person on the right is abbreviated as RFC (right front coordinate) and RBC (right back coordinate). We also use the term "person front" when referring to the coordinates that correspond to the person's front and "person back" when referring to the coordinates that correspond to the person's back.

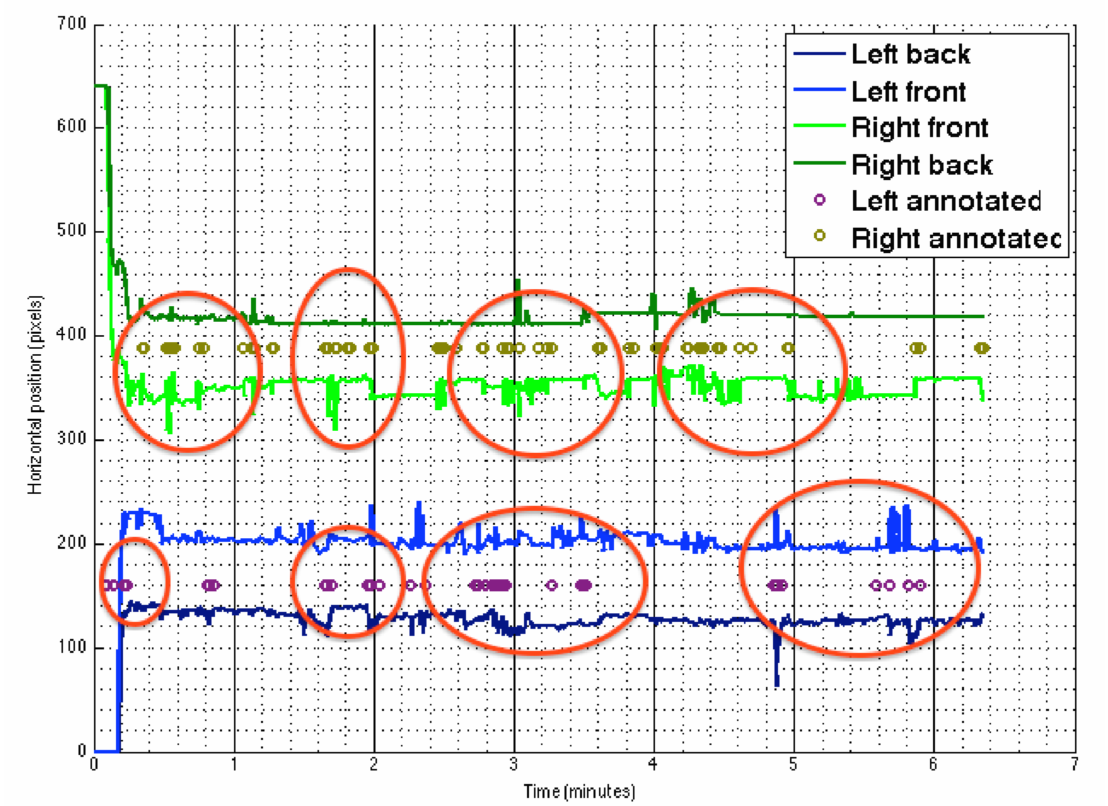


Figure 8: Person movement diagram of the whole video.

The diagram (see Figure 8) shows the movements of the left and the right person in time. Both conversation partners are described with two lines, first showing the back and next the front coordinate of the person. In the middle of these two lines, the hand movements of that particular individual are shown in small circles. Hand movement time was manually marked using annotation software—ANVIL (Kipp, 2001). By comparing the diagrams it can be seen that most of the hand movements can be detected using the front coordinates of the surrounding boxes of the moving persons. However, we also note that there are many cases where movements seem to be detected by the algorithm but are not annotated as communicatively important in the data. This is due to the fact that the size of the surrounding box does not change only due to hand gesturing but also due to person leaning over or moving leg etc. Sudden movements of the head or the whole body of the person are also detected in the changes of the coordinates. The LFC and RFC changes do not automatically indicate hand movements but need to be interpreted in the context of back coordinates.

It must also be emphasized that the comparison of the algorithm with manual hand gesture annotation is meant to visualize the functioning of the algorithm rather than to evaluate its performance against manually annotated data. The large number of "false positive" detections does not indicate the algorithm's oversensitivity to hand gestures, but rather, that the algorithm detects movement in general, and needs to be further tuned in order to detect hand gestures.

Comparing the manually annotated gesture tags with the automatically detected movements of the persons, we notice, however that in many cases it is possible to detect hand gestures from LFC and RFC. Especially if a person stands still, the back coordinate (BC) is steady, too, while the front coordinate (FC) changes rapidly because of the hand movement (see Figure 9 left back and left front). On the other hand, if the person moves frequently, then the FC does not reliably indicate the hand gestures (see Figure 9 right back and right front). As mentioned, the FC does not move only because of the hand gestures but also if the person moves her leg or head or bends forward. Thus, if we only look at the FC changes, the technique is ambiguous between the hand gesture, leg or head movement, and body bending.

We can also detect the handshake of the conversation partners in the beginning of the videos: in Figure 8 the persons do not shake hands, but in Figure 9 they do, as can be seen from the touching of the curves during the starting seconds of the video.

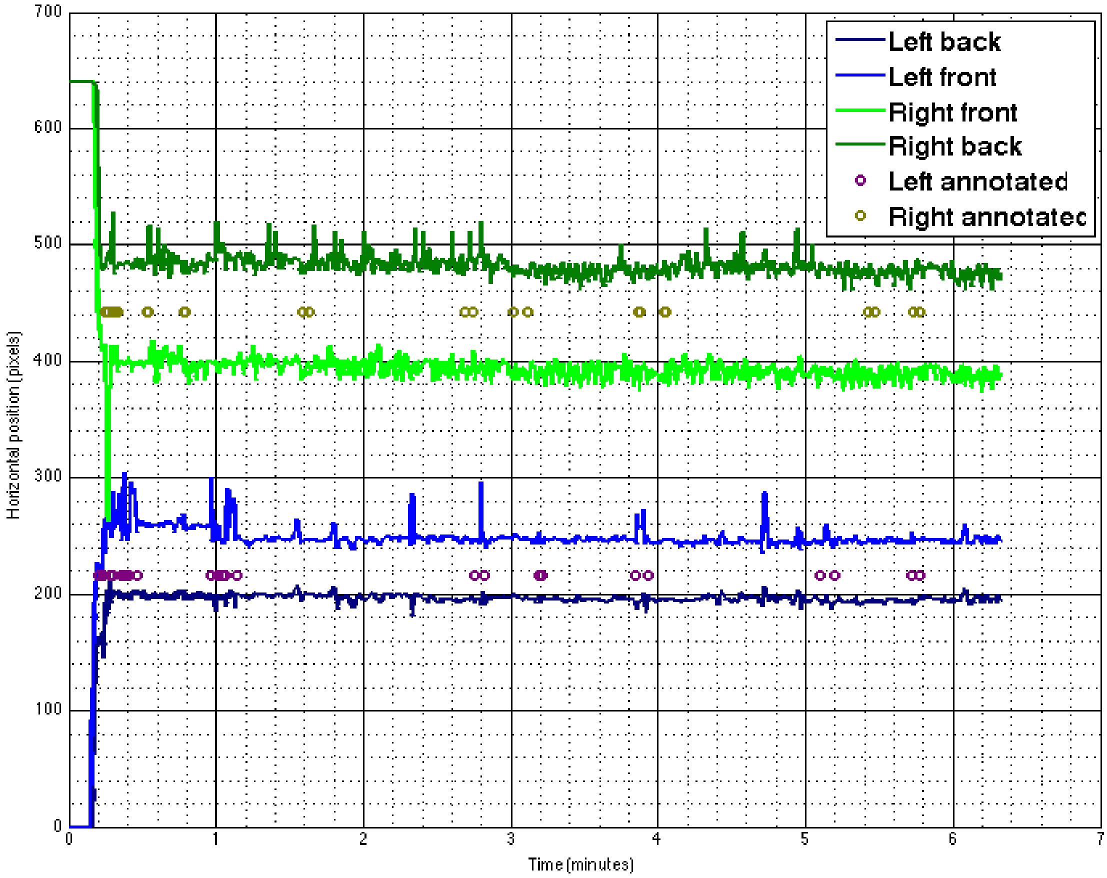


Figure 9: Rhythmic body movement of the person on the right.

4.2 Differences in individual body movements

The method is meant to study the human behavior as a whole, and another interesting behaviour that our diagrams can clearly visualize, is the differences between the conversational style of the interlocutors. We can see that if the partner is mostly standing still or performing many small movements during the conversation (left person in Figure 9). The rhythmic body movement, as can be easily seen in Figure 9 (right front), is an individual property that distinguishes people in the conversation. If the person is using her hands a lot during the conversation we can detect this by comparing the back and front coordinate movements (left person in Figure 10).

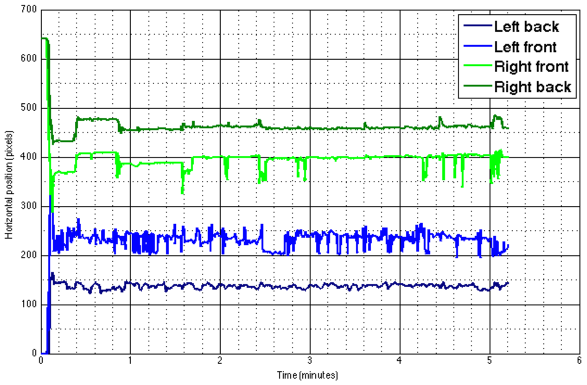


Figure 10: Frequent hand movement of the person on the left. (Front-coordinate changes while back-coordinate keeps still).

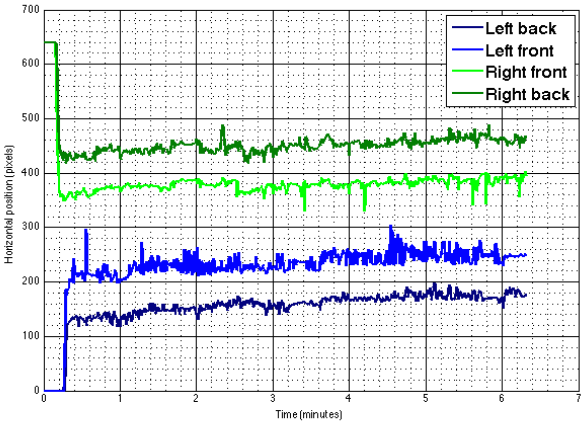


Figure 11: Movement trends. Both persons are moving to the right of the scene during the conversation.

4.3 Synchrony and irregularities

Finally, it is also possible to detect larger trends in position changes of the interlocutors during the conversation, e.g. if one individual is moving away from the other and the second one is following. A nice example of this kind of synchrony can be seen in Figure 11: the participants slowly move to the right of the scene which can be seen from the rising curves. As the participant on the left moves forward, the right participant moves further to the right trying to keep the same distance. The participants thus intuitively adapt their position so that the speaking distance remains constant. The participants' awareness to maintain a comfortable speaking distance intuitively can be used as an indirect measure of synchrony between the participants and of their adaption to the conversational situation. Such subtle movement may not be obvious by looking at the videos only, but our technique makes it concretely visible.

Another possible use of the algorithm is to identify large irregularities in positions of the conversation partners. For instance, if a participant performs a large step away from the regular position, this can be seen as a clear change in the magnitude of the movement among the normal moving of the participants (see Figure 12 and Figure 13).

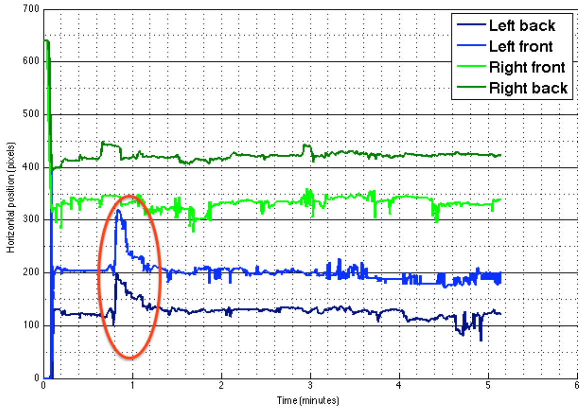


Figure 12: The person on the left performs a sudden movement towards the other person.

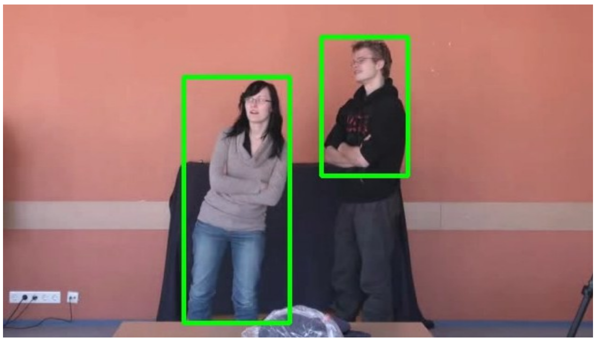


Figure 13: The person on the left is performing an unordinary movement.

5 Conclusion

In this paper we studied human behavior in communication and used image processing techniques to recognize body movement in video conversations. The algorithm focuses on the changes in human movements through the front and back coordinates of the box that surrounds the detected human body. We compared the movement visualization diagrams by the algorithm with the manually annotated gesture tags, and noticed that the algorithm can indicate hand gesturing, but it needs to be further developed to distinguish hand gestures from other movement types which also cause the coordinate of the surrounding box change. However, the technique provides an easy and helpful way to compare the participants' individual conversation styles. Moreover, as a novel contribution, the method can also show the participant's movement trends and irregularities in their behavior, which can effectively be used to study synchrony and adaption between the participants.

The results can be used to gain deeper understanding of human movements and body posture in natural interactions. The method can be used by annotators to find interesting gestures that may not be obvious from videos. They can also consolidate their annotations and check consistency of the annotations with respect to the automatically recognized movements.

The technique is based on empirically collected objective data and applying automatic signal analysis to the data, then comparing and combining this analysis with human perception and top-down analysis of annotated data. The work will thus also contribute to technical development of movement detection, and automatic scene analyses.

We will continue work on these lines to improve the algorithm on the conversational data, especially focusing on the specification concerning gesture recognition. We will also investigate further the synchrony of the participants as observed through their movements

References

Allwood, Jens, Cerrato, Loredana, Jokinen, Kristiina, Navarretta, Costanza & Paggio, Patrizia. 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. In Martin, J.C. et al (eds) *Multimodal Corpora for Modelling Human Multimodal Behaviour*. Special issue of the International Journal of Language Resources and Evaluation, 41(3–4), 273–287, Springer.

Argyle, Michael. 1975. *Bodily Communication*. London: Routledege.

Barnich, Olivier and Van Droogenbroeck, Marc 2011. M., *ViBe: Universal Background Subtraction Algorithm for Video Sequences.* Image Processing, IEEE Transactions on, Vol. 20, pp 1709-1724.

Tobias Baur, Ionut Damian, Florian Lingenfelser, Johannes Wagner and Elisabeth André. 2013. *NovA: Automated Analysis Of Nonverbal Signals In Social Interactions*, HBU'13 Proceedings of the Third international conference on Human Behavior Understanding

Caridakis, G., Raouzaiou, A., Karpouzis, K., Kollias, S. 2006. *Synthesizing Gesture Expressivity Based on Real Sequences*, Proceedings of the LREC 2006 Conference

Chellappa, Rema, Chen, Tsuhan and Katsaggleos, Angelo 1997. *Audio-visual interaction in multimodal communication*, IEEE Signal Processing Mag., pp 37-38.

Gonzales, Rafael C. and Woods, Richard E. 2010. *Digital Image Processing (3rd edition).* Pearson Education, Inc.

Goodwin, Charles 1981. *Conversational Organization: Interaction between Speakers and Hearers.* Academic Press, New York.

Jokinen, Kristiina and Tenjes, Silvi, 2012. *Investigating Engagement - intercultural and technological aspects of the collection, analysis, and use of the Estonian Multiparty Conversational video data*. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC). Istanbul, Turkey.

Jokinen, Kristiina and Wilcock, Graham, 2012. *Multimodal Signals and Holistic Interaction Structuring*. Procs of the 24th International Conference on Computational Linguistics (COLING). Mumbai, India.

Kendon, Adam. 2004. *Gesture: Visual Action as Utterance.* Cambridge University Press.

Kipp, Michael. 2001. Anvil - *A Generic Annotation Tool for Multimodal Dialogue*. Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech), pp. 1367-1370.

Michelet, Stephane, Karp , Koby, Delaherche , Emilie, Achard , Catherine, and Chetouani , Mohamed, 2012. *Automatic Imitation Assessment in Interaction*, HBU'12 Proceedings of the Third international conference on Human Behavior Understanding.

Mitra Sushmita and Acharya, Tinku 2007. *Gesture Recognition: A Survey.* Trans. Sys. Man Cyber Part C 37, 3, pp 311-324.

Oikonomopoulos , Antonios, Patras , Ioannis, and Pantic , Maja, *Spatiotemporal Salient Points for Visual Recognition of Human Actions*, IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics, Vol. 36, No. 3, June 2006

Santhanam, T., Sumathi, C. P. and Gomathi, S. 2012. *A survey of techniques for human detection in static images*. In Proceedings of the Second International Conference on Computational Science, Engineering and Information Technology (CCSEIT '12). ACM, New York, NY, USA, 328-336.

Suzuki, Satoshi. and Abe, Keiichi, 1985. *Topological Structural Analysis of Digitized Binary Images by Border Following.* CVGIP 30 1, pp 32-46.

Breathing in Conversation: an Unwritten History

Marcin Włodarczak, Mattias Heldner
Department of Linguistics
Stockholm University
Stockholm, Sweden
{wlodarczak,heldner}@ling.su.se

Jens Edlund
Speech, Music and Hearing
KTH Royal Institute of Techonology
Stockholm, Sweden
edlund@speech.kth.se

Abstract

This paper attempts to draw attention of the multimodal communication research community to what we consider a long overdue topic, namely respiratory activity in conversation. We submit that a turn towards spontaneous interaction is a natural extension of the recent interest in speech breathing, and is likely to offer valuable insights into mechanisms underlying organisation of interaction and collaborative human action in general, as well as to make advancement in existing speech technology applications. Particular focus is placed on the role of breathing as a perceptually and interactionally salient turn-taking cue. We also present the recording setup developed in the Phonetics Laboratory at Stockholm University with the aim of studying communicative functions of physiological and audio-visual breathing correlates in spontaneous multiparty interactions.

1 Introduction

Human face-to-face communication is known to be inherently multimodal. Specifically, multimodal features have been demonstrated to be closely linked to such basic mechanisms of interaction as turn-taking, grounding and interpersonal coordination. In addition, they have also proved useful in developing dialogue systems and computational models of interaction.

At the same time, while some multimodal cues (gaze, manual gestures, head movements, body posture) have received much attention, others remain as yet unexplored, despite their great potential in highlighting important aspects of human-human and human-computer interaction. In this paper we address one such feature. Namely, we argue that studying breathing in conversation is crucial for understanding how speech production is employed in the coordinated and highly context-sensitive domain of conversation, and call for more research in the field. In particular, in the light of perceptual salience of speech breathing suggested by earlier studies (Whalen et al., 1995; Whalen and Sheffert, 1996), we focus on the role of kinematic and audio-visual correlates of respiration in coordination of speaker change in spontaneous conversation.

In the remainder of this paper we briefly discuss earlier research on speech breathing (Section 2) as well as its possible extensions to the domain of spontaneous conversation (Section 3). Subsequently, in Section 4 we describe our newly established respiratory lab at the Department of Linguistics, Stockholm University.

2 Historical look

Breathing is a primary mechanism of voice generation maintaining a suitable level of subglottal pressure required for momentary production needs. As such, it is implicated in many aspects of speech production, such as voice quality (Slifka, 2006), voice onset time (Hoit et al., 1993) and loudness (Huber et al., 2005). Similarly, breathing has been claimed to enter into processes of speech planning and structuring

K. Jokinen and M. Vels. 2015. Proceedings of The 2nd European and the 5th Nordic Symposium on Multimodal Communication.

(Fuchs et al., 2013). However, in line with the methodological stance dominant in traditional phonetics, breathing has been studied almost exclusively in tightly controlled experiments decoupled from communicative context. Consequently, while these and other studies have made important contributions to speech science, they have largely ignored interactive factors at play in conversation, the most common language use.

At the same time, certain findings stirred by the recent wave of interest in speech respiration indicate that breathing plays an important interactional role. For instance, McFarland (2001) observed that speakers synchronise their respiratory cycles prior to speaker change. It was subsequently shown that the synchronisation is brought about by performing a shared task (Bailly et al., 2013) and is therefore similar to other known examples of interspeaker coordination (Shockley et al., 2009). Indeed, there is some evidence than breathing is linked to synchronisation of speech and gesture (Hayashi et al., 2005) and might even be the basis for synchronisation of movement in general (Pellegrini and Ciceri, 2012).

In addition, the listener's breathing cycle was reported to change depending on such properties of perceived speech as tempo or vocal effort (Rochet-Capellan and Fuchs, 2013). While there is considerable controversy as to the exact nature of the underlying alignment mechanism (or mechanisms), it suggests that breathing is implicated in processes of speech perception. Similarly, on the production side, a variety of kinematic adjustments were found depending on where speech was initiated within the respiratory cycle (McFarland and Smith, 1992), thus indicating sensitivity of the respiratory apparatus to the demands of an upcoming vocal task. Clearly, these mechanisms could be also exploited for conversational needs, for instance to coordinate speaker change.

Last but not least, respiratory data have been demonstrated to improve performance of speech and language technology applications. In particular, including breathing noises in synthetic speech enhances its naturalness (Braunschweiler and Chen, 2013) and recall (Whalen et al., 1995). Improvements in performance were also noted for automatic speech recognition (Butzberger et al., 1992) and automatic annotation of prosody (Wightman and Ostendorf, 1994). Finally, respiratory data were successfully used to detect conversational episodes by automatic discrimination between periods of quiet breathing, listening and speaking (Rahman et al., 2011).

3 Conversational perspectives

In spite of the interactional salience of breathing suggested by the work outlined above, studies of breathing in spontaneous conversation are strikingly rare. Conversation analysis has presented some evidence of how audible inspirations and expirations are used as turn-taking and turn-yielding cues, and how breath holds function as a turn-holding device (Schegloff, 1996; Local and Kelly, 1986). However, these findings have so far not been backed up by a comprehensive quantitative analysis of conversational corpora. Moreover, earlier attempts at quantifying breathing in interaction were based on material which was often not entirely spontaneous (McFarland, 2001; Winkworth et al., 1995). Two notable exception is are recent studies by Rochet-Capellan and Fuchs (2014) and Ishii et al. (2014), which measured breathing patterns during pauses coinciding with speaker change or followed by more speech from the previous speaker.

We argue that breathing in dialogue is a potentially fruitful line of research likely to highlight fundamental principles underlying interspeaker coordination and collaborative human action. Respiratory data could be particularly instructive for investigating mechanisms of turn management. Specifically, as turns are normally preceded by easily perceivable inhalations and followed by equally salient exhalations, audio-visual correlates of respiratory events could be an important extension of the set of the more familiar multimodal turn-taking cues. In addition, respiratory data should allow detecting "hidden events" otherwise not easily available for analysis, e.g. abandoned speech initiation attempts (sharp audible inhalations not followed by speech), thus offering more direct access to speakers' intention to initiate or terminate a turn. Similarly, adaptations of the respiratory cycle prior to speaker change, whose preliminary account was presented by McFarland (2001), could shed new light on the long-standing question of mechanisms behind the observed distributions of gaps and overlaps. Importantly, as breathing is by its very nature an embodied activity, it is also likely to provide a valuable insight into interdepenen-

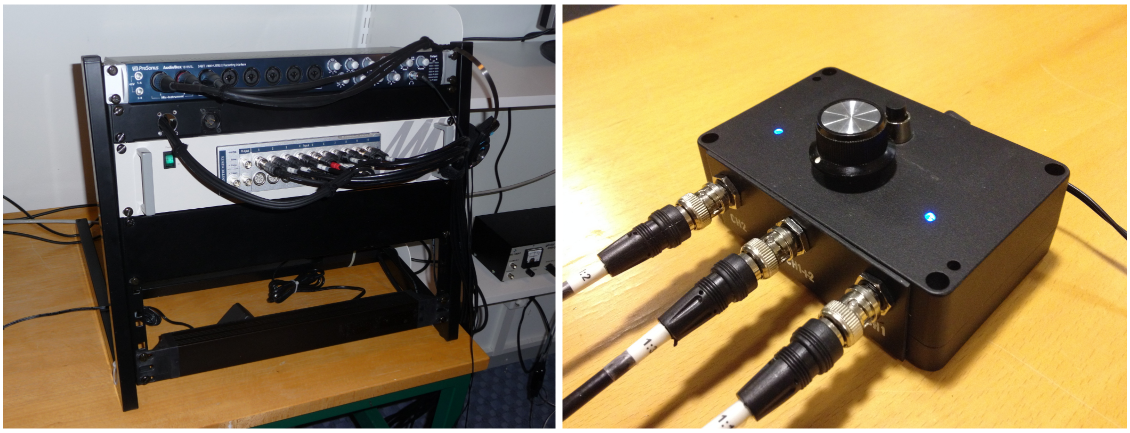


Figure 1: Data acquisition system: PowerLab alongside an audio interface (left) and a RespTrack belt processor (right).

cies between physical and communicative constraints operating in dialogue, for instance the relationship between momentary lung volume and kinematic adaptations prior to speech initiation similar to those found by McFarland and Smith (1992) but set in the fully interactive domain of conversation and subject to temporal constraints of the turn-taking system. Lastly, the links between breathing and other modalities implied by cross-modal synchronisation reported in literature should inform models of sensorimotor coordination both within and between individuals.

In addition to their theoretical significance, studies of respiratory activity in conversation should also help solve some of the key problems in speech and language technology. In particular, loud inhalations might facilitate inferring speaker's intention to initiate a turn and, consequently, provide a shallow, signal-based solution to detecting user barge-ins before their actual onset. Similarly, presence of audible exhalations and breath holds could be used to reason about turn completeness and avoid pause interruptions, which are common in dialogue managers using pause duration as the only turn-yielding cue.

4 Stockholm University Respiratory Lab

In order to answer the questions related to interactional functions of breathing discussed in the previous section, we have developed the following recording setup in the Phonetics Laboratory at Stockholm University. The core of the design is a respiratory inductance plethysmograph (Watson, 1980), which consists of two elastic transducer belts (Ambu RIPmate) measuring changes in cross-sectional area of the rib cage and the abdomen due to breathing. Before each recording, the belts are calibrated using isovolume manoeuvres (Konno and Mead, 1967), which allow estimating contributions of individual belts to the total lung volume change. In addition, vital capacity and resting expiratory levels are also recorded for reference. In order to minimise noise in the signal produced by body movement, participants are recorded standing at a table (about 90 cm high). As the range of respiratory patterns is likely to be sensitive to complexity of turn negotiation and the degree of dialogue competitiveness, we base our studies on multiparty dialogues between three communicative partners.

The belts are connected to dedicated RespTrack processors developed in the Phonetics Lab (see the right panel of Figure 1). The processors were designed for ease of use, and optimised for low noise recordings of respiratory movements in speech and singing. In particular, DC offset can be corrected simultaneously for the rib cage and abdomen belts using a "zero" button. Unlike in the processors supplied with the belts, there is no high-pass filter, thus the amplitude will not decay during breath-holding. A potentiometer allows the signals from the rib cage and abdomen belts to be weighted so that they give the same output for a given volume of air, as well as for the summed signal, enabling direct estimation of lung volume change (see Figure 2).

The signal is recorded by a data acquisition system (PowerLab 16/35 by ADInstruments, left panel

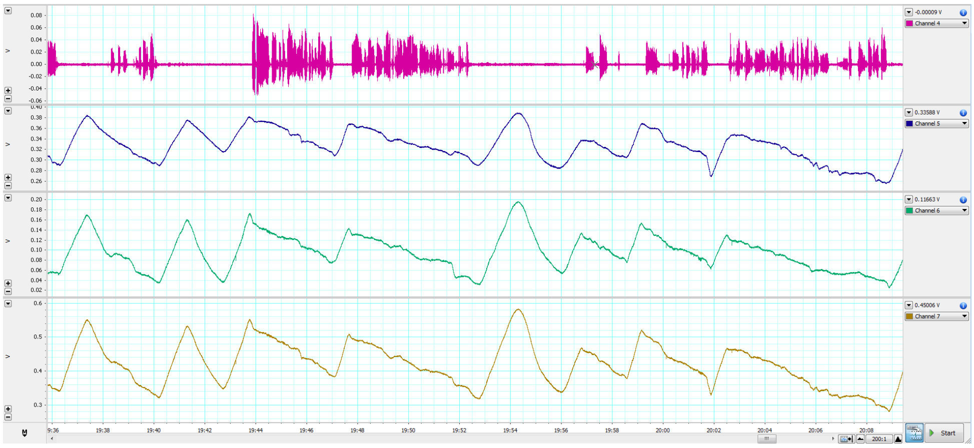


Figure 2: Sample recording for a single speaker: speech (channel 1), respiratory signal from the rib cage and abdomen belts (channels 2 and 3) and the summed respiratory signal (channel 4).

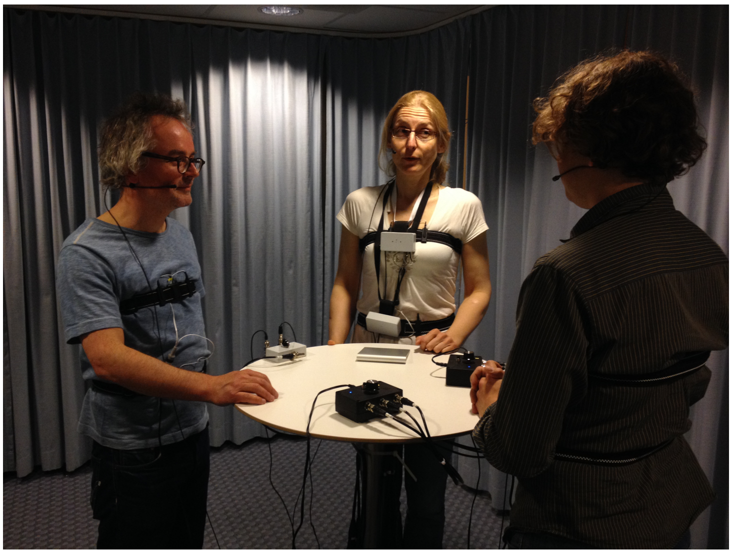


Figure 3: Recording setup. The white boxes are earlier prototypes of the RespTrack processors.

of Figure 1). The system is essentially an analogue-to-digital converter which synchronises the inputs and works with dedicated recording and analysis software (LabChart by ADInstruments). Notably, the system allows connecting other measuring devices, such as airflow masks, which are potentially useful for calibrating the belts. A sample signal is shown in Figure 2.

The setup can be easily adapted to specific recording conditions. For instance, making field recordings is possible by replacing our lab-based data acquisition system with a portable USB-powered unit (DLP-IO8-G Data Acquisition Board by DLP Design). Given the low cost of such devices, they could be also useful for educational purposes, such as student projects.

High quality audio is recorded by close talking microphones (Sennheiser HSP 4) connected to an audio interface (PreSonus AudioBox 1818). The signal is additionally routed to PowerLab to ensure synchronisation with the respiratory trace. As breathing is not only audible but also visible, GoPro Hero3+ cameras are used to record the video.

Our present setup is shown in Figure 3. We are currently conducting a series of pilot studies related to respiratory turn-taking cues as well as temporal patterns of speech initiation within the respiratory cycle. Preliminary results were presented in Aare et al. (2014).

Given that we are particularly interested in communicative functions of audible inhalations and exhalations, we are experimenting with alternative methods of recording clear respiratory noises. Two variants are being assessed: one in which a dedicated close-talking microphone is placed directly in front of the mouth and one which uses a contact microphone placed on the neck near the larynx (throat microphone). A further extension of the recording setup consists in using thermistor probes placed in speakers' nostrils, which should allow differentiating between breathing through the nose and through the mouth.

The resulting corpus will be segmented into (semi-)automatically derived stretches of speech and silence in the audio signal, and inhalations and exhalations in the respiratory signal. In addition, selected dialogue act categories (interruptions, backchannels, disfluencies) will be annotated. The data set will be made public for research use.

5 Conclusions

This paper has aimed at pointing out potential interest and relevance of respiratory activity to fundamental mechanisms of conversation related to turn management. We have argued that the topic has been long overlooked in breathing research and is ripe for systematic quantitative investigation, especially in the light of the existing evidence of multifaceted interactions between breathing and speech production and perception as well as its possible applications in speech technology. We have also described a recording setup developed at Stockholm University required for such a data collection and analysis effort. We hope to see respiratory activity taking its legitimate place among other better studied multimodal features in the nearest future.

Acknowledgements

The research presented here was funded in part by the Swedish Research Council project 2014-1072 *Andning i samtal* (*Breathing in conversation*).

References

Kätlin Aare, Marcin Włodarczak, and Mattias Heldner. 2014. Backchannels and breathing. In *Proceedings of FONETIK 2014*, pages 47–52, Stockholm, Sweden.

Gérard Bailly, Amélie Rochet-Capellan, and Coriandre Vilain. 2013. Adaptation of respiratory patterns in collaborative reading. In *Proceedings of Interspeech 2013*, pages 1653–1657, Lyon, France.

Norbert Braunschweiler and Langzhou Chen. 2013. Automatic detection of inhalation breath pauses for improved pause modelling in HMM-TTS. In *Proceedings of the 8th ISCA Speech Synthesis Workshop*, pages 1–6, Barcelona, Spain.

John Butzberger, Hy Murveit, Elizabeth Shriberg, and Patti Price. 1992. Spontaneous speech effects in large vocabulary speech recognition applications. In *Proceedings of the workshop on Speech and Natural Language*, pages 339–343. Association for Computational Linguistics.

Susanne Fuchs, Caterina Petrone, Jelena Krivokapić, and Philip Hoole. 2013. Acoustic and respiratory evidence for utterance planning in German. *Journal of Phonetics*, 41(1):29–47.

Koji Hayashi, Nobuhiro Furuyama, and Hiroki Takase. 2005. Intra-and inter-personal coordination of speech, gesture and breathing movements. *Transactions of the Japanese Society for Artificial Intelligence*, 20(3):247–258.

Jeannette D. Hoit, Nancy Pearl Solomon, and Thomas J. Hixon. 1993. Effect of lung volume on voice onset time (VOT). *Journal of Speech, Language and Hearing Research*, 36(3):516–521.

Jessica E. Huber, Bharath Chandrasekaran, and John J. Wolstencroft. 2005. Changes to respiratory mechanisms during speech as a result of different cues to increase loudness. *Journal of Applied Physiology*, 98(6):2177–2184.

Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. 2014. Analysis of respiration for prediction of “who will be next speaker and when?” in multi-party meetings. In *Proceedings of the 16^th ACM International Conference on Multimodal Interaction (ICMI 2014)*, pages 18–25, Istambul, Turkey.

Kimio Konno and Jere Mead. 1967. Measurement of the separate volume changes of rib cage and abdomen during breathing. *Journal of Applied Physiology*, 22(3):407–422.

John Local and John Kelly. 1986. Projection and 'silences': Notes on phonetic and conversational structure. *Human studies*, 9(2):185–204.

David H McFarland and Anne Smith. 1992. Effects of vocal task and respiratory phase on prephonatory chest wall movements. *Journal of Speech and Hearing Research*, 35(5):971–982.

David H. McFarland. 2001. Respiratory markers of conversational interaction. *Journal of Speech, Language and Hearing Research*, 44(1):128–143.

Raffaella Pellegrini and Maria Rita Ciceri. 2012. Listening to and mimicking respiration: Understanding and synchronizing joint actions. *Review of Psychology*, 19(1):17–27.

Md. Mahbubur Rahman, Amin Ahsan Ali, Kurt Plarre, Mustafa al'Absi, Emre Ertin, and Santosh Kumar. 2011. mConverse: Inferring conversation episodes from respiratory measurements collected in the field. In *Proceedings of the 2nd Conference on Wireless Health*, pages 1–10, San Diego, CA.

Amélie Rochet-Capellan and Susanne Fuchs. 2013. Changes in breathing while listening to read speech: the effect of reader and speech mode. *Frontiers in Psychology*, 4(906):1–15.

Amélie Rochet-Capellan and Susanne Fuchs. 2014. Take a breath and take the turn: How breathing meets turns in spontaneous dialogue. *Philosophical Transactions of the Royal Society B*, 369(1658):1–10.

Emanuel A. Schegloff. 1996. Turn organization: One intersection of grammar and interaction. *Studies in Interactional Sociolinguistics*, 13:52–133.

Kevin Shockley, Daniel C. Richardson, and Rick Dale. 2009. Conversation and coordinative structures. *Topics in Cognitive Science*, 1(2):305–319.

Janet Slifka. 2006. Some physiological correlates to regular and irregular phonation at the end of an utterance. *Journal of Voice*, 20(2):171–186.

H. Watson. 1980. The technology of respiratory inductive plethysmography. In F. D. Stott, E. B. Raftery, and L. Goulding, editors, *Proceeding of the Second International Symposium on Ambulatory Monitoring (ISAM 1979)*, pages 537–563, London. Academic Press.

Doug H. Whalen and Sonya M. Sheffert. 1996. Perceptual use of vowel and speaker information in breath sounds. In H. Timothy Bunnell and William Idsardi, editors, *Proceedings of ICSLP 96*, pages 2494–2497.

Doug H. Whalen, Charles E. Hoequist, and Sonya M. Sheffert. 1995. The effects of breath sounds on the perception of synthetic speech. *The Journal of the Acoustical Society of America*, 97:3147–3153.

Colin W. Wightman and Mari Ostendorf. 1994. Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing*, 2(4):469–481.

Alison L. Winkworth, Pamela J. Davis, Roger D. Adams, and Elizabeth Ellis. 1995. Breathing patterns during spontaneous speech. *Journal of Speech, Language and Hearing Research*, 38(1):124–144.

Towards gesture-based literacy training with a virtual agent

Kirsten Bergmann
Bielefeld University
Faculty of Technology, CITEC
P.O. Box 100 131
33501 Bielefeld, Germany
kirsten.bergmann@uni-bielefeld.de

Abstract

Illiteracy remains a persistent problem all over the world. Especially people with cognitive impairments are affected and, at the same time, often excluded from interventions because courses and e-learning materials are not tailored to the particular needs of these people. A teaching method that is occasionally applied in special schools is a gesture-based training: Learning of letter-sound pairs is supported by associated gestures. In this paper first steps towards realizing this methodology with a *virtual* teacher are presented. We apply a user-centered approach and present a first prototype system which is used to set up an evaluation with learners from the target audience. In the study, participants learned with the system in consecutive learning sessions over three days. We present results with regard to the general acceptability of the system, problems the learners had as well as issues of the system they liked. The outcome of the study serves to inform future versions of the prototype system.

1 Introduction

Illiteracy is a serious problem given that reading and writing are basic skills in our contemporary information society. According to the United Nations, 15.9% of the world population is illiterate[1]. In the European Union, illiteracy in the strict sense that people are not able to read and write a single word, is nearly eradicated. However, the phenomenon of *functional illiteracy* in adults is becoming increasingly serious, i.e., people can read or write single words and short sentences, but no longer sentences or continuous text. One in five young people in Europe has such poor reading and writing skills[2]. Insufficient literacy has severe consequences. People with low literacy are less likely to finish school, more likely to be unemployed, especially in times of crisis, and more likely to suffer from poor health. Although various intervention programs have been set up over the past decade (2003-2012 was the United Nations Literacy Decade), very little progress has been made in terms decreasing illiteracy rates, so far. In other words, there is an urgent need for effective literacy programs.

One group that is particularly affected by illiteracy are people with learning difficulties or cognitive impairments. For these people, the situation is further hampered because courses and training materials are not tailored with regard to their special needs. They are typically reliant on individual tutoring which is expensive and rarely available–even in specialized schools. This target group is, at the same time, very interesting to develop courses or other kinds of literacy training materials with. The outcome of projects developing for and with handicapped people are typically systems which are particularly easy to use and can, hence, also be used easily by others (contrary to the other way around).

An interesting approach to overcome illiteracy has been described in the literature for a long time: gesture-based literacy training (Koch, 1939; Kossow, 1979; Bleidick and Kraft, 1966; Radigk, 1975). Basically the idea is to support letter-sound integration by gestures. That is, students learn to associate

[1] UNESCO Institute for Statistics, September 2012

[2] Programme for International Student Assessment (PISA), 2009

K. Jokinen and M. Vels. 2015. Proceedings of The 2nd European and the 5th Nordic Symposium on Multimodal Communication.

not only a particular letter with a sound, but also a related gesture. The methodology has found its way especially into schools specialized for handicapped students. There is, however, no standard system available and every teacher employs the methodology with her own gesture repertoire. So pupils are running into problems when they have to change classes or schools. Likewise, adult learners are struggling when hey change courses or when their course it taught by a different teacher for any reason.

At the same time, research on pedagogical agents is providing virtual characters to be applied in the role of teachers, tutors or trainers (Heidig and Clarebout, 2011). These characters can train learners with flexible and customized multimodal materials, and can motivate them through supportive feedback (Baylor and Kim, 2009). Research with virtual agents engaged as trainers in learning tasks has actually shown that gesture-supported learning can increase memory performance for vocabularies (Bergmann and Macedonia, 2013). Moreover, with regard to acceptance, previous research has shown that people with cognitive impairments show a high degree of acceptance and desirability of personal virtual assistants for daily use (Yaghoubzadeh et al., 2013; Kramer et al., 2013).

In this paper, we aim to integrate these two directions in a gesture-supported literacy training with a virtual human for people with cognitive impairments. Applying a virtual character as a teacher in the sensitive domain of literacy education for adults might be particularly advantageous as illiterate people often feel embarrassed when they have to admit their illiteracy to other people. This is one of the reasons why they do not attend courses or make demands on other kinds of instruction. To tailor the system to the particular needs of the target group, we apply a user-centered design approach (Abras et al., 2012). In this methodology, one builds prototypes already at an early stage and improves these prototypes iteratively on the basis of tests and evaluations with users. Within this methodological framework, we set up a prototype system for several short training sessions and evaluated this with participants from the target audience. In the evaluation we were particularly interested in the general feasibility of the approach, in participants' acceptance and appreciation, and potential problems or challenges from the participants' perspective. The remainder of the paper is structured as follows. In Sect. 2 we provide background information about gesture-based literary training. In Sect. 3 we present a first prototype system which allowed us to set up consecutive three training sessions. In Sect. 4 an evaluation study and first results thereof are reported. We conclude in Sect. 5 with a summary and discussion.

2 Gesture-based literacy training

Gesture-based systems to overcome illiteracy have their origin in very early work, for instance, by Ickelsamer (1530), Grosselin (1866) and Piper (1898) (cf. Schäfer (2011)). At the beginning of the 20th century, Franz-Joseph Koch proposed the 'Fingerlesemethode' (engl. 'finger reading method'; Koch (1939)) which spread the idea of becoming literate with gestures into schools. In Koch's approach speech sounds are associated with natural sounds from the environment or from everyday life and accompanied with related hand and arm movements. For instance, the IPA 'ʃ' is associated with clapping hands because the 'ʃ' occurs in the word 'to shoo' (German: 'scheuchen'). The gestures are performed in front of the face while performing the very sound.

Subsequently, others started to develop various gesture-based systems that were based on Koch's method. Kossow (1979), for instance, set up a gesture set consisting of different signs for vocals and consonants, respectively. Signs for vocals depict the mouth posture while the sound is being articulated. For consonants, the hands are brought into specific position and configurations close to the mouth such that the learner is enabled to sense the air stream, lip configuration and larynge movement. Overall, Koch's and Kossows systems are rather *fine-motoric*, characterized by particular handshapes and hand configurations.

By contrast, others have developed gesture sets which are rather gross-motoric. That is, the employed gestures are characterized by rather large movements of hands and arms. Bleidick & Kraft (1966), for instance, combined hand and arm gestures established in music education and eurythmy (a movement art developed at the beginning of the 19th century). In Bleidick & Kraft's approach the employed gestures fall into the two categories for vocals and consonants. Vocals are depicted statically depending on their pitch (e.g., 'e' is high, 'u' is low). Consonants are depicted dynamically. Another gross-motoric system has been developed by Radigk (1975). Here, the selected gestures come from different domains. Some are taken from speech therapy, others bear some resemblance with the shape of letters, e.g., the gesture for 'T' is a T-shaped configuration of both arms.

These methods have been employed successfully in many schools in the 1970s and 1980s, but they were displaced over the years by more modern teaching methods. Only in special schools for handicapped children they have been kept to date. Here, gesture-based methods to become literate got integrated into the curricula because they are a major help for weak pupils in particular. Recently, the gesture-based literacy teaching is becoming more popular again. For instance, private schools in Germany ('Hasenschulen'[3]) are established where children who did not manage to become literate at their regular school are taught successfully with Koch's finger reading method.

It appears that especially the fine-motoric gesture sets (especially Koch's approach) found their way into usage. So the question might come up whether these have any advantages for the learners. To date (and to the best of the author's knowledge), there are no comparing studies of the two approaches. To explore this question, we will realize parts of both, fine- and gross-motoric approaches, in our prototype system. We can, hence, evaluate whether learners have particular problems with one system or the other, or whether they have preference for any of the two approaches.

3 Prototype System

We set up a prototype system to realize the basic idea of a virtual agent engaging in the role of a teacher for literacy education with cognitively impaired learners. The prototype employs the virtual character 'Billie' driven by the ASAP realizer system for multimodal behavior realization (van Welbergen et al., 2014). The agent's behavior is specified in the Behavior Markup Language (BML; Vilhjalmsson et al. (2007)). Speech is synthesized with the text-to-speech system MaryTTS employing a German voice (Schröder and Trouvain, 2003). The current prototype is implemented as a wizard-of-oz system. That is, instead of automatic perception components (speech recognition, gesture recognition etc.) a human wizard interprets the learners' verbal and nonverbal behavior and initiates the agent behavior. Basically, we set up four different kinds of teaching units: (i) Letter introduction units in which the agent introduces a new letter to the learner, (ii) Letter repetition units in which an already introduced letter was further trained in interaction between agent and learner, (iii) Joint reading units in which agent and learner read words (consisting of known letters) together, and (iv) Independent reading units in which the learner is encouraged to read words on her own. In the following we present the main features of these four units.

Introducing a letter. The letter to be learned is displayed on a blackboard behind the virtual character. The agent uses a pointing gesture to refer to the letter and names it for the learner. Subsequently the agent lists some example words in which the letter occurs (see Table 1). Next, the agent performs the gesture for the letter to be learned and encourages the learner to perform the same movement together with him. Before executing the gesture once again, the agent explains some peculiarities of the movement and then performs the gesture three more times together with the learner. Next, the learner is assigned to repeat the gestural movement another three times on her own while also performing the speech sound associated with the letter. Finally, the agent asks the learner whether she wants to repeat the letter once again and if the learner repeats with 'yes', another three repetitions of gesture and spoken sound are conducted.

Repetition of a letter. To repeat a particular sound-letter-gesture combination, the letter was again displayed on the blackboard. The agent encourages the learner to perform the very movement together with him. Then the virtual character performs the gesture to be imitated or performed simultaneously.

Joint reading of a word. The letters of the word to be read (consisting of known letters) are displayed on the blackboard. The virtual character performs the sequence of gestures and speech sound associated with the letters respectively. Subsequently, learners are encouraged to read words together with the agent by performing the gestures and speech sounds simultaneously.

[3] www.hasenschule.de

Independent reading of a word. The letters of the word to be read (consisting of known letters) are displayed on the blackboard. The agent invites the learner to try to read the words on her own. Whenever learners need help, they can request support by the virtual teacher who can either help out with the spoken sound of a letter and its accompanying gesture, or with the complete (spoken and gesticulated) word. For the case that no help was necessary, the agent just stands still. When the learner reads the word successfully, the virtual character provides positive feedback and compliments the learner. When the word is not read out correctly by the learner, the agent encourages the learner to try it once again.

4 Evaluation study and first results

In the evaluation study we tested the prototype system described above. Note that this evaluation is not a controlled study testing a readily implemented system. It is rather a first qualitative evaluation of a prototype within the scope of a user-centered design. That is, the results are to gain insights into general acceptability of the system, to identify potential problems the learners might have as well as favorable issues of the system. The outcome of the study serves to inform future versions of the prototype system.

Materials The prototype system was brought to application with content in terms of six letters: the three vowels ‘A’, ‘E’, and ‘U’ as well as the three consonants ‘B’, ‘S’, and ‘T’. The gestures for these letters were realized in two ways: (1) in the gross-motoric approach by (Bleidick & Kraft, 1966) and (2) in the fine-motoric approach by Koch (1939). According to the latter, the ‘S’ is taught in two variants, namely voiced and voiceless. See Figure 1 for visualizations of the gestures as performed by the virtual agent. This distinction increases the number of letters being realized in the fine-motoric approach to seven. The gestures are taught together with particular approach-specific example words. Here we employed those words that have been which are provided by Koch (1939) and Bleidick & Kraft (1966) for their proposed gestures sets, respectively (see Table 1). For Koch’s fine-motoric approach these example words were children’s names and names for everyday objects. For Bleidick & Kraft’s gross-motoric approach the words referred to the shape of the letters and gestures.

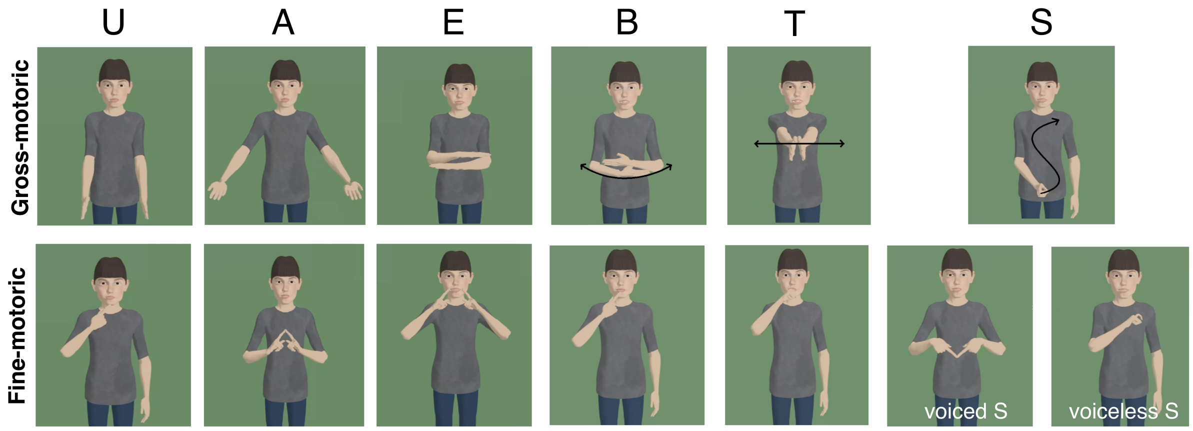


Figure 1: Gestures performed by the virtual agent for the letters taught in the evaluation study: Gross-motoric gestures (Bleidick & Kraft, 1966) in the upper row, fine-motric gestures (Koch,1939) in the bottom row.

Procedure The experimental procedure comprised three learning sessions taking place on three consecutive days. Upon arrival on day one, participants were informed about the course of the study by the experimenter. Next, they fulfilled a standardized pre-test to assess their literal skills[4] and were assigned for training either with the fine-motoric or the gross-motoric system. Subsequently, the first lesson started with a welcome by the virtual agent. The further course of the session follows the schedule as

[4] PROSON Kompetenzfeststellung Sprechen, Lesen und Schreiben (Engl.: PROSON Assessment of competence in speaking, reading and writing)

Table 1: Example words for the letters employed in the letter introductions. The fine-motoric system (Koch, 1939) employs names for everyday objects and children's names. The words in the gross-motoric system (Bleidick & Kraft, 1966) refer to the shape of the gesture for the very letter.

Letter	**Example words**	
	Fine-motoric approach (Koch, 1939)	**Gross-motoric approach** (Bleidick and Kraft, 1966)
A	Arm (Engl. 'arm')	Ankunft (Engl. 'arrival')
	Augen (Engl 'eyes')	Abend (Engl. 'evening')
	Anna	Abschied (Engl. 'farewell')
E	Elefant (Engl. 'elephant')	Wehe (Engl. 'woe')
	Emil	Gefahr (Engl. 'danger')
	Esel (Engl. 'donkey')	Erschrecken (Engl. 'frighning')
U	lustig (Engl. 'funny')	gruselig (Engl. 'creepy')
	Fuchs (Engl. 'fox')	unten (Engl. 'down')
	Schule (Engl. 'school')	dunkel (Engl. 'dark')
S		sofort (Engl. 'immediately')
		Stille (Engl. 'silence ')
		Pssst (Engl. 'Shhh')
S (voiceless)	Essen (Engl. 'food')	
	anders (Engl. 'different')	
	Horst	
S (voiced)	Summen (Engl. 'buzzing')	
	Senf (Engl. 'mustard')	
	Sardine (Engl. 'sardine')	
T	Tanz (Engl. 'dance')	Trommel (Engl. 'drum')
	Traum (Engl. 'dream')	Ton (Engl. 'tone')
	Tino	Tuba (Engl. 'tuba')
B	Baum (Engl. 'tree')	Bauch (Engl. 'belly')
	Ball (Engl. 'ball')	Bett (Engl. 'bed')
	Bernd	Baby (Engl. 'baby')

summarized in Table 2 consisting of several of the teaching units described in Sect 3. For instance, session 1 comprised of three introductory units for the vowels ‘A’, ‘E’ and ‘U’, respectively. Each session was completed with a post-session questionnaire which was read to the participants. Participants’ oral answers were written down by the experimenter. Questions covered the following issues:

- Assessment of the gestures (e.g., clarity of the agents’ gestures, difficulty of performing the gestures)
- Assessment of the agent’s overall behavior (e.g., comments given by the agent)
- Overall assessment of the system (e.g., joy of interaction with ‘Billie’, joy of learning with body movements)

Participants Cognitively impaired participants were recruited from the clientele of an institution where people of all ages with various cognitive impairments can attend computer and photography courses (PIKSL lab[5]). Participants (n=5, 3 male, 2 female, aged 24 to 57) had different degrees of illiteracy. As measured with the PROSON literacy test (see above) two of them were illiterate (lowest level A01), and three of them were classified as literate (highest level A03). As we were mostly interested in acceptance and potential problems with the learning approach, we did not restrict participation to illiterate people.

Table 2: Course of the three training sessions taking place on three consecutive days.

Session 1	Introduction: ‘A’, ‘E’, ‘U’ Questionnaire
Session 2	Repetition: ‘A’, ‘E’, ‘U’ Introduction: ‘B’, ‘S’, ‘T’ Questionnaire
Session 3	Repetition: ‘A’, ‘E’, ‘U’, ‘B’, ‘S’, ‘T’ Joint reading: ‘BUS’, ‘AST’, ‘TABU’, ‘UTE’, ‘ASS’ Independent reading: ‘BETT’, ‘TEST’, ‘SAAT’, ‘TUBA’, ‘TUBE’ Questionnaire

Results First of all, all participants managed to interact with the system easily. All of them saw the sessions to the finish and no one quit her participation early, although the study stretched across several days.

One goal of the evaluation was to gain insights into problematic and challenging issues from the participants’ perspective. It turned out that training with the gross-motoric system by Bleidick & Kraft (1966) resulted in problems with one out of six letters, namely the ‘S’. All three participants who trained with the gross-motoric gestures had problems with this letter. They stated that the S-gesture was either difficult to recognize when performed by the virtual character, or hard to remember. This subjective impression of the participants is in line with observations from the interactions. No one of the three participants was immediately able to imitate the S-gesture correctly. All participants repeated the S-gesture more often than the other gestures in order to memorize it. The direction of movement (from bottom to top) seemed to be counterintuitive for the participants and was, in particular, hard to memorize. For one of the three participants, the difficulties in gesture performance and memorization had an impact on the feeling of joy during learning. This learner reported less fun while learning and repeating the letter ‘S’ as compared to the other five letters. The other two participants who also had problems with the S-gesture, nevertheless, reported a comparable degree of fun for all six gestures.

For participants training with the fine-motoric system by Koch (1939) the amount of problematic gestures/letters was higher as compared to the gross-motoric system. Both participants had problems with the vowel ‘A’. They reported that the ‘A’ was particularly difficult and their gesture performance

[5] http://www.piksl.net

for the ‘A’ was not exact. One of the participants, moreover, reported that the gestures for the other two vowels ‘E’ and ‘U’ were hard to recognize because the gestures were performed directly in front of there agent’s face. In session 2, when the consonants were trained, both participants had problems with the letters ‘B’ and ‘U’ which were apparently too similar. These letters were mixed up by both participants. Another problem was the voiceless ‘S’. One participant reported problems to recognize this gesture, especially when it had to be distinguished from the voiced ‘S’ gesture. Both participants had obvious problems with the performance of the voiceless ‘S’ gesture such that the gesture was not performed correctly.

Another aspect to look at is participants’ independent reading ability in the end. Were the learners able to read the five selected words at the end of session 3? As expected, the three literate participants had no problems reading the words. Among the illiterate participants, the person who trained with the gross-motoric system, managed to read three out of five words in the end. The person who trained with the fine-motoric system, was able to read one out of five words in the end. These results have to be seen in relation with the proportion of gestures learned correctly. Here, the participants who learned the gross-motoric gestures were able to remember the majority of gestures correctly. The illiterate participant who learned the gross-motoric gestures performed 80% (five out of six) gestures correctly. Among the learners with the fine-motoric system, the literate person was able to perform four out of seven gestures correctly (57%) and the illiterate person three out of seven gestures (43%). See Figure 2 for a visualization of learning results.

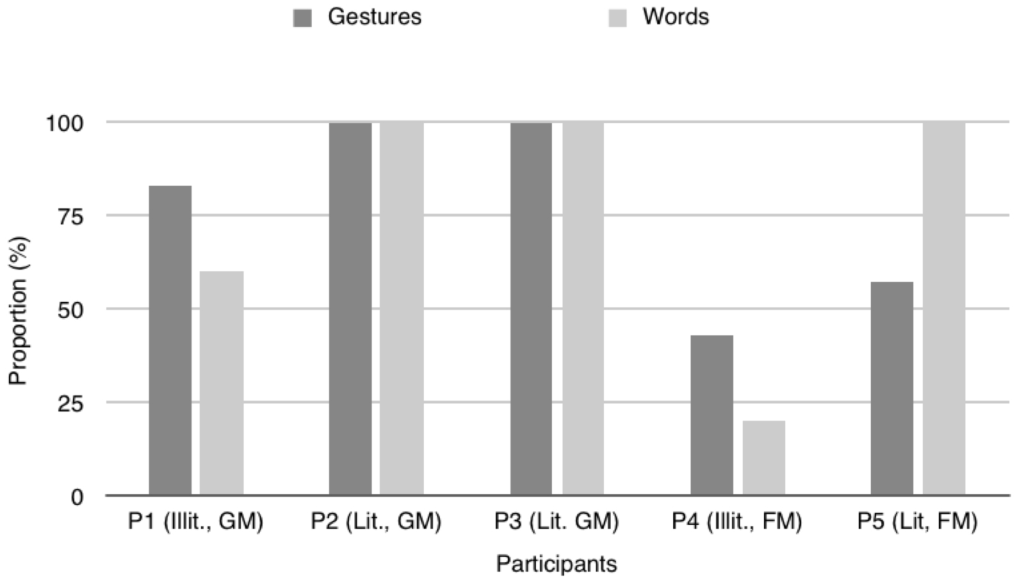


Figure 2: Learning outcome: Proportion of gestures performed correctly and words read independently and correctly at the end of training session 3. Participants were classified as literate (Lit.) or illiterate (Illit.) and learned either with a gross-motoric gesture set (GM) or a fine-motoric gesture set (FM).

Moreover, we were interested in the learning experience: How much did the participants enjoy the sessions? And is there a motivational effect to train with a gesture-based program beyond the scope of the study? Generally, all participants reported high degrees of fun. They enjoyed both, interacting with the virtual character as well as using gestures to become illiterate. One participant (who had significant problems with performing and remembering the fine-motoric gestures) reported a low degree fun with the gesture-based learning at the end of session 2, but still enjoyed the interaction with the virtual agent. Moreover, the majority of participants evinced to further train their reading abilities with gesture support apart from the fixed training sessions with the virtual agent. Only one participant (again the illiterate person who had problems with the fine-motoric gestures) raised concerns that this might be difficult especially when Billie would not be present.

We were further interested in issues that participants found helpful with respect to learning, motivation etc. Four out of five participants noted that the virtual character's explanations of letter shape and comments on peculiarities of the movement were particularly helpful. One participant put it like this "Billie's explanations were great. They helped to recognize and understand everything". It could also be observed that gestures were performed more exactly after the agent had given/repeated his explanations. Moreover, participants judged the repetitions of previously learned letters and gestures at the beginning of sessions 2 and 3 as very helpful. One participant noted that these repetition helped to gain certainty. Another comment by one participant highlighted the motivational effect of positive feedback provided by the agent: "When Billie said that I did well, I felt especially motivated. Then I wanted to proceed immediately.".

Finally, another comment by a participant stressed the beneficial effect of an *artificial* character in the role of a teacher: 'It is very embarrassing for me that I'm illiterate. In interaction with Billie this does not matter, but I do not like to talk about this with other people. And when I'm making mistakes, this does not matter in interaction with Billie as well.". This assessment highlights the advantage of a virtual agent over a human teacher given that people feel unpleasant in their role of illiterate learners.

5 Conclusion

In this paper, we presented first work towards bringing gesture-supported literacy training together a virtual human in the role of a teacher for people with cognitive impairments. In a user-centered design approach we set up several short training sessions and evaluated these with participants from the target group. Two different gesture sets came to application: A gross-motoric set developed by Bleidick & Kraft (1966) and a fine-motoric set developed by Koch (1939) (often applied in past and present classroom teaching). Overall, participants' training with the prototype was successful in the way that all of them could read at least some short words at the end of the third session. Results of the evaluation study will inform future prototype developments and improvements. These can be summarized in two major points.

First, participants enjoyed the interaction with the virtual agent and had, likewise, fun with the gesture-based training program. The majority of participants were motivated to continue the gesture-based training apart from the study. So the idea of bringing gesture-supported literacy training together a virtual human in the role of a teacher, as realized in the prototype, receives distinct support in terms of sensed fun and motivation. These issues are of particular importance for the development of a comprehensive literacy training because most of the intervention methods (courses, e-learning) which currently available are affected by decreasing motivation of participants and high dropout rates.

Second, problems arose especially from particular gestures. In the gross-motoric gesture set, only one gesture appeared to be problematic. In the fine-motoric set, there were several gestures which caused trouble. People who trained with the latter reported repeatedly that gestures were relatively hard to recognize and to imitate. We take these observations and comments as an indication to bank on a rather gross-motoric gesture set in future development. There are several likely explanations for this finding (cf. Kraft (1971)). One reason might be that people with cognitive impairments are often also motorically disabled so that gross-motoric movement are less challenging for them than fine-motoric movements. Another reason could be that gross-motoric systems provide motoric distinctions between vocals and consonants which is not present in the fine-motoric systems. This distinction might be an effective help for reading and sound synthesis. Further, in the gross-motoric gesture set, the shape of movements is related to letter shape which might also support memory performance. Moreover, evidence from human-robot interaction also provides evidence for large and exaggerated gestures to enhance memory performance, engagement and perceived entertainment value (Gielniak and Thomaz, 2012).

On the basis of these insights we will further improve and extend the prototype system and transform the present wizard-of-oz into an autonomous system. Besides work on significant challenges in speech and gesture recognition this work should also comprise work on the agent's overall verbal and nonverbal behavior. An advanced prototype should also be subject to further controlled evaluations measuring both short- and long-term learning success. Moreover, as we could substantiate in the work presented here, that employing gesture supported literacy training with a virtual character is a promising direction, we can think of applying a similar approach for other content or domains – as one participant put it: “Maybe we could also train sign language with Billie. In gestures he is already great”.

Acknowledgements

The author is grateful to Katharina Suhre. This research is supported by the Cluster of Excellence in “Cognitive Interaction Technology” (CITEC), both funded by the German Research Foundation (DFG).

References

C. Abras, D. Maloney-Krichmar, and J. Preece. 2012. User-centered design. In *Encyclopedia of Human-Computer Interaction*. Sage Publications.

A.L. Baylor and S. Kim. 2009. Designing nonverbal communication for pedagogical agents: When less is more. *Computers in Human Behavior*, 25:450–457.

K. Bergmann and M. Macedonia. 2013. A virtual agent as vocabulary trainer: Iconic gestures help to improve learners' memory performance. In *Proceedings of the 13th International Conference on Intelligent Virtual Agents*, pages 139–148, Berlin/Heidelberg. Springer.

U. Bleidick and W. Kraft. 1966. *Lesen und Lesenlernen unter erschwerten Bedingungen [Reading and learning to read under difficult conditions].* Neue Deutsche Schule Verlagsgesellschaft.

M.J. Gielniak and A.L. Thomaz. 2012. Enhancing interaction through exaggerated motion synthesis. In *Proceedings of the 7th annual ACM/IEEE international conference on Human-Robot Interaction*, pages 375–382.

S. Heidig and G. Clarebout. 2011. Do pedagogical agents make a difference to student motivation and learning? a review of empirical research. *Educational Research Review*, 6(1):27–54.

F.J. Koch. 1939. *Fingerlesen. Lesen als Gebärdenspiel [Finger-reading. Reading as a gesture-based game].* Schwann, Düsseldorf.

H.J Kossow. 1979. *Zur Therapie der Lese-Rechtschreibschwäche [About the therapy of dyslexia].*

W. Kraft. 1971. Lautgebärden im Erstleseunterricht der Lernbehinderten- und Geistigbehindertenschule [gestures for sounds in reading eductation in specilaized schools for pupils with learning difficulties and cognitive impairments]. *Zeitschrift für Heilpädagogik*, 22:1–19.

M. Kramer, R. Yaghoubzadeh, S. Kopp, and K. Pitsch. 2013. A conversational virtual human as autonomous assistant for elderly and cognitively impaired users? social acceptability and design considerations. In *Lecture Notes in Informatics (LNI). Series of the Gesellschaft für Informatik (GI)*, volume P-220, pages 1105–1119, Bonn. Köllen Druck + Verlag GmbH.

W. Radigk. 1975. *Lesenlernen: Unter besonderer Berücksichtigung der Arbeit mit lernbehinderten Schülern [Learning to read: Under particular consideration of the work with learning-disabled pupils].* Marhold, Berlin.

H. Schäfer. 2011. Zur wirkungsweise von lautgebärden im förderschwerpunkt geistige entwicklung [about the effectiveness of gestures in the area of mental development]. *Behindertenpädagogik*, 2:200–121.

M. Schröder and J. Trouvain. 2003. The german text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology*, 6:365–377.

H. van Welbergen, R. Yaghoubzadeh, and S. Kopp. 2014. AsapRealizer 2.0: The next steps in fluent behavior realization for ECAs. In *Proceedings of the 14th International Conference on Intelligent Virtual Agents*, LNCS, Berlin, Germany. Springer.

H. Vilhjalmsson, N. Cantelmo, J. Cassell, N. Chafai, M. Kipp, S. Kopp, M. Mancini, S. Marsella, A.N. Marshall, C. Pelachaud, Z. Ruttkay, K.R. Thorisson, H. van Welbergen, and R.J. van der Werf. 2007. The Behavior Markup Language: Recent developments and challenges. In C. Pelachaud, J.-C. Martin, E. Andre, G. Chollet, K. Karpouzis, and D. Pelé, editors, *Proceedings of the 7th International Conference on Intelligent Virtual Agents*, pages 99–111, Berlin/Heidelberg. Springer.

R. Yaghoubzadeh, M. Kramer, K. Pitsch, and S. Kopp. 2013. Virtual agents as daily assistants for elderly or mentally handicapped persons - studies on acceptance and interaction feasibility. In *Proceedings of the 13th International Conference on Intelligent Virtual Agents*, pages 79–91, Berlin/Heidelberg. Springer.