Investigating Engagement
Intercultural and technological aspects of the collection, analysis, and use of Estonian Conversational Video Data

Kristiina Jokinen[1], Silvi Tenjes[2]
University of Tartu
[1]Department of Computer Science, J.Liivi 2, Tartu
[2]Department of Estonian as a Foreign Language, Jakobi 2-432, Tartu
E-mail: kristiina.jokinen@helsinki.fi, silvi.tenjes@ut.ee

Abstract

In this paper we describe the goals of the Estonian corpus collection and analysis activities, and introduce the recent collection of Estonian First Encounters data. The MINT project aims at deepening our understanding of the conversational properties and practices in human interactions. We especially investigate conversational engagement and cooperation, and discuss some observations on the participants' views concerning the interaction they have been engaged.

Keywords: multimodal corpora, conversational video data, interaction engagement

1. Introduction

In this paper we describe the Estonian corpus collection and analysis activities and especially focus on the project MINT (Multimodal INTeraction), and its collection of the Estonian First Encounters Dialogues. The aim of the MINT project is to create multimodal database which will enable researchers to study interaction behaviours concerning gesturing, synchrony and engagement in particular, and also allow systematic comparison between Nordic and Baltic multimodal conversational strategies. In this way, the project is connected to the work in the Nordic context, where the MUMIN network (Allwood et al. 2007) and the ongoing NOMCO project (Paggio et al. 2010) have already paved the way to collect and annotate comparable data that can be used to investigate communicative phenomena, especially feedback, turn management and sequencing. The data will make it possible to empirically study multimodal signals (head movements, facial displays, hand gestures and body postures) and their relation to spoken utterances, and moreover, to compare and analyse similar phenomena and communicative strategies in different but neighbouring languages.

In this paper we describe our corpus collection activities and especially introduce the recent collection of the Estonian First Encounters data. We also discuss some issues related to conversational engagement and cooperation, and provide observations on the participants' views concerning the interaction they have been engaged.

2. The Corpus Collection Projects

At the University of Tartu we have two research groups working on multimodal communication and the analysis of conversational video data. The groups have different focus points but enjoy synergy built on mutual cooperation.

Within the project MINT (*Multimodal Interaction – intercultural and technological aspects of video data collection, analysis, and use)* we have collected a corpus of Estonian First Encounter dialogues. The project is funded by the Estonian National Science Foundation (ETF), and it focuses on the interlocutors' multimodal means and strategies for building shared understanding, on the basis of their participation in a particular activity. Important aspects deal with intercultural comparisons concerning the participants' engagement and synchrony in various communicative activities, and building models for their automatic processing. As human-machine interactions get more complex, such models are crucial in designing and constructing different applications, e.g. interactive robot agents (Wilcock and Jokinen, 2011).

The goals of the MINT project are:

a) to create Estonian multimodal video corpus on various conversational activities,

b) to provide analysis and annotation of the data that contributes to the previous work on annotation standards, guidelines, and schemes,

c) to study multimodal signals, especially gesturing, in social communication and conversation management, and indicating the interlocutors' engagement and synchrony in communicative activity,

d) to build (computational) models for the coordination and controlling of interaction (e.g. taking turns, giving feedback), and constructing shared understanding, and

f) to investigate techniques and means for automatic recognition of multimodal signals, especially gestures.

This MINT project is supported by the multimodal communication research group (MUSU), and another ETF project *The structure of multimodal communication and the choice of communication strategies*. This project aims to identify the communicative behaviours important from the perspective of a particular communicative situation in social interaction, and to analyse the choice and use of means of communication within this context. The MUSU group is currently compiling a multimodal communication database, the Multimodal Communication Research Corpus. The corpus has two sub-corpora: corpus of interactive communicative

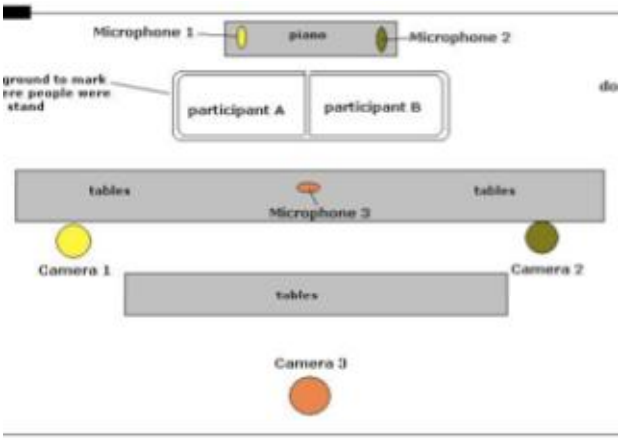

Figure 1 The data collection setup

Figure 2 A mosaic view of the video recordings.

situations (ISU) and corpus of contextualized written texts (KOK). The ISU sub-corpus contains data in the form of video and audio recording from real-life Estonian communicative settings, e.g. multiparty casual conversations in the borderland of East-Estonia, situations in language learning classroom, and specific material such as communication situations with Patau syndrome subject, etc. The KOK sub-corpus contains data in the form of written texts, e.g. literal translations, historical manuscripts, etc.

3. First Encounters Data Collection

In the first phase, the MINT project has collected a corpus of first encounters following the guidelines of the project NOMCO. The first encounter dialogues engage participants, who do not know each other in advance, in an activity where their task is to chat and make acquaintance with each other. Original data was collected in the Estonian language, and the data is annotated and analysed using an annotation scheme which is co-measurable with the annotations used in NOMCO.

Each participant was given a short presentation of the project and the goals of the data collection before the recording, and they were also asked to sign a consent form (in the Estonian language) that grants permission for their video data to be used for research purposes, and to be shown to third parties without further permission.

The collection setup is shown in Figure 1. The participants entered the recording environment through the doors at both ends of the video setup room, and if they arrived too early and needed to wait for their time, it was made sure that the pairs did not see each but in the experiment room. They were asked to proceed to the line marked on the floor. This was to ensure that both participants were approximately in the middle of the video camera views.

Three cameras were used: one recording each of the two partners (marked by yellow and green balls in Figure 1, and one recording both (marked red in Figure 1). We used SonyHDR-XR550V cameras with three external Sony ECM-HW2 wireless microphones. The microphones were paired with cameras so that each camera had its own audio track. As for the video recording, we chose the full HD quality mode, although it turned out that the standard quality would have been good enough.

The camera views were cut, edit and merged via Sony Vegas Pro 11, and they were syncronised and integrated into one single video film providing a mosaic view of the situation (as in the similar Finnish encounters). This is shown in Figure 2.

We have a total of 23 participants (12 male and 11 female), with age ranging between 21 and 61 years. The participants are native speakers of Estonian and they are students or university employees. Each participant took part in two encounters, i.e. with two different partners. The corpus contains 23 encounters, and each encounter is about 8 minutes long. They balanced with gender distribution and we have 8 female-female encounters, 7 female-male encounters, and 8 male-male encounters.

4. Participants' Views of the Interaction

We also conducted a small questionnaire regarding the participants' impressions and feelings of the interaction. The questionnaire was in a web format and the participants answered the questions immediately after each interaction. The questionnaire was in Estonian, and asked if the participants considered the interaction enjoyable, friendly, impressive, nice, interesting, relaxed, anxious, natural, happy, tense, awkward, angry in a 5-point Likert scale, where 5 indicates agreement and 1 disagreement with the adjective in question.

The average ratings among the 23 participants are given in Table 1 (next page). As can be seen, the participants have had rather positive experience of the interactions, with the top impression (4.2/5) being happy. Also the adjectives enjoyable, nice, and interesting are found appropriate in describing interaction experience, while the participants did not regard their interactions as being angry at all, and, somewhat surprisingly, they didn't seem to consider the interactions very awkward, tense, or anxious either.

Since we also have demographic information about the participants' gender, age group, education, self-estimated knowledge of the computers and self-estimated familiarity with videos, we did some detailed studies to check if there are differences between the participants' experience along

these factors. Using Student's t-test, we first compared the evaluation values with respect to gender.

descriptive feature	average	min	max
enjoyable	**4.1**	1	5
friendly	3.0	1	5
impressive	3.7	1	5
nice	**4.1**	**2**	**5**
interesting	**4.1**	**2**	**5**
relaxed	3.6	1	5
anxious	2.3	1	5
natural	3.4	1	5
happy	**4.2**	**2**	**5**
tense	2.0	1	4
awkward	**1.9**	1	5
angry	**1.0**	1	2
Average	3.1		

Table 1 Participants' average impressions of the interaction. The values are given in a 5-point Likert scale with 5 indicating agreement and 1 disagreement.

It turned out that differences between the average male and female participants are not statistically significant except for the value "interesting": the male participants consider their interactions more interesting than female participants (means: 4.4 vs. 3.6) at $p < 0.01$ ($t= 2.68$, degrees of freedom 44, standard deviation 0.775). In general, male participants also considered the dialogues more friendly and impressive, but also more anxious and tense than females, while female participants considered the dialogues slightly more enjoyable and natural.

Concerning age groups, 3/4 of the participants are between 20-30 years of age, so there was not enough data to find significant differences between the age groups. However, on average, the 20-30 years old considered their interactions slightly nicer and more relaxed than did those over 30 years of age, while the latter found their interactions more enjoyable, friendly and interesting.

As for differences in regard to education levels (graduate students vs. those who study on the Master's level having completed their bachelor degree vs. those who had completed Master's degree), they were not significant either. However, some divergence was obvious on individual aspects and e.g. graduate students regarded their interactions less relaxed and more anxious, yet more interesting than those who had bachelor degree. Those with master's degree were in general more positive than the others, but as said the differences were not significant.

Neither were deviations with respect to computer knowledge significant. Participants who estimated themselves experts seem to find interactions slightly more enjoyable, friendly, interesting, natural, and happy than those who estimated themselves as advanced users, while the latter tend to rate the interactions more awkward and anxious than the experts. It is interesting that none of the participants self-estimated themselves as "ordinary" users, i.e. IT knowledge was generally regarded as common rather than a special skill.

However, an intriguing result is that the participants' self-estimated familiarity with video recordings, video analysis and videos in general (beginner, middle, advanced) caused statistically significant differences in their experience. In particular, those who regarded themselves as having advanced familiarity with the videos found interactions more interesting (mean: 4.5) than those who regarded themselves as beginners (mean: 3.7) at $p > 0.05$, $t=2.37$, degrees of freedom 18, standard deviation 0.770. Those with advanced familiarity also contrasted with participant who considered themselves in the middle, in that the interactions were happy (mean: 4.4 vs. 3.6) at $p < 0.05$

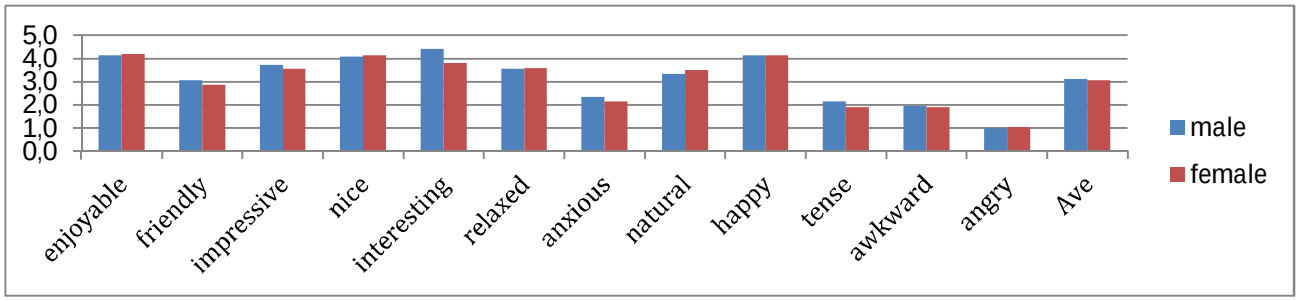

Figure 3 No significant difference between gender impressions except for interesting ($p < 0.01$)

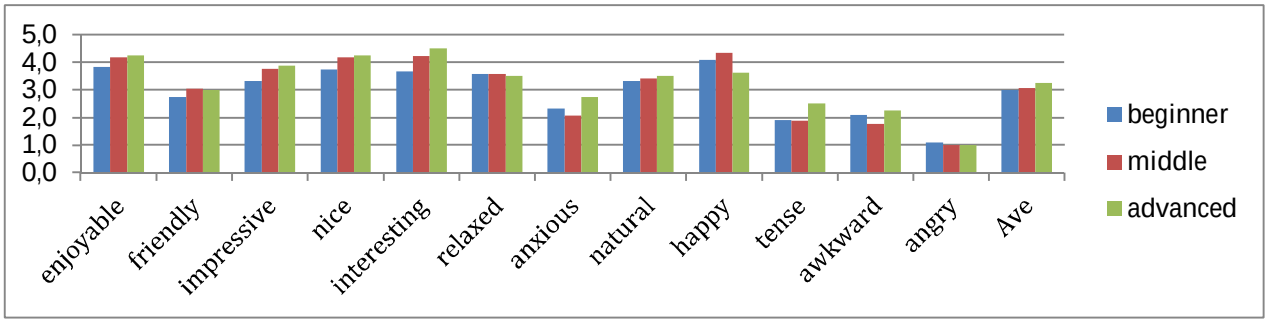

Figure 4 Significant differences between advanced and beginners on interesting, and between advanced and middle on happy ($p < 0.05$).

(t=2.27, degrees of freedom 32, standard deviation 0.786). Big deviations were also found in that the advanced participants found interactions more awkward, anxious, and tense than the middle-level participants, and they considered interaction more tense, yet nicer than the beginners.

As said, not all differences are significant. However, what is significant is the contrast between participants who evaluate themselves as experts vs. not experts on videos and video recordings. Apparently the participants' interest and knowledge of the novel technology also carries over to situations where the novel technology is used, and thus also affects their experience of the communication in that particular situation. It is good to remember that the participants' knowledge of computers and information technology as such did not cause a similar effect in the participant's experience. This supports the hypothesis that new expertise or knowledge which deviates from the usual knowledge in conversational situations tends to have a positive effect on the evaluation and experience of the situation as a whole.

5. Engagement

Conversations are cooperative activities through which the participants aim at achieving some underlying goals of the interaction. The goals can range from specific task-related goals (e.g. to instruct someone to use annotation software, to provide information about bus time, to learn to know each other) to an intention just to keep the "channel open". The interlocutors react to each others' actions and coordinate their turns in a manner that allows both to present their message in a cooperative manner. Engagement is an important sign of this kind of cooperation, and it is related to the interlocutors' experience of the interaction in general: more engaged the interlocutors are in the conversation, the more positively they may experience the interaction.

Our interest in studying conversational engagement goes back to intelligent systems and interaction technology, where engagement is used to describe the user's willingness and involvement in the interaction with the automatic interactive system. If it is possible to measure the interlocutors' engagement level, it is easier to adjust the system's conversational strategies accordingly.

An intuitive definition of engagement is that it refers to the situation where the participants are involved in a conversation and show basic willingness to listen to the partner and provide coherent contributions. However, the more active the interlocutors become, the more engaged they seem to be in the conversation: their speaking frequency, tone of voice and body posture indicate interest and commitment to the topic of the conversation: they engagement becomes embodied. Usually such activity is reinforced by the partner's actions on a similar level of engagement, so we can talk about mutual engagement. Such interactions are described as pleasant, inspiring, and fun.

We will investigate various action patterns and behaviours that are typical for the participants when they are engaged in interaction they describe as nice, enjoyable, interesting, natural, pleasant. We will explore some measures of engagement in terms of the interlocutors' verbal and non-verbal communicative activity, in particular their gesture activity and study if engagement can be operationalised through these signals. Previous work has used such measures as frequency counts and the utterance density (Campbell and Scherer, 2010; Jokinen, 2011), and we will follow these lines as is appropriate.

We also emphasise the close relationship between speech and multimodal information in the processing of human conversational interactions. The semantic content of linguistic utterances is accompanied by hand gestures, body movement, eye-gaze, and non-speech vocalisations which are used as tacit signals to indicate the speaker's focus of attention, intentions, emphasis, emotional state, etc. They should, of course, be processed alongside the language expressions.

Moreover, the participants' synchrony with each other is regarded as one of the pertinent signs of cooperation: the interlocutors intuitively tend to follow the partner's communication and produce similar behaviour, thus contributing to the construction of the shared context and mutual understanding. This kind of adaptation to each other's behaviour is often called alignment (Pickering and Garrod, 2004), or mimicry, and it can take place verbally (words, prosody) or non-verbally (gestures, body posture). It is thus an important sign of the partners' engagement in the interaction, and our previous works on this can be found in Jokinen and Pärkson (2011) as well as in Rummo and Tenjes (2011).

We also have data and first results about specific type of communication – subject with the Patau syndrome. The preliminary results of the analysis of data (see Jokinen et al. forthcoming) reveal meaningful nonverbal behaviour through touching. For instance, the Patau subject put her hand on her partner's shoulder, creating her own communicative space. From this act we can conclude that *touch* is clearly one component of Patau subject's language. This type of nonverbal interactive behaviour guarantees to subject that her partner is involved in the interaction as well as supplies her necessity of adjacency.

When people engage in dialogue, they use verbal and non-verbal cues to structure the conversation flow and provide feedback about the current understanding of the discourse. An intuitive measure of cooperation between interlocutors is their verbal and non-verbal communicative activity, and we hypothesize that engagement can be measured analogously, with respect to their cooperation. It can be estimated by measuring the participants' verbal and non-verbal activity. In this we use both quantitative and qualitative analysis; the latter gives an opportunity to show variety of the research material.

In another previous study, we used annotated multiparty conversations where three participants (who are familiar with each other) had been assigned certain roles related to a simulated school inspection situation. The analysis

shows that participants use a full repertoire of different non-verbal signals as signs of their engagement in the conversation: hand gestures, facial displays, nods, and body movement. For instance, the speakers animate and emphasise their speech by hand gestures, so as to give importance to a particular part of their speech, and they also use gestures to control and coordinate conversation flow. Also the speakers' gaze is used as a pointing gesture: it can mark the speaker who is expected to take the next turn (mutual gaze). It is also related to the processing of the given information and is effectively used to create social bonds between the interlocutors. Some body movements are also used to fill pauses in conversation: if the speaker does not want to take the turn or is unable to take the turn, they usually withdraw from the centre of the conversational space. By body movement, the participants tacitly indicate that they are present in the conversation.

Other communicative signals include nodding and laughing. Nodding is related to the interlocutor's engagement because the very act of nodding signals that the person takes part in a conversation and is ready for cooperation. Laughing often occurs in smooth conversations at the same time. On other hand, laughing can also illustrates the case where engagement is unevenly distributed among the interlocutors: not all are engaged to the same level at the same time.

The fresh First Encounters corpus will provide valuable reference material for these earlier corpora and related studies. Moreover, it may help us also to shed light on the general question of whether the participants' nonverbal behaviour differs in role-playing and spontaneous situations, and if so what are the characteristics.

6. Conclusions and Future Work

Several interesting issues emerged from the corpus collection and preliminary data analyses, which can lead to subsequent research. For instance, we have studied the interlocutors' self-evaluation of the interaction they were engaged in, and although the dataset is relatively small, we found statistically significant differences concerning how the participants experienced the interaction. Interestingly, the participants' expertise on the novel technology that was used in the data collection seems to be one of the distinguishing factors, indicating that engagement and experience are complex issues, related to intrigue aspects of individual skills and knowledge.

We will also continue engagement studies and aim at building models that would allow us to better understand the complex process of cooperation and coordination of interactions. For this, we will especially study gesturing and gesture recognition, and can use all the that has been collected in various activities.

The research will also focus on intercultural comparison of the collected conversational data. Systematic multimodal communication studies are useful, and they can provide valuable empirical evidence for the observations and views concerning different communication styles and strategies in different cultures. Especially in the Nordic context, comparisons between the first encounter dialogues in the neighbouring countries will be particularly interesting because the countries have a long history of various interactions and encounters on political and cultural levels, while the languages are pair-wise linguistically related. It is also possible to continue this kind of comparison by extending our research to corpora that represent larger cultural differences, such as Japanese (Jokinen et al. 2010).

We are currently in the process of transcribing and annotating the First Encounters Data. As future work, we plan to collect more data, and may especially focus on multi-party conversations where the participants have different roles: the speaker, the main addressee, and the side addressee(s). In these situations the participants' role is important and their level of engagement can differ depending on whether the participant is actively engaged in the interaction or listening to a conversation as a side participant. For instance, the side participant need not react at the same time as the main addressee, and still be engaged in the conversation.

7. Acknowledgements

We would like to thank the participants who took part in the video recordings and Sven Laater who took care of the practical video setup. We would also like to thank Mare Koit and her project in the Estonian Centre of Excellence in Computer Science (EXCS) for encouragement and financial support for the work. The MINT project is financially supported by the Estonian Science Foundation grant ETF8958.

8. 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*.

Argyle, M. (1975) Bodily Communication. Methuen & Co. Ltd. London.

Boersma, P. and D. Weenink (2009). Praat: doing phonetics by computer (version 5.1.05). Retrieved May 1, 2009, from http://www.praat.org/

Campbell, N., Scherer, S. 2010. Comparing Measures of Synchrony and Alignment in Dialogue Speech Timing with respect to Turn-taking Activity. *Proceedings of Interspeech*. Makuhari, Japan

Jokinen, K. (2011). Turn taking, Utterance Density, and Gaze Patterns as Cues to Conversational Activity. *ICMI-MMI 2011*, November 14-18, 2011, Alicante, Spain.

Jokinen, K., Tenjes, S., Rummo, I. forthcoming. Embodied interaction and semiotic categorization: communicative gestures of a girl with Patau syndrome. In C. Paradis et al. (eds) *The Construal of Spatial Meaning: Windows into Conceptual Space*. Oxford: Oxford University Press, 38 pp.

Jokinen, K. and S. Pärkson, 2011. Synchrony and Copying in Conversational Interactions. The 3rd Nordic Symposium on Multimodal Interaction, Helsinki, May 2011.

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

Navarretta, C., E. Ahlsén, J. Allwood, K. Jokinen, P. Paggio (2011) Creating Comparable Multimodal Corpora for Nordic Languages. In *Proceedings of the 18th Nodalida*., 153-160.

Nezlek, J. B. (2010). Multilevel modeling and cross-cultural research. In D. Matsumoto and A. J. R. van de Vijver (Eds.) *Cross-Cultural research methods in psychology*. Oxford.

Paggio, P., J. Allwood, E. Ahlsén, K. Jokinen, C. Navarretta (2010). The NOMCO multimodal Nordic resource - goals and characteristics. In *Proceedings of LREC 2010*, 2968-2973.

Pickering, M, Garrod, S. 2004. Towards a mechanistic psychology of dialogue, Behavioral and Brain Sciences 27, 169– 226.

Rehm, M., E. André, N. Bee, B. Endrass, M. Wissner, Y. Nakamo, A. Akhter Lipi, T. Nishida and H.H. Huang (2009). The Intercultural Dimension of Multimodal Corpora. In Kipp et al. (eds) *Multimodal Corpora*. LNAI 5509. Springer,138–159.

Rummo, I. and S.Tenjes (2011). AJA mõistestamine Patau sündroomiga subjekti suhtluses. (Conceptualization of TIME in the context of Patau syndrome). In H. Metslang et al. (eds) *Eesti Rakenduslingvistika Ühingu aastaraamat 7 / Estonian Papers in Applied Linguistics 7*. Tallinn: Eesti Keele Sihtasutus, pp 231-247.

Wilcock, G. and K. Jokinen (2011). Adding Speech to a Robotics Simulator. *Proceedings of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop (IWSDS 2011)*, Granada, Spain, pp 371-376.