The NOMCO multimodal Nordic resource - goals and characteristics

Patrizia Paggio*, Jens Allwood†, Elisabeth Ahlsén†, Kristiina Jokinen‡, Costanza Navarretta*

*University of Copenhagen, Centre for Language Technology
†University of Gothenburg
‡University of Helsinki
E-mail: paggio@hum.ku.dk, jens@ling.gu.se, elisabeth.ahlsen@ling.gu.se, kristiina.jokinen@helsinki.fi, costanza@hum.ku.dk

Abstract

This paper presents the multimodal corpora that are being collected and annotated in the Nordic NOMCO project. The corpora will be used to study communicative phenomena such as feedback, turn management and sequencing. They already include video material for Swedish, Danish, Finnish and Estonian, and several social activities are represented. The data are being annotated following an annotation scheme that provides attributes concerning the shape and the communicative functions of head movements, face expressions, body posture and hand gestures. The paper also discusses how the corpora will be used to study the way feedback is expressed in speech and gestures, and reports results from two pilot studies that have been conducted on this issue. The annotated corpora will be valuable sources for research on intercultural communication as well as for interaction in the individual languages.

1. Introduction

Through the collaborative Nordic project NOMCO we are building corpora of annotated multimodal videos for Swedish, Danish, Finnish and Estonian. The purpose is to provide comparative annotated data on which to base investigations of communicative phenomena, especially feedback, turn management and sequencing. The data will make it possible to verify empirically how gestures (head movements, facial displays, hand gestures and body postures) and speech interact in all the three mentioned aspects of communication. The project aims to (i) create multimodal corpora for the languages involved with a number of standardised coding features, (ii) perform a number of studies testing hypotheses on multimodal interaction, (iii) develop, extend and adapt models of multimodal communication management that can provide the basis for interactive systems, and (iv) apply machine learning techniques to create support for automatic recognition of gestures with different communication functions.

NOMCO positions itself in the research area carved by many other projects and networks both inside and outside of Europe, e.g. ISLE, HUMAINE, SIMILAR, CHIL, AMI, CALO, VACE, CALLAS. In the Nordic countries, the network on Multimodal Interfaces MUMIN (2002-2004) stimulated cooperation among research groups that were and still are working with multimodal resources and systems. NOMCO is the first effort to build parallel Nordic multimodal corpora.

In this paper, we start in Section 2 by describing the NOMCO corpora collected so far. Then in Section 3 we explain how we plan to annotate the gesture behaviour. Finally, in Section 4 we discuss how the corpora will be used to study feedback phenomena, and in Section 5 we conclude.

2. The corpora

The NOMCO corpora will eventually span over annotated video material from different social activities. The difference between activities will to some extent lead to a difference in multimodal behaviour. Such differences can, for example, result from:

i. Whether the activity allows the participants to stand up or sit down.
ii. The degree of freedom allowed participants in different roles: usually roles connected with more power lead to greater freedom in both speech and gesturing.
iii. The purpose of the activity: some activities necessitate multimodal exchange, while others have the opposite effect.
iv. The type of emotions and effect connected with the activity.

In what follows, we will first describe studio-recorded corpora belonging to the two types “first encounters” and “group interaction”. They either have been or are being collected, and will be made available for research purposes through the project website http://sskkii.gu.se/nomco/.
Then we will briefly mention existing corpora which the project has access to, but which are subject to more restricted availability constraints.

2.1 First encounters

The first section of comparable material consists of videos from “first encounters” interactions. This type of activity has also been studied by other projects dealing with cross-cultural multimodal studies (Rehm et al., 2009). Different cultures have in fact different ways of dealing with social status, familiarity and other social and psychological factors that play a role in first encounters,

and that have an influence on the linguistic and gestural behaviour. Nordic cultures are generally regarded as relatively similar: our data will allow us to find empirical evidence for similarities as well as differences.

The current "first encounters" corpus involves so far 30 Swedish speakers and 12 Danish speakers. The subjects are instructed to get to know each other in about 3-5 minutes. They are recorded in a studio standing in front of a light background, in order to facilitate automatic registration of body movements. The subjects are also given a questionnaire about their reactions to the other person and the interaction as a whole after the recording. The videos are being annotated according to the NOMCO annotation specifications, as explained in Section 3 below. Thus, the choice of a common social activity, but also the use of similar setups, equipments and annotations makes the two corpora comparable.

A statistical analysis was carried out on the questionnaires filled in by the Danish participants. The subjects, 6 men and 6 women, were all native speakers of Danish, and either university students or people with a higher education. Ages ranged from 21 to 36. Each subject participated in two videos, one with a male and one with a female partner. The subjects were asked to characterise their experience in terms of 12 parameters concerning setting and interaction (Nezlek, in press). For each parameter, they could choose a value from a 5-point scale. The purpose was to assess the naturalness of the collected data, as well as their homogeneity in terms of subject experience. The results are shown in Table (1).

Variable	Mean	Sd
Enjoyable	4.42	0.72
Intimate	2.71	1
Influence	3.75	0.79
Liked	4.04	0.91
Interesting	4.17	0.76
Free	4.12	0.74
Perturbed	2.54	1.06
Natural	2.33	1.05
Happy	4.58	0.58
Tense	2.42	1.06
Awkward	2.17	0.82
Angry	1.38	0.49

Table 1: Subjects' interaction experience: mean scores and standard deviations.

A high score is positive when associated with certain parameters (e.g. enjoyable), and negative with others (e.g. angry). In general, the participants show a positive response. Even though the score for how perturbed they felt by the camera is slightly above average, at the same time they report feeling influential, well-liked, interesting and free to express themselves. The data are quite homogeneous across individual variation, gender and age, and can be considered a relatively valid exemplification of natural interaction with participants positively oriented towards the task albeit somewhat tense. The data have not yet been analysed with regard to linguistic and gestural behaviour variation. Awareness of individual variation, however, is crucial in cross-cultural studies, where individual and culturally determined behaviours must be teased apart from one another.

2.2 Group interaction

The term *group interaction* here refers to three or more people speaking to each other, either in formal meetings or in more informal settings.

Group interactions have been recorded in Sweden and Estonia. The Estonian data were recorded in a studio with three participants. The corpus contains two half-an-hour long scenario-based meetings where the participants sit around a table sideways so that they are half facing one another. The task of the participants in the first dialogue was to discuss the design of a new school building; the task in the second dialogue was to discuss the inspection of the new school building. Participants were students who assumed the roles of an architect, a building designer, and a council representative. The dialogues proceed naturally, and although the interaction is based on controlled scenarios, the participants' gestural behaviour is natural. Two video clips of about two minutes from both conversations have been annotated following the NOMCO annotation specifications, and work is ongoing to annotate the whole corpus. The corpus belongs to the activity type of Task-Oriented Dialogues with Role Play, and can be contrasted with two-party dialogues where both partners have roles with similar responsibility for the smooth continuation and turn-taking in the interaction: in the Estonian group interactions, the responsibility for the dialogue is dependent on the different roles the participants have, and their gestural behaviour may reflect this difference.

We have also collected, in collaboration with the Doshisha University in Japan, a corpus of group interactions where the three participants are chatting freely about issues that interest them. The corpus has 10 conversations with familiar interlocutors and 10 with unfamiliar interlocutors (see more in Jokinen et al., 2010). The unfamiliar conversations are comparable to the first encounters topic-wise, but differ in that there are three participants instead of two. The corpus also has eye-tracking information so that the gaze of one of the partners can be accurately studied. Part of the corpus is annotated according to the NOMCO specifications. In addition, there is annotation of dialogue acts following the guidelines developed in the AMI project (see www.amiproject.org). The Japanese corpus will provide us with an opportunity to compare communicative functions beyond the Nordic sphere in the context of Eastern and Western cultures, along the lines outlined in Jokinen and Allwood (2010).

2.3 Field recordings

The existing Gothenburg Spoken Language Corpus (GSLC), consisting of 1.4 million words from 25 different types of video and audio recorded social activities, will also be used in the project. Parts of it will be annotated in the same way as the new material we are collecting. The GSLC mainly contains recordings with high ecologic validity, i.e. field recordings. The following social activities are represented: Discussion, Retelling of Article, Interview, Task-Oriented Dialogue, Informal Conversation,

Role Play, Trade Fair, Arranged Discussions, Formal Meeting, Consultation, Shop, Dinner, Market, Auction, Factory Conversation, Party, Games & Play, Phone, Travel Agency, Court, Church, Lecture, Hotel, Therapy, Bus Driver-Passenger interaction. For comparative purposes, we will be especially interested in formal meetings and informal conversation. An overview of the entire corpus is at http://www.ling.gu.se/projekt/tal/index.cgi?PAGE=3.

The GSLC Corpus has been partly analysed for linguistic variation only in spoken language, not much for multimodal variation. This means that such an analysis would be a valuable addition to the corpus planned in NOMCO.

Another example of field recordings is the Finnish data on card-playing situations. These are group interactions among four participants, and they belong to the activity type of Games & Play. We aim at collecting more dialogue data that would be available freely for the purposes of comparison planned for the NOMCO project. Recent recordings in this direction comprise material about interactions between a Finnish teacher and an immigrant student. These dialogues are in terms of activity type a kind of Consultation, where the teacher gives feedback and discusses the results of the language test the student has taken.

An overview of the properties of the various corpora is provided in Table (2).

Corpus Id	Language	No. of subjects per video	No. of interactions	Average video duration	Studio/Field recording	Activity type	Setup	Equipment
DA first encounters	Danish	2	12	5 min.	studio	getting to know each other	standing in front of a light background	three TV cameras two overhead microphones
SE first encounters	Swedish	2	30	6 min	studio	getting to know each other	standing in front of a light background	three TV cameras
SE formal meetings	Swedish	average 6.1	15	120 min	field	Formal meetings	sitting around a table	one or two cameras, sometimes separate audio
SE informal conversations	Swedish	average 2.4	23	26 min	mixed	Informal conversations	mixed	one or two cameras, sometimes separate audio
EE group	Estonian	3	2	30 min	studio	task-oriented, role-play	sitting around a table	one camera
FI teacher student	Finnish	2	18	6 min	field	consultation, semi-formal interview	sitting at a square table	one camera

Table 2: Details of the NOMCO corpora.

3. Annotation

Gestures in the NOMCO data are annotated according to the MUMIN annotation scheme (Allwood et al., 2007), where each modality is described by means of a list of attributes. The scheme is a general framework for the study of gestures in interpersonal communication that has been applied to multimodal data in several languages within the context of the Nordic MUMIN network (www.cst.dk/mumin). It concerns face expressions and head movements – both subsumed here under the term *head gestures*, hand gestures and body posture, and it provides attributes for shape as well as function. The original MUMIN scheme has been adapted for the purposes of NOMCO. The most notable change is the fact that gaze has not been included since its annotation had proved too unreliable in previous studies. Gaze attributes can be inferred from the face and head attributes, or obtained through gaze tracking.

Gesture annotation is performed with the ANVIL tool (Kipp, 2001).

Head gestures	
Face	Smile, Laughter, Scowl, Other
FaceInterlocutor	ToInterlocutor, AwayFromInterlocutor
Eyebrows	Frown, Raise, Other
HeadMovement	Nod, Jerk, Backward, Forward, Tilt, SideTurn, Shake, Waggle, Other
HeadRepetition	Single, Repeated
Body posture	
BodyDirection	Forward, Backward, Up, Down, Side, Other
BodyInterlocutor	ToInterlocutor, AwayFromInterlocutor
Hand gestures	
Handedness	SingleHand, BothHands
Trajectory	Forward, Backward, Up, Down, Sideways, Complex, Other

Table 3: Attributes for the annotation of gesture shape.

The attributes for the annotation of gesture shape are shown in Table (3). The granularity of the annotation categories has been determined on two grounds. First of all, the main purpose of the annotation is to be able to distinguish different communicative functions rather than providing precise morphological descriptions. Furthermore, it must be possible for the annotators to complete the annotation task in a reasonable timeframe. The communicative phenomena we are mostly interested in can be captured on the basis of the MUMIN attributes. To study other phenomena where iconicity and hand gestures are more central, more detail would probably be necessary in the hand gesture annotation. For example, attributes could be added to describe the position and orientation of the palm, the fingers and the forearm, and the amplitude of the movement (McNeill 1992). Separate descriptions for right and left hands could also be considered following McNeill.

The functional annotation features concern feedback, turn management and sequencing, and only gesturing that is relevant to one of these phenomena is annotated. These attributes are shown in Table (4).

Feedback	
FeedbackBasic	CPU, Other
FeedbackDirection	Give, Elicit, GiveElicit, Underspecified
FeedbackAgreement	Agree, NonAgree
Turn management	
Turn	Take, Accept, Yield, Elicit, Complete, Hold
Sequencing	
Sequencing	Open, Resume, Continue, Close

Table 4: Attributes for the annotation of gesture function.

Semiotic categories following the distinctions by Peirce (1931), are also provided in the annotation scheme. The categories are *IndexicalDeictic*, for gestures pointing to objects in the conversation situation; *IndexicalNonDeictic*, for gestures directly connected in some way causally, (although not through pointing) to their meaning; *Iconic*, for gestures building on similarity and *Symbolic*, for gestures building on arbitrary conventional relations. For each gesture, a relation with the corresponding speech expression, if one such exists, is also annotated by means of a link. The link can point to a speech segment uttered by the person producing the gesture, or to a speech segment in the interlocutor's vocal stream.

Inter-coder agreement scores were calculated in an earlier study based on the annotation of Danish and Finnish TV multimodal data (Allwood et al., 2007). The *k* scores obtained on the attributes concerning facial display attributes indicate substantial agreement (0.83-0.96). Those for hand gesture attributes show moderate agreement (0.55-0.88). The inter-coder agreement for the annotation of communicative functions also varies from moderate for sequencing to substantial for feedback. However, it must be noted that the material used in these tests was rather limited. Fresh scores will be calculated for the NOMCO corpora.

The speech stream is orthographically transcribed, and we plan to add an annotation of dialogue acts. A possibility we are studying is the emerging ISO 24617-2 standard for dialogue acts annotation.

4. A pilot project – feedback in speech and gestures

The "first encounters" part of the corpus will be used to carry out investigations on the combined use of speech and gestures to signal feedback – investigations that will provide the basis for comparing the communicative functions across the neighbouring cultures. In these studies we focus on how gestures and speech interact, although gestures can express feedback even on their own.

Feedback is about expressing to each other in conversation whether we are willing and able to perceive, understand and accept what the interlocutor is communicating (Allwood, 2002). It is expressed in speech, but often also by gestures, especially head movements and facial expressions. Kendon (2004), Jokinen and Vanhasalo (2009) and Jokinen (2010) emphasise that feedback gestures have functions that are related to controlling the dialogue flow and progress.

It has been observed that 70% of all head movements in a subset of the Swedish GSLC corpus are related to feedback, and that most of these are nods and up-down movements (Cerrato, 2007). In general, feedback gestures are less consciously controlled than spoken words and phrases. Several studies have been done on the use of gestures in human-machine interaction. Some suggest that users like gestural feedback by a talking head (Edlund and Nordstrand, 2002). Others note that gestures can be a distraction if not seamlessly integrated with speech output (Piwek et al., 2005). Interesting attempts have also been made to develop automatic recognition of human feedback gestures to ECAs (Morency et al., 2006). Generally, there is a need for more empirical studies of relevant multimodal data to use as a basis for more

complex and realistic models.

Two issues will be investigated in our project. The first concerns the relation between features characterising gesture shape and dynamics on the one hand, and different feedback functions on the other. Not much empirical evidence has been given on the issue. However, Cerrato (2007) notes that in Swedish data, single and repeated nods have different functions, i.e. a basic continuation function (which we call *Continuation/contact, perception and understanding, or CPU)* as well as agreement. To model functional variation in the feedback dimension, we already saw that the MUMIN coding scheme distinguishes the three different features *FeedbackBasic*, *FeedbackDirection* and *FeedbackAgreement*. We have conducted preliminary machine learning experiments on limited Danish and Estonian data (Jokinen et al., 2008). The study shows that head features are quite important to discriminate between feedback types in both datasets, the strongest association being the one between nods and *FeedbackAgreement* values. In NOMCO, all head movements will be coded with feedback values, as well as with shape features. We are interested in what correspondences holding in the larger material will allow us to generalise over individual variation and in whether there are significant differences related to the different languages.

The other issue to be looked at is whether there are systematic co-occurrence patterns between feedback gestures and different prosodic realisations of feedback words and phrases. We know that the content of feedback expressions is highly context-dependent. Allwood et al. (1992) i.a. note that the interpretation of these expressions must, for example, take into account the preceding dialogue act and the polarity of the preceding utterance. It is reasonable to assume that prosody and gestures also contribute to the interpretation.

To study this, we annotated facial expressions and head movements in part of the videorecorded Danish map-task dialogues from the DanPass corpus (Grønnum, 2005). The entire corpus contains about 4,100 token *yes* and *no* phrases of one to four words enriched with phrase prosody and stress information, and it is therefore an interesting dataset to study the relation between feedback expressions and gestures. The gestures in the videos have been annotated according to the MUMIN scheme, and the feedback phrases with relevant dialogue act categories (such as *Answer, Accept* and *RepeatRephrase*). The annotated data consists of approximately one hour of video showing interactions between four different speaker pairs. The total number of head gestures annotated is 236. Of these, however, only 56 (21%) co-occur with feedback expressions. The results (for more detail, see Paggio and Navarretta, 2010) indicate that the dominating patterns of feedback phrase, stress and pitch information correlate with different types of feedback. In particular, nods and jerks are associated with *Answer* rather than *Accept*, and with *RepeatRephrase* more strongly than either *Answer* or *Accept.*

The results provided by this initial experiment are only indicative due to the limited size of the material, and also given the fact that the two subjects in the DanPass videos do not see each other. We intend to replicate the investigation on the "first encounters" section of the NOMCO corpus, where we expect to find many more and more varied movements.

We have also carried out a pilot study on repeated head movements (head-nods and head-shakes) and the speech co-occurring with them in three of the Swedish spontaneous "first encounters" interactions (Boholm and Allwood, 2010). There are a total number of 89 repeated head movements in the three recordings: 75 repeated head nods, 13 repeated head-shakes and one repeated tilt. The main function of such repeated head movements is found to be communicative feedback. This is also the most frequent function of the speech co-occurring with the head movements.

However, there is no 1-1 relation between repetition in head movement and vocal words, even if a majority of the repeated head movements (68, i.e. 76%) are produced simultaneously with speech. Repeated head movements are more often accompanied by single than repeated words.

Both repeated head movements and repeated vocal words can also occur without accompaniment in the other modality. In such cases, the most frequent function for the head movements is still communicative feedback. However, the most frequent function of repeated words without accompaniment in the other modality is own communication management.

5. Conclusions

The NOMCO project is based on existing resources and research results, but it is also collecting and annotating new multimodal corpora for several Nordic languages. It is in fact the first collaborative work directed to collect comparable Nordic multimodal corpora. The project is unique also in its efforts to design an annotation scheme that will allow comparison of data and enable quantitative measures on the communicative activity concerning feedback, turn management and sequencing. Work is in progress to collect and annotate more material, and studies on how feedback is expressed in speech and gestures are already being conducted on parts of the data. The annotated corpora will be valuable sources for research on intercultural communication as well as for interaction in the individual languages.

6. Acknowledgements

The NOMCO project is funded by the NORDCORP programme under the Nordic Research Councils for the Humanities and the Social Sciences (NOS-HS).

7. References

Allwood, J. (2002). Bodily Communication – Dimensions of Expression and Content. In B. Granström, D. House and I. Karlsson (Eds.) *Multimodality in Language and Speech Systems*. Dordrecht: Kluwer Academic Publishers, pp. 7–26.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C. and Paggio, P. (2007) The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing. In J. C. Martin et al. (eds) *Multimodal Corpora for Modelling Human Multimodal Behaviour*. Special issue of the International Journal of Language Resources and Evaluation. Springer.

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

Boholm, M. and Allwood, J. (2010). Repeated head movements, their function and relation to speech. In Kipp, M., Martin, J. C., Paggio, P. and Heylen, D. (eds.) *Proceedings of the Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*. LREC 2010. Valletta, Malta, May.

Cerrato, L. (2007) *Investigating Communicative Feedback Phenomena across Languages and Modalities.* PhD Thesis in Speech and Music Communication, Stockholm, KTH.

Edlund, J., and Nordstrand, M. (2002). Turn-taking gestures and hour-glasses in a multi-modal dialogue system. In Proc of ISCA Workshop Multi-Modal Dialogue in Mobile Environments. Kloster Irsee,Germany.

Grønnum, N. (2005) DanPASS – Danish Phonetically Annotated Spontaneous Speech. In *Proceedings of FONETIK 2005*.

Jokinen, K. (2010). Gestures and Synchronous Communication Management. In: Esposito, A., Campbell, N., Vogel, C., Hussain, A., and Nijholt, A. (Eds.) *Development of Multimodal Interfaces: Active Listening and Synchrony*. Springer Publishers.

Jokinen, K. and Allwood, J. (2010). *Hesitation in Intercultural Communication: Some observations on Interpreting Shoulder Shrugging.* Proceedings of the International Workshop on Agents in Cultural Context, The First International Conference on Culture and Computing 2010. Kyoto, Japan. pp.25-37.

Jokinen, K., Navarretta, C. and Paggio, P. (2008) Distinguishing the communicative functions of gestures. In Proceedings of the 5th Joint Workshop on Machine Learning and Multimodal Interaction, 8-10 September 2008, Utrecht, The Netherlands.

Jokinen, K., Nishida, M. and Yamamoto, S. (2009). Eye-gaze Experiments for Conversation Monitoring. The 3rd International Universal Communication Symposium, Tokyo, Japan. pp. 303-308.

Jokinen, K. and Vanhasalo, M. (2009) Stand-up Gestures — Annotation for Communication Management. In Proceedings of the NODALIDA 2009 workshop Multimodal Communication - from Human Behaviour to Computational Models. NEALT Proceedings Series, Vol. 6 (2009), 15–20.

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

Kipp, M. (2001). Anvil – A Generic Annotation Tool for Multimodal Dialogue. In *Proceedings of Eurospeech* 2001, pp. 1367 – 1370.

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

Morency, L.P., Sidner, C., Lee, C. and Darrell, T. The Role of Context in Head Gesture Recognition, *Proceedings of the Twenty-first National Conference on Artificial Intelligence* (AAAI), 2006.

Nezlek, J. B. (in press). 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. and Navarretta, C. (2010) Feedback in Head Gestures and Speech. In Kipp, M., Martin, J. C., Paggio, P. and Heylen, D. (eds.) *Proceedings of the Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*. LREC 2010. Valletta, Malta, May.

Peirce, C. S. (1931*). Collected Papers of Charles Sanders Peirce, 1931–1958*, 8 vols. Edited by C. Hartshorne, P. Weiss and A. Burks. Camrbidge, MA: Harvard University Press.

Piwek, P., Masthoff, J. and Bergenstrahle, M. (2005). Reference and gestures in dialogue generation: Three studies with embodied conversational agents. In *AISB05 Virtual Social Agents Symposium* (Hatfield, England).

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. M, J.C. Martin, P.Paggio and D. Heylen (eds) *Multimodal Corpora*. *From Models of Natural Interaction to Systms and Applications*. LNAI 5509. Springer, pp.138–159.