Lang Resources & Evaluation (2017) 51:463–494
DOI 10.1007/s10579-016-9371-6

ORIGINAL PAPER

The Danish NOMCO corpus: multimodal interaction in first acquaintance conversations

Patrizia Paggio[1,2] · Costanza Navarretta[1]

Published online: 19 October 2016

Abstract This article presents the Danish NOMCO Corpus, an annotated multimodal collection of video-recorded first acquaintance conversations between Danish speakers. The annotation includes speech transcription including word boundaries, and formal as well as functional coding of gestural behaviours, specifically head movements, facial expressions, and body posture. The corpus has served as the empirical basis for a number of studies of communication phenomena related to turn management, feedback exchange, information packaging and the expression of emotional attitudes. We describe the annotation scheme, procedure, and annotation results. We then summarise a number of studies conducted on the corpus. The corpus is available for research and teaching purposes through the authors of this article.

Keywords Multimodal corpora · First acquaintance conversations · Gestural annotation

1 Introduction

The past few decades have seen the emergence of a new research paradigm, which considers human communication as a multimodal system, so that communication is increasingly being studied by considering gesture alongside speech (see e.g. Kendon 2004; McNeill 2005; Duncan et al. 2007; Poggi 2007; Cienki and Müller 2008; Gullberg and de Bot 2010; Gibbon 2011; Enfield 2012). The general effort made by

✉ Patrizia Paggio
paggio@hum.ku.dk; patrizia.paggio@um.edu.mt

[1] University of Copenhagen, Copenhagen, Denmark

[2] University of Malta, Msida, Malta

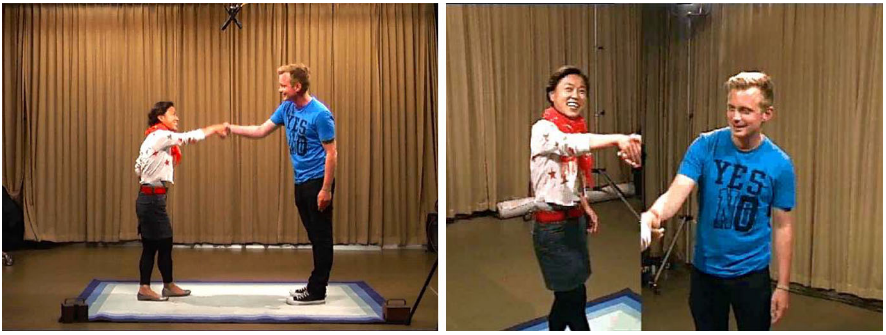


Fig. 1 Recordings from the Danish NOMCO dialogues: total and split views

theoreticians in gesture studies to re-define the realm of linguistic analysis to encompass gestural behaviour goes hand in hand with the development of multimodal corpora, where subjects are video-recorded while they interact in different types of communicative situation, and their speech and gesture behaviour is annotated with rich descriptive features. The existence of such corpora, and of specialised tools for their annotation and analysis, provides a unique opportunity for researchers from different fields to work on naturally occurring multimodal data.

In this article, we describe the Danish NOMCO corpus, which we consider an important contribution to the fields of multimodal corpora and gesture studies, not only because the corpus has specific and interesting properties related to the communicative situation in which it has been collected, but also because we believe the methods used to annotate and analyse it will be helpful to the research community. Note that the term *modality* in this work is used to refer to production modality (speech, and different types of gestural behaviour, e.g. head movements, facial expressions, and body posture). Following this definition, an annotated *multimodal* corpus is a video-recorded collection in which contributions in two or more of these modalities are annotated.

We start in Sect. 2 by describing the way the data were collected. Then in Sect. 3 we describe the annotation methods used to annotate speech as well as gestural behaviour and give counts of the various annotation features. We also analyse the relation between gesture and speech, and the way emotional attitudes are expressed. Finally, we provide an account of how inter-coder agreement was measured. In Sect. 4, we discuss a number of phenomena in light of the annotated data. These include the issue of temporal coordination between speech and gesture, the relation between gestures and focusing, and the mechanisms of multimodal feedback and turn management. The last part of the section is dedicated to an overview of results from machine learning studies carried out on the NOMCO data. Section 5 contains the conclusions.

Table 1 Self-assessment scores (Likert scale 0–5, N = 12)

Variable	Mean	SD
Enjoyable	4.42	0.72
Intimate	2.71	1
Liked	4.04	0.91
Interesting	4.17	0.76
Influence	3.75	0.79
Free	4.13	0.74
Not affected	3.46	1.06
Natural	2.33	1.05
Pleased	4.58	0.58
Relaxed	3.58	1.06
At ease	3.83	0.82
Content	4.46	0.88

2 The recordings

The Danish NOMCO corpus is one of a collection of first acquaintance dialogues created under the auspices of the Nordic NOMCO project. The collection consists of video-recorded and annotated conversations in Danish, Swedish, Finnish, and Estonian (Paggio et al. 2010), comparable with one another for the type of dialogues, the recording setting, and the annotation methodology. Recently, a similar corpus was also recorded for Maltese (Paggio and Vella 2014).

The Danish corpus, which is the focus of this article, consists of twelve recordings, featuring six male subjects and six female subjects of age 21–36, each taking part in a dialogue with a female and one with a male, for a total of about an hour of interaction. The two conversations took place on different days, and in both cases the dialogue participants had never seen each other before. They were told that they had about five minutes to get to know each other, as if they had been at a party, or a similar situation. As a consequence, they spoke freely about any topic they wanted. The dialogues were recorded in a studio, with the participants standing in front of each other on a carpet to delimit the possible distance between them. Each dialogue was filmed by three cameras, as shown in Fig. 1. The video format is MOV (six files) and AVI (six files), both with CINEPAK as codecs. The audio is uncompressed (44.100 Hz), and five files were recorded in stereo while seven are in mono format. The three camcorders and two cardioid microphones used were synchronised by the IT and Media group at the faculty of the Humanities of the University of Copenhagen.

It was a goal of the project to create a multimodal corpus of natural and free conversations. It was important, therefore, to ensure that the conversations proceeded in as natural a way as possible in spite of their being recorded in a studio. Therefore, the participants were not made to wear any kind of equipment, not even microphones. In addition, to assess how much they were affected by the artificial setting, after each conversation the subjects filled in a questionnaire containing questions about setting, interaction and emotional attitudes felt during

Table 2 Word statistics

Statistic	Count
Total no. speech tokens	18,556.00
Total no. speech types	3002.00
No. tokens per speaker per dialogue (mean)	1546.33
No. tokens per speaker per dialogue (SD)	173.77

Table 3 Frequent speech tokens

Frequency	Token	Gloss
698	Breath	Breath
527	j,a	Yes (stressed)
333	så	Then
330	det	It
326	jeg	I
298	og	And
284	øh	Oh (interjection)
222	laugh	Laugh
203	[false_start]	False start
197	smack	Smack
167	men	But
166	ok,ay	Okay (stressed)
157	i	In
152	det_er	It is
147	sådan	So
132	på	On
121	ja	Yes
113	med	But
111	man	One
110	der	There
105	eller	Or
103	n,ej	No (stressed)
101	mm	Um (filled pause)
99	ikke	Not
99	du	You

the dialogues (Paggio and Diderichsen 2010). Each question had to be answered by assigning a score on a Likert scale from 0 to 5. The results are shown in Table 1. In general, the participants were positive about the interaction and not too affected by the setting. They felt quite free to express themselves (average score for *Free* is 4.13), and relaxed (average score for *Relaxed* and *Not affected* are 3.58 and 3.46, respectively). It is crucial that the subjects scored these dimensions positively even though they judged the setting slightly unnatural (average score for *Natural* is 2.33).

To sum up, based on the results of the questionnaires, and given the important fact that they were not scripted, the NOMCO dialogues can be considered close to naturally occurring conversations, and studied in view of understanding naturally occurring multimodal behaviour in first acquaintance encounters.

3 The annotation

3.1 Transcription and annotation of speech

An orthographic transcription of the spoken contributions was done using PRAAT (Boersma and Weenink 2009). The transcription includes word boundaries as well as word stress, indicated by a "," before the stressed vowel. Pauses are represented by a "+", and filled pauses transcribed as fillers, e.g. *mm*, or glossed with English words, e.g. *laugh, breath*. The PRAAT transcriptions were then imported into the ANVIL tool (Kipp 2004), which was used for the gesture annotation.

Statistics concerning speech tokens and types are shown in Table 2. If we look at frequency, we see that fillers such as *breath, laugh, smack, mm*, but also feedback words like *ja, okay, nej*, are all among the 25 most frequent speech tokens, shown in Table 3. False starts are also quite frequent. These frequency patterns are typical of spoken language, and conversational data in particular. We will discuss the role of feedback words and feedback gestures in more detail in Sect. 4.3.

In order to be able to investigate the relation between gestures and focusing, the transcription was added an annotation of information structure following a methodology used in previous studies on different Danish data (Paggio 2006a, b). First of all, utterance boundaries[1] were found and annotated with the attribute "boundary true". For this annotation, syntactic cues, but also pauses, repairs etc. were considered. Secondly, for each sentence-like utterance, topic and focus were identified, and the attributes "topic true" and "focus true" were added to the corresponding words in the ANVIL annotation. In short, *topic* indicates the presupposed entity about which the sentence predicates something new, while *focus* indicates non-presupposed information. Words that do not belong to either topic or focus, are considered background and left untagged. Not all sentences have a topic, whereas the focus is always present. The annotation guidelines include principles for how to assign topic and focus in general, as well as in specific syntactic constructions such as clefts, epistemic constructions, and topicalised sentences.

For a simple example, consider the following short exchange from one of the conversations (',' indicates stress, "+" stands for a pause, and boldface marks topicality):

Speaker A: + **jeg** hedder Chr,esten +
Speaker B: + h,ej + **jeg** hedder T,anja +
(my name is Chresten
hi my name is Tanja)

[1] We relied on the definition of utterance proposed in Levinson (1983), where an utterance is defined as "the issuance of a sentence, a sentence-analogue, or sentence-fragment, in an actual context" (p. 18).

```
<el index="2" start="98.61222" end="99.22767">
  <attribute name="token">+_</attribute>
  <attribute name="boundary">true</attribute>
</el>
<el index="3" start="99.22767" end="99.3199">
  <attribute name="topic">true</attribute>
  <attribute name="token">jeg_</attribute>
</el>
<el index="4" start="99.3199" end="99.49494">
  <attribute name="token">hedder_</attribute>
</el>
<el index="5" start="99.49494" end="99.9156">
  <attribute name="token">T,anja_</attribute>
  <attribute name="focus">true</attribute>
</el>
<el index="6" start="99.9156" end="100.05865">
  <attribute name="token">+_</attribute>
  <attribute name="boundary">true</attribute>
</el>
```

Fig. 2 Orthographic transcription example

Table 4 Distribution of focus and topic in the corpus

Speech type	Proportion
Focus words	0.36
Topic words	0.06
Background words	0.58
Total	1

In both turns, the subject *jeg* is the topic, while the focus is the name of each person. In speaker B's turn, the word *hej* is also in focus. Pauses correspond to sentence boundaries. Figure 2 displays the orthographic transcription of the first turn in the ANVIL XML format. Each element, enclosed by "el" tags and marked by start and end time points, corresponds to a speech token.

Table 4 shows the distribution of words with respect to the three information structure categories. Words belonging to the focus make up for more than one third of the material, whereas topic words, as expected, are only a small percentage. The total number of syntactic clauses in the corpus is 2955, with 6.255 words per clause on average. The average length of a focus phrase is 2.26 words.

3.2 Annotation of gestural behaviour

The gestural behaviour annotated in the corpus concerns head movements, facial expressions, and body posture. In our terminology, they correspond to different modalities of expression, or modalities of production, as also proposed in Allwood (2002). Modalities of expressions should not be confused with sensory modalities (vision, hearing, etc.), which are relevant when discussing perception or reception of

Table 5 Annotation features for gesture shape and dynamics

Attribute	Value
HeadMovement	Nod, Jerk (Up-nod), HeadBackward, HeadForward, Tilt, SideTurn, Shake, Waggle, HeadOther
HeadRepetition	Single, Repeated
Eyebrows	Frown, Raise, BrowsOther
General face	Smile, Laughter, Scowl, FaceOther
BodyDirection	BodyForward, BodyBackward, BodyUp, BodyDown, BodySide, BodyTurn, BodyDirectionOther
BodyInterlocutor	BodyToInterlocutor, BodyAwayFromInterlocutor
Shoulders	Shrug, ShouldersOther

communicative signals. The annotation was done using the ANVIL annotation tool (Kipp 2004), and following the MUMIN coding scheme, which has proven useful for annotating gestural behaviour in terms of its shape and dynamics as well as communicative function (Allwood et al. 2007). Only a subset of the attributes defined in MUMIN was used for the modalities considered in NOMCO: eye-gaze features, for example, were not assigned, due to the difficulty of doing so manually given the quality of the recordings and the angle from which the subjects have been filmed. Moreover, MUMIN also provides annotation features for the hand gesture modality, which was not targeted in the project. On the other hand, new features were developed for the annotation of emotional attitudes, as will be detailed below.

3.2.1 Gesture shape and dynamics

The attributes and values used to annotate gesture shape and dynamics, shown in Table 5, are relatively coarse-grained. They are meant to capture the various gesture types that serve different communicative functions rather than provide a detailed description of the movement. For example, the scheme distinguishes between a *nod* and a *shake*, but does not give the possibility to describe qualities of nodding, or shakes of different sizes. If a more fine-grained annotation is needed, it could be provided by computer vision analysis techniques.[2]

For most of the attributes, a category called *Other* (HeadOther, FaceOther, etc.) is available to the annotator for cases in which no other value seems to fit the data. As far as shape is concerned, Other was mainly used to annotate cases in which two of the other categories were combined (e.g. BodyForward and BodySide). The frequency of the category is 6 % for head movements, 6 % for facial expressions, and 27 % for body direction. This indicates that the range of categories for the annotation of body direction could be developed further.

[2] A step in this direction was taken by developing a face and head tracker ANVIL plugin-in (Jongejan 2010) which can be used to further annotate the corpus.

The total number of gestural behaviours, together with average and standard deviation per speaker per conversation, is shown in Table 6. More detailed counts of the various gesture types are provided in the Appendix. Box plots of the distribution of behaviours in the three modalities across speakers are also shown in Fig. 3. Head movements constitute by far the most frequently occurring behaviour type, followed by facial expressions and body movements. There is also quite a bit of speaker variation, especially in the use of facial expressions, where we see outliers at both ends of the distribution. As already mentioned, the gestural behaviours annotated only refer to movements or expressions that were considered communicative by the annotators. In other words, so-called adaptors (e.g. scratching one's nose) were not considered. The distinction between communicative and non-communicative gestures was proposed already by Kendon (1978). For a more recent discussion of its reliability in annotation of head movements, see Kousidis et al. (2013).

An interesting question is whether there are correlations between richness in spoken language output and gestural expressivity. This was measured by testing for correlations between number of words and number of gestural behaviours of each type produced by the individual speakers. There is a low to moderate correlation (Pearson's r = 0.42) (Dancey and Reidy 2004) between number of words spoken and number of head movements produced, which explains 58 % of the variation. In turn, this means that there must also be genuine individual variation in how much speakers use head movements when they speak. The correlation is even weaker between speech and the other two modalities.

Interdependencies between the number of behaviours produced in the three gestural modalities (independently of speech) were also investigated, but no correlations were found. This may be due to different reasons. Behaviours in different modalities, for example head movements and facial expressions, may on average have quite different durations: for example, a facial expression may overlap with several head movements, so that they will not correlate in number. In addition, genuine individual differences may also be at stake here. Thus, some people may tend to use their heads a lot, while their faces are not very expressive. Others may change their posture often without necessarily moving their heads, or viceversa.

3.2.2 Gesture function

Of particular interest in NOMCO was the annotation and analysis of communicative functions, mainly related to the regulation of the conversation interchange between speakers, in other words the diverse roles carried out in conversations by *interactive* gestures, also known as *regulators* (Ekman and Friesen 1969). Such communicative functions are indeed the main focus of the MUMIN coding scheme. Thus, the scheme provides features to annotate gestures related to *feedback*, *turn management*, and *sequencing*. The features were streamlined and slightly simplified in order to be applied to the NOMCO data, but the original spirit of the annotation scheme was basically maintained. Several of the features defined in the scheme are very similar to standard categories for dialogue act analysis, as proposed in Bunt et al. (2012).

Table 6 Gesture statistics

Modality	Sum	Mean per speaker	SD
Head movements	3117	129.88	34.59
Facial expressions	1448	60.33	24.89
Body postures	982	40.92	17.44
All modalities	5547	231.13	55.69

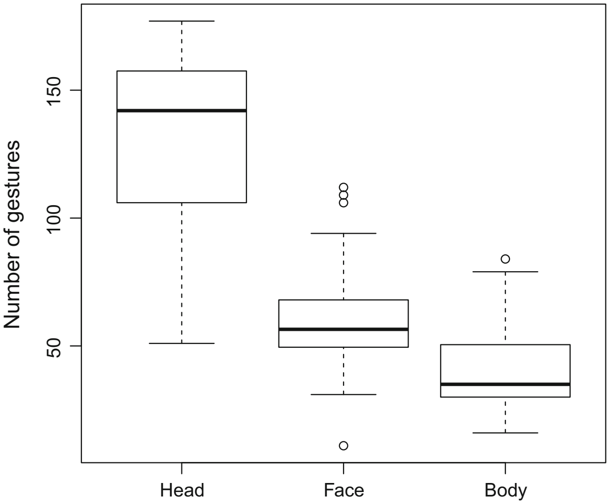


Fig. 3 Distribution of gesture types across speakers in the NOMCO corpus

Feedback is defined in Allwood et al. (1992: p. 1) and Allwood et al. (1993), as the mechanism through which speakers exchange information about (1) *contact*, the fact that participants are willing and capable of continuing to interact; (2) *perception*, the fact that they are willing and capable of perceiving what is being communicated; and (3) *understanding*, the fact that they are able to understand the message that is being communicated. In practice it is very difficult to discern these specific aspects, and therefore our scheme combines them into the unified value *CPU*, which stands for *contact, perception* and *understanding*. In addition, feedback has a *direction* depending on whether it is being given or elicited (or both, or unspecified), and an *agreement* feature to specify agreement or disagreement.

The annotation of turn management relies on a distinction between turn change achieved in agreement or as a result of an interruption. In the former situation, the two relevant values are *TurnElicit* and *TurnAccept*, in the latter *TurnTake* and *TurnYield*. The value *TurnHold* is used for a behaviour which indicates that the speaker wants to keep the turn, and *TurnComplete* for one where the speaker stops speaking without the turn being picked up by the interlocutor.

Table 7 Annotation features for gesture functions and semiotic classes

Attribute	Value
FeedbackBasic	CPU, SelfFeedback, FeedbackOther
FeedbackDirection	FeedbackGive, FeedbackElicit, FeedbackGiveElicit, FeedbackUnderspecified
FeedbackAgreement	Agree, NonAgree
Turn	TurnTake, TurnAccept, TurnYield, TurnElicit, TurnComplete, TurnHold
Sequencing	SeqOpen, SeqResume. SeqContinue, SeqClose
SemioticType	IndexDeictic, IndexNon-deictic, Iconic, Symbolic, SemioticOther

Sequencing is concerned with the structuring of the discourse. Thus, discourse sequences can be opened and closed. They can be resumed after an interruption or continued. All these situations can be signalled by gestural behaviour.

Head movements, facial expressions and body posture movements were all annotated with features referring to the three communicative functions just discussed. In addition, the annotation scheme also provides features to assign gestural behaviours to semiotic classes. As suggested in Allwood (2008), the scheme builds on Peirce's three categories *indexical, iconic* and *symbolic* (Peirce 1931). Indexicals are sub-divided into *deictic* gestures, which point to an entity in the conversation situation, and *non-deictic* gestures, which include *displays*, *beats* (also sometimes called *batonic*), and other indexical gestures with *interactive* function, e.g. head movements used to give and elicit feedback. It must be remembered, however, that the same gesture type can play different functions (and belong to different semiotic classes), depending on the context. A nod, for example, can be a symbol when it corresponds to an acceptance or agreement act, or it can function as a beat that accompanies a stressed word. In fact, one and the same gesture can often be interpreted at different levels: to stay with the same example, an affirmative nod will typically also function as a beat. In the current version of the NOMCO corpus, the annotation of semiotic classes is limited to the two classes deictic (101 total occurrences) and iconic (16 total occurrences).

Attributes and values referring to communicative functions and semiotic classes are displayed in Table 7. For some of the functional features, the annotators could use the value *Other*. However, the value was never used in the annotation of feedback. In the annotation of emotions it was employed by the annotators to mark when they wanted to use a new emotion label. Then, three annotators adjudicated whether the marked emotion was missing from the existing emotion list, and if so a new label and corresponding PAD values were created (the PAD values are explained in Sect. 3.2.4).

Statistics concerning the number of behaviours associated with feedback, turn management and sequencing in the three gestural modalities are shown in Table 8. For each modality, the table shows the total number of gestures and the proportions of these gestures that have been annotated with features related to the three communicative functions. Note that in many cases, the same gesture may be

Table 8 Gesture function statistics

Modality	Total	Feedback (%)	Turn (%)	Sequencing (%)
Head movements	3117	0.53	0.27	0.07
Facial expressions	1448	0.73	0.17	0.05
Body postures	982	0.47	0.23	0.06

Table 9 Speech sequences linked to gestures

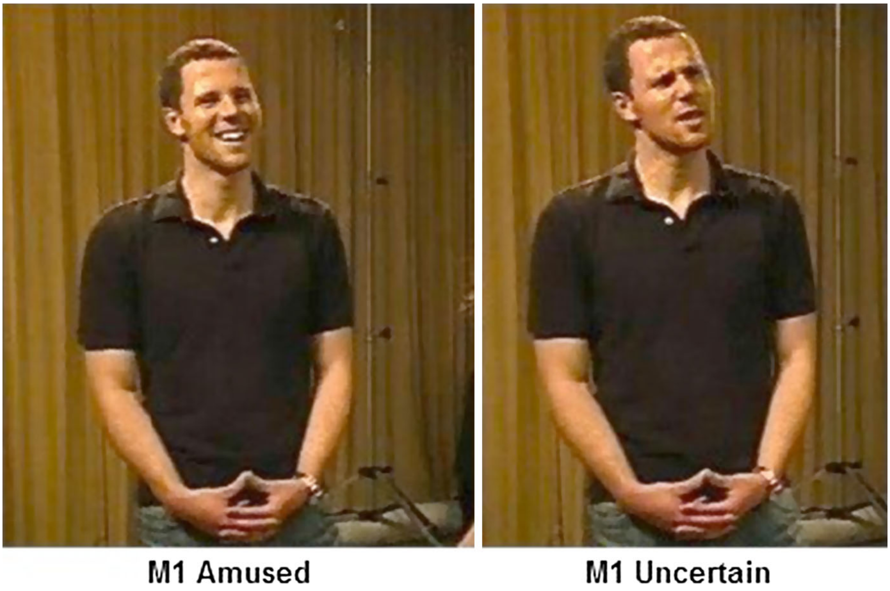
Speech	Face (%)	Head (%)
Pause	0.08	0.06
Filled pause	0.16	0.09
One unstressed word	0.07	0.08
One stressed word	0.21	0.30
Several words	0.48	0.47
Total	1	1

annotated with a feature in either two or even three of the functions. In other words, the functions are not mutually exclusive, and the proportions do not therefore add up. It can be noted that, in terms of functional content, head movements and body postures are similar in that in roughly 50 % of the cases they serve a feedback function, and in about 25 % of the cases they are used in turn management. Facial expressions, on the other hand, are related to feedback more often (73 %), and to turn management less so (17 %). The proportion of behaviours related to sequencing is quite low for all three modalities (5–7 %).

3.2.3 Relation between gesture and speech

In addition to shape and function features, for each gesture under consideration, a relation with the corresponding speech expression is also explicitely annotated. Two attributes are used for this purpose. One is *MMRelationSelf*, which establishes a link between the gesture under consideration and the semantically related speech token or tokens in the orthographic transcription of the gesturer's speech. The other is *MMRelationOther*, which is used to codify a relation between a gesture and the interlocutor's speech in cases where the gesturer is silent.

The adoption of these two relations is motivated by a wish to provide a relation between gestures and corresponding speech based on semantics rather than mere temporal alignment between the two elements. The relations are thus reminiscent of the notion of *lexical affiliate* (Schegloff 1984; Kipp 2004), although they are applied to any kind of gesture, not just iconic ones. Note also that in almost half of the cases, they link the gesture to a speech sequence consisting of two or more speech tokens, as shown in Table 9. Having defined an explicit link between gestures and the speech sequence they are associated with, allows us to study the issue of how the



M1 Amused M1 Uncertain

Fig. 4 Facial expressions of emotion in the NOMCO corpus

two modalities are coordinated both from a content-oriented point of view (based on the explicit link), and at a temporal level (based on time stamps in the annotation).

3.2.4 Emotional attitudes

Emotion studies often deal with the basic emotions described by Ekman and Friesen (1975), and Ekman (1992), and are studied in acted data (Bourbakis et al. 2011; Kipp and Martin 2009), or in specialised situations like clinical settings (Lucey et al. 2012; Aung et al. 2014), or computer games (Savva et al. 2012). Emotional behaviour in the NOMCO conversations mainly consists of emotional attitudes (Allwood et al. 2007), also called affective epistemic states (Allwood et al. 2014), which concern the way people feel about the communicative situation, the interlocutor and the content of the ongoing conversation. Such emotional attitudes are quite different from the basic emotions, and also from the reactions expressed by subjects in clinical settings or during games: examples are the feelings of being *amused*, *uncertain*, *engaged* or *surprised*. They are also often not so strong or easily identifiable as acted basic emotions, so generalisations made on the basis of acted data or data concerning different communicative and interactional settings may not carry over to the expression and effect of emotional attitudes in normal face-to-face conversation. Figure 4 illustrates facial expressions of amusement and uncertainty in our data.

Emotions have often been classified via emotion labels (Ekman and Friesen 1975), but they have also been described in terms of their position in more or less complex dimensional spaces. One of the dimensional models which has been used

Table 10 Emotion labels and corresponding PAD values

P	A	D	Emotions
+	+	+	Amused, Excited, Happy, Interested, Ironic, Joking, Proud, Satisfied, Self-Confident, Supportive
–	–	–	Disappointed, Hesitant, Unconfident, Uncomfortable, Uninterested
+	–	+	Certain, Friendly
–	+	–	Awkward, Embarrassed, Puzzled, Shy, Uncertain, Uneasy
+	+	–	Engaged, Surprised
+	–	–	Docile, Thoughtful
–	+	+	Irritated
–	–	+	None

to describe emotions in communication is the three-dimensional model proposed by Russell and Mehrabian (1977) with the three emotions being *P*leasure, *A*rousal and *D*ominance—the PAD model henceforth. In the MUMIN framework, which is at the basis of our work, emotions are annotated via an open list of emotion labels to reflect the fact that affective states and attitudes, corresponding to minor emotions in Ekman (1992), are often more frequent in communication than the six basic emotions, so that which emotion labels are relevant depends on the communicative situation.

In our corpus, an annotation of emotions and emotional attitudes was added to facial expressions based on both coding styles: the annotators had to choose a label from a list of 28 emotions at the same time as picking a value for each of the three dimensions in the PAD scheme. Emotion labels were added incrementally as needed during the annotation. The resulting list of annotated emotions and their PAD values is shown in Table 10. In general, it was found that using the PAD values in combination with the emotion labels ensured better inter-coder agreement than using the emotion label list alone (Studsgård and Navarretta 2013; Navarretta 2014). Note that only the emotions conveyed by facial expressions were annotated. However, the annotators used the whole context In order to decide whether a facial expression conveyed an emotion and how to classify it.

Ten of the emotion labels in the corpus have positive values and five labels have negative values in all three of the dimensions; six have a positive value for *Arousal* and negative values for the other two dimensions. The remaining PAD combinations match fewer emotion labels, with the last possibility in the table (negative *Pleasure* and *Arousal* with positive *Dominance*) matching no label and never occurring in the corpus.

The total number of facial expressions annotated with an emotion feature constitutes 70 % of the facial expressions in the entire corpus. The remaining ones were judged to be neutral. Not all emotions are equally represented since conversation participants mainly express positive emotional attitudes towards each other.

Table 11 Results of the first inter-coder agreement experiment: head and face attributes

Attribute	Kappa score		
	Pair 1	Pair 2	Pair 3
Head			
HeadMovement	0.57	0.56	0.52
HeadRepetition	0.62	0.59	0.57
FeedbackBasic	0.62	0.55	0.50
FeedbackDirection	0.62	0.54	0.51
FeedbackAgreement	0.67	0.59	0.59
Face			
General face	0.67	0.49	0.50
Eyebrows	0.71	0.56	0.59
FeedbackBasic	0.67	0.40	0.40
FeedbackDirection	0.67	0.37	0.38
FeedbackAgreement	0.71	0.51	0.56

3.2.5 Annotation procedure and inter-annotator agreement

Six annotators, students and researchers in linguistics or multimodal communication, were involved in the transcriptions and annotations of the shape and functions of gestures.

Speech was transcribed by three expert students who had previously participated in projects involving the transcription of Danish speech using PRAAT. The students corrected each other's transcriptions, discussed and resolved problematic speech segments.

Gestural behaviour, that is communicative head movements, facial expressions and body postures, were annotated one track at the time in this order. However, in all cases the annotators also considered concomitant speech and other behaviours. In other words, the entire context was used to analyse and code the gestural behaviour.

Before annotating the gestures, the annotators were trained in the use of the ANVIL tool and in the MUMIN model and annotation scheme. The training programme included a group annotation of the head movements and facial expressions in one of the videos. The attributes chosen for the inter-coder annotation test were the shape and feedback-related ones. Successively, three coders annotated a second video independently in order to run the test. The inter-coder agreement results of this experiment, expressed in terms of Cohen's *kappa* (Cohen 1960), are in Table 11.

As shown in the table, the inter-coder agreement is not the same for all pairs. The most frequent disagreement cases were: (a) head movements which were identified as one repeated gesture by one coder and as sequences of single gestures by another; (b) the segmentation of facial expressions; (c) the distinction between jerks and nods; (d) the choice of the primary value in cases of multifunctional gestures, especially feedback and self-feedback signals, feedback giving and eliciting signals[3]; and (e) categories which were left uncoded.

Disagreement cases were discussed, and annotation strategies as well as new annotation guidelines were developed. Furthermore, a procedure for checking contradicting or missing values was agreed upon. A second inter-coder agreement experiment involving ony two of coders (pair 2) was then run. The results, given in Table 12, constitute an average improvement of 0.9 (Navarretta et al. 2012).

The figures indicate that the annotators' agreement is 0.63 on average (0.55–0.69). Given the difficulty of the task and the fact that the figures cover both the segmentation and classification of gestures, this level of agreement is acceptable, and compares well with measures provided in connection with other multimodal corpus annotation tasks (Cavicchio and Poesio 2009).

Inter-coder agreement was also measured for the annotation of emotional behaviour, as described in detail in (Navarretta 2012). The annotators could choose between 26 available emotion labels and their corresponding PAD values. Sixteen emotion labels were assigned during the experiment. Seven PAD combinations were annotated. The resulting Cohen's kappa scores, which are shown in Table 13, are in line with results reported in other studies on the annotation of the six basic emotions in acted data.

To produce the final annotated corpus, one coder annotated a behaviour and a second one checked the annotations. Disagreement cases were resolved by a third coder (Paggio and Navarretta 2011). Procedures for checking missing or inconsistent annotations in the corrected and final version were then applied by an expert annotator.

4 Studies

In this section we discuss several characteristics of the NOMCO corpus that shed light on theoretical aspects of multimodal communication. Some of these characteristics have been treated in previous publications, which we mention along the way. However, more detail is provided here, and this is also the first time that these analysis results are reported together in a systematic way. We start by analysing the temporal coordination between speech and gesture, and the related issue of how gesture contributes to the expression of focus. Then we describe the role of gesture in the mechanisms of feedback and turn management. Finally, we summarise a number of machine learning experiments that were conducted on the corpus to predict several communicative behaviours.

[3] In most cases one coder chose one category as the primary and indicated another possible category in the comment field, while the second coder chose the second category as the primary and mentioned the first one in the comment field.

Table 12 Results of the second inter-coder agreement experiment: head and face attributes

Attribute	Kappa score
Head	
HeadMovement	0.60
HeadRepetition	0.62
FeedbackBasic	0.64
FeedbackDirection	0.64
FeedbackAgreement	0.68
Face	
General face	0.61
Eyebrows	0.67
FeedbackBasic	0.57
FeedbackDirection	0.55
FeedbackAgreement	0.68

Table 13 Results of final inter-coder agreement experiment on the annotation of emotions

Attribute	Kappa
Emotion labels	0.61
Pleasure	0.67
Arousal	0.54
Dominance	0.64

4.1 Coordination of speech and gesture

Many studies have claimed that speech and gesture, in particular hand gestures, are two manifestations of the same underlying cognitive mechanism (McNeill 1992, 2005; Kendon 2004; Kita and Özyürek 2003; De Ruiter 2000). One aspect of this tight relation is the temporal coordination between the two modalities. There seems to be general agreement about the fact that hand gestures are coordinated with prosodic events, such as pitch accents and prosodic phrase boundaries (Bolinger 1986; Kendon 1980; Loehr 2004, 2007). This temporal coordination has also been studied experimentally by manipulating the synchronisation between the visual and the auditive streams in video-recorded stimuli. The results indicate that subjects are sensible to asynchrony, especially when gesture strokes are made to lag behind the accompanying speech (Leonard and Cummins 2010), and also that coordination with prosody contributes to the well-formedness of multimodal signals (Giorgolo and Verstraten 2008).

These studies deal with hand gestures, especially those that are used as beats. Head movements often have the same quality of manual beats, by being rapid, simple and often repeated movements. Therefore, we would expect them also to show tight temporal synchronisation with the words they co-occur with. Temporal synchronisation between head movements and speech is dealt with in Hadar et al. (1985), where it is argued that coordination with speech, together with physical

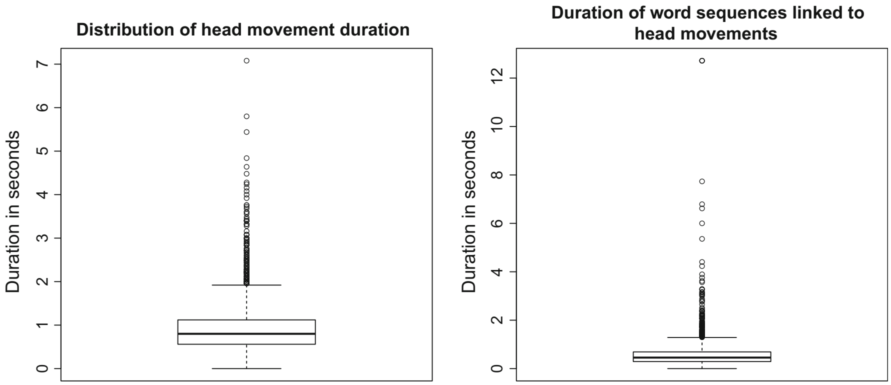


Fig. 5 Duration of head movements and associated speech sequences in the NOMCO dialogues

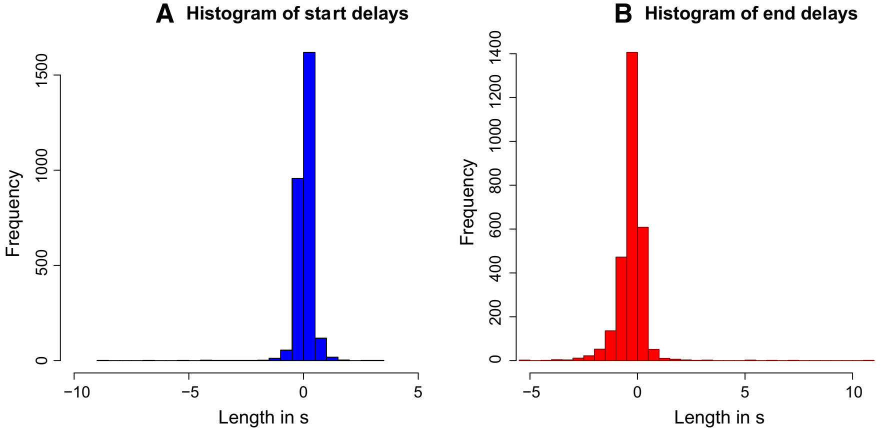


Fig. 6 Start and end delays in the NOMCO corpus. In histogram **a**, *bars* to the left of zero (negative) correspond to speech preceding the onset of the corresponding movements, and those to the right to speech onset following movement onset. In histogram **b**, *bars* to the *left* of zero count speech ending before, and those to the *right* speech ending after movement offset. Histogram bins correspond to intervals of half a second

properties of the movements (cyclicity, amplitude, duration) are indicative of the diverse communicative functions of head movements.

As we saw earlier, in the annotation of the NOMCO corpus an explicit link is established between gestural behaviours and the speech sequences they are semantically and temporally related with. The links were used to derive measures of start and end delays between head movements and associated speech. We are only interested in head movements that are linked to word sequences in the gesturer's own speech stream, which are in total 2795. The remaining head movements are unimodal signals, and are ignored here. As shown in the left-hand graph in Fig. 5, the duration of most head movements in the NOMCO corpus is

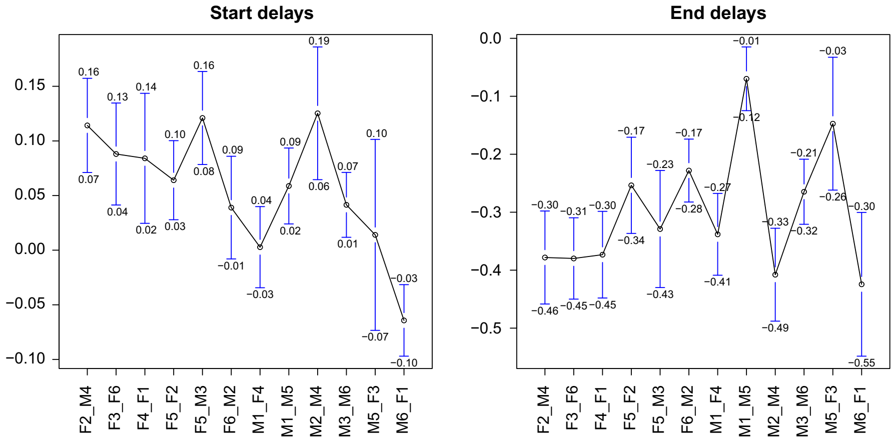


Fig. 7 Start and end delays in the NOMCO dialogues: means and confidence intervals

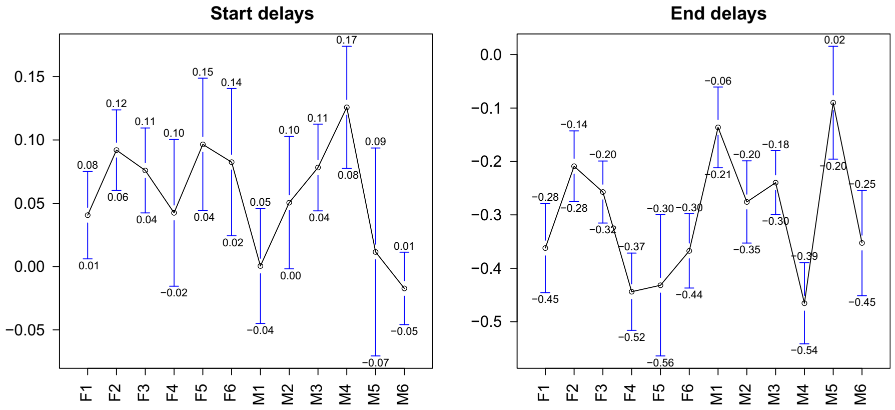


Fig. 8 Start and end delays across speakers: means and confidence intervals

around 1 s, although there are occurrences up to 7 s (mean = 0.93 s, SD = 0.58 s). The duration of the word sequences linked with the head movements (see same figure) is on average slightly shorter with single outliers, however, up to 8 and 12 s (mean = 0.59 s, SD = 0.67 s). On average, head movements tend to start 0.05 s before the onset of the associated speech sequence (SD = 0.40 s), and to end 0.28 s after its offset (SD = 0.64 s).

The histograms in Fig. 6 show that in more than 2500 cases, delays range between −0.5 and 0.5, and that about 1750 delays are in fact positive delays between 0 and 1. In other words, in almost two thirds of the cases head movements start before the corresponding speech. As for the end delays, slightly more than 1800 are between −1 and 0, showing that in almost two thirds of the cases, the head movement ends up to 1 s after speech offset. To have an intuition of what a one second delay means, in Leonard and Cummins (2010) it is found that subjects are

Table 14 Co-occurrence of head movements with focused words

	No.	%	Exp. (%)
Head on focus	1959	0.63	0.36
Head on non-focus	1158	0.37	0.64
Total	3117	1	1

sensible to asynchrony of as little as 0.2 s if a gesture lags behind speech, whereas in Giorgolo and Verstraten (2008) it is claimed that subjects react to gesture-speech misalignments of at least 0.5 s. Thus, a delay of 1 s is not negligible.

If we look at the delay data in the individual dialogues, shown in Fig. 7, we see that there are slight differences between the individual dialogues, with the mean start delay varying from 0.13 to −0.06, and the mean end delay in the range −0.07 to −0.42.

The variation of the length of the delays in the dialogues reflects variation across the individual speakers, which is shown in Fig. 8.

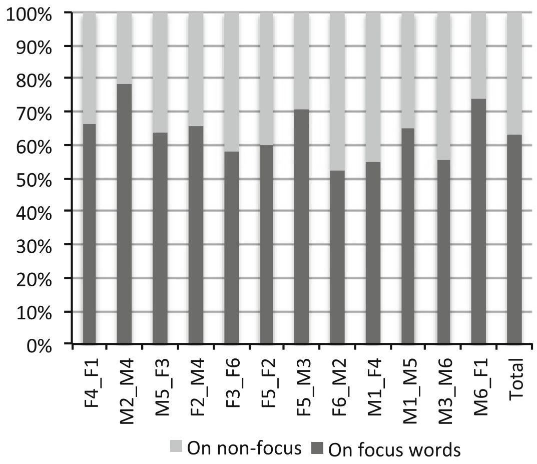
In spite of the variation, however, the general picture seems to indicate that head movements tend to start slightly before the onset of the corresponding speech sequence and to end slightly after. We would expect the length of the speech sequence (consisting of more than one word in about 50 % of the cases, see Table 9), and the position of pitch accent in the sequence, to affect the length of the delays in both directions. The function of head movement may also play a role, as argued in (Hadar et al. 1985). For a discussion of these effects, see Paggio (2016).

4.2 Focus and gesture

We saw in the preceding section that gesture strokes and pitch accents are described in the literature as being correlated, in particular it has been claimed that hand gesture strokes occur slightly before the intonation peak of a clause (McNeill 1992). Indeed, in a recent formal semantic approach, it is considered a grammatical constraint for hand gestures to co-occur with words, that such words should be prosodically prominent (Alahverdzhieva and Lascarides 2010). In many languages, prosodic prominence is an indicator of sentence focus (Vallduví and Engdahl 1996 among many others), thus it is reasonable to ask the question whether gesture is related to focus. The issue is investigated in an empirical study of 276 hand gestures in German by Ebert et al. (2011), where the authors look at the relation between whole gesture phrases (including preparation and retraction) and focus phrases. The study shows that on average, gesture strokes tend to precede sentence accent by 0.36 s; and that the onsets of gesture phrases and new-information foci align with a time lag of −0.31 s. No alignment is observed, on the other hand, between gestures and contrastive foci.

All these studies deal with hand gestures. Here, we approach the topic of gesture-focus alignment from the point of view of head movements. We saw earlier

Fig. 9 Distribution of head movements on focus and non-focus words in the NOMCO dialogues



(Table 9) that head movements mostly co-occur with speech elements containing at least one stress. However, not only focused words are stressed in Danish sentences. Thus, in order to understand whether there is a relation between head movements and focus phrases, we counted how many of the head movements in the corpus are associated with words belonging to a focus phrase (tagged as "focus true"). Table 14 shows the counts. Head movements tend to co-occur with focused words about twice as often as it would be expected had the distribution been random ($\chi^2 = 975.2301$, $df = 1$, p value <2.2e−16, expected frequency proportions reflect the relative frequency of each category in the corpus).

There is some, but not a lot of variation in the distribution if we consider the individual dialogues. On average, 163.25 head movements in each conversation are linked to words belonging to the focus (SD: 29.04) and 96.5 on words outside of the focus domain. Figure 9 shows the proportions for the individual dialogues.

It is difficult to provide data from NOMCO to verify the patterns discovered for German in the Ebert et al. study because focus phrases are not annotated. In other words, the annotation marks, for each word, whether it is or it is not part of the focus, but not where the left-hand boundary of the focus domain is. To make a better comparison possible, the beginning of all focus phrases was marked in one of the dialogues. This shows, in fact, that the majority of head movements in that dialogue occur in conjunction with the first word in the focus domain, as can be seen in Table 15. The trend is highly statistically significant ($\chi^2 = 70.9878$, $df = 2$, p value = 3.848e−16).

In conclusion, we can observe a relation between focus and head movements in the Danish dialogues in the sense that most head movements occur in conjunction with words in the focus domain. There is also indication, from one of the dialogues, that head movements in fact align with the first word of the focus domain. In future, a more precise analysis of this relation will be carried out on the entire dataset, and the temporal alignment between the onset of the movement and the onset of the focus domain will also be measured.

Table 15 Head movements and left-hand focus boundary

	No.	%
On first focus word	134	0.58
On other focus words	33	0.14
On non-focus words	64	0.28
Total	231	1

Table 16 Head nods with feedback function

Head nod	Counts (#)	Proportion (%)
Nods with speech	451	0.71
Nods no speech	186	0.29
Total	637	1

4.3 Feedback by speech and head movements

As we saw in Section 7, feedback is one of the most frequent communicative functions of gestural behaviours, in particular facial expressions and head movements (73 and 53 % of the cases, respectively). In this section we look at feedback by head movement, in particular the relation between head movements and feedback words, such as *ja/jo* (yes), *nej/næ* (no), *okay*, and *mhm*. Firstly, we focus on the occurrence of nods and up-nods (rapid down-up movements called “jerks” in the MUMIN coding scheme), since nods in general are the most common and frequently studied head movement type (Duncan 1972; Hadar et al. 1985; McClave 2000). In general, there are 926 nods (746 up-down and 180 down-up) in the corpus, one every 4.27 s on average. Compared with the frequency of nods reported for other cultures, e.g. the study by Maynard (1987), where Japanese speakers are reported to produce a nod every 5.57 s in contrast to only one every 22.5 s for Americans, it would appear that Danish speakers nod quite frequently. Our data cannot be directly compared to those used by Maynard, since the situation and the setup are different. Indeed, the NOMCO speakers may be giving a lot of feedback by head movement because they are particularly polite in first acquaintance dialogues. However, even allowing for such an effect, Japanese speakers do not seem alone in their frequent use of head nodding (Paggio and Navarretta 2011).

About 68 % of the nods in the NOMCO data are used to signal feedback, mostly together with a feedback word but also without, as shown in Table 16.

Table 17 shows how combined speech and nod feedback is distributed across four different types of nods. The up-nod type is typical of Scandinavian languages, and has been described especially for Swedish, where it is even more common than in Danish both in the single and the repeated variant (Cerrato 2007; Navarretta et al. 2012). Interestingly, the distribution of repeated versus simple nods is slightly different depending on whether the head movement is accompanied by a word or not. Thus, repeated nods are relatively more frequent in the absence of words. As far

Table 17 Distribution of nods and up-nods in multimodal feedback

Head nod	With FB words (%)	Without word (%)
Repeated nod	0.472	0.73
Simple up-nod	0.271	0.20
Single nod	0.253	0.07
Repeated up-nod	0.004	0.00
Total	1	1

as up-nods are concerned, there is no difference between multimodal and unimodal signals.[4]

All types of nod occur mostly in conjunction with positive feedback words, as shown in Fig. 10, where it can also be noted that repeated nods occur mostly together with *yes* words.

Another interesting question is whether the distinction between stressed and unstressed words play a role in whether feedback words are accompanied by feedback gestures. To investigate the issue we looked at a larger set of feedback phrases, including repeated words, e.g. *ja ja*, or sequences, e.g. *ja okay*. This resulted in a list of 1382 examples, which were examined to see whether stress has an effect on the occurrence of an accompanying head movement (not necessarily nods).

The results are shown in Table 18. In general, there are many more stressed feedback phrases than unstressed ones. Therefore, they have much higher probability of occurrence both together with a head movement and without. If we look at unstressed feedback phrases, however, the probability of them occurring together with a head movement is nearly half as high as the probability of them occurring without, and the difference is highly statistically significant ($\chi^2 = 25.3187$, $df = 1$, p value $= 4.86\text{e}{-07}$). In other words, although feedback is expressed by means of speech alone more or less as often as together with a head movement, head movements are rarer in conjunction with an unstressed feedback expression. This tendency is not surprising since gestural behaviour in general tends to be associated with stressed words, as was shown in Sect. 3.2.3 (Table 9).

4.4 Multimodal turn management

The multimodal quality of turn management has been pointed out in numerous studies, starting with Kendon (1967), Duncan Jr and Fiske (1977), Goodwin (1981). The focus of these studies has been on specific features, such as the role of gaze (Kendon 1967), mutual gaze (Argyle and Cook 1976), hand gestures (Duncan 1972) and types of head movement (Hadar et al. 1984).

Concerning the relation between speech and gesture in turn management, the study in Duncan (1972) identifies verbal and non-verbal turn giving cues in dyadic conversations. The cues comprise (a) intonational cues, (b) the use of hedges, such as

[4] Unimodal here is intended in the sense of a gesture not accompanied by a word. We do not investigate whether the nod occurs together with other gestural behaviours.

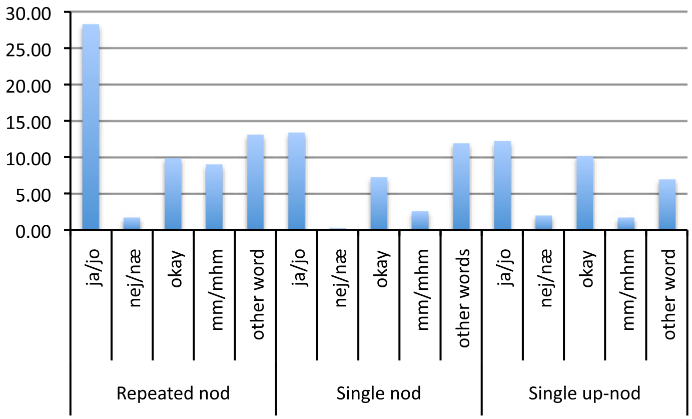


Fig. 10 Distribution of nod types in connection with feedback words: values are given as proportions (%). Repeated head nods are omitted because of their rarity

Table 18 Stress on feedback phrases and accompanying head movements

Stress pattern	Head (#)	No head (#)
Stressed	529	489
Unstressed	134	230
Total	663	719

you know and *I guess*, (c) the syntactic completion of an utterance, and d) the completion of on-going hand gestures as signals that the speaker wants to pass the turn.

In Hadar et al. (1984) it is found that postural shifts of the head tend to occur after "grammatical" pauses and towards the initiation of speech turns or syntactic phrases inside a turn.

Inspired by these studies, we have looked at what types of gesture are related to turn management features in the first encounter corpus, and made a qualitative study of turn shifts and co-occurring gestures in two of the dialogues (Navarretta and Paggio 2013a, b).

As we saw in Sect. 3, 24 % of the occurrences of head movements, 17 % of the occurrences of facial expressions, and 23 % of the occurrences of body posture in the corpus have a turn management function. In Table 19, we show how the turn management function is distributed across the three different types of gesture. It must be noted, however, that turn features are often expressed by several modalities at the same time. In other words, head movements, facial expressions and body postures may reinforce each other in the expression on a turn behaviour.

Figure 11 shows the types of gestural behaviour most often associated with turn management. As can be seen, many types of head movement have a turn management function, not only side turns, shakes and nods as proposed in Hadar et al. (1984). Figure 12 illustrates how the most frequent turn management categories are distributed across the three modalities.

Table 19 Turn management distribution across gesture types

Body behaviour	Proportion (%)
Head movements	0.61
Facial expressions	0.21
Body postures	0.18
Total	1

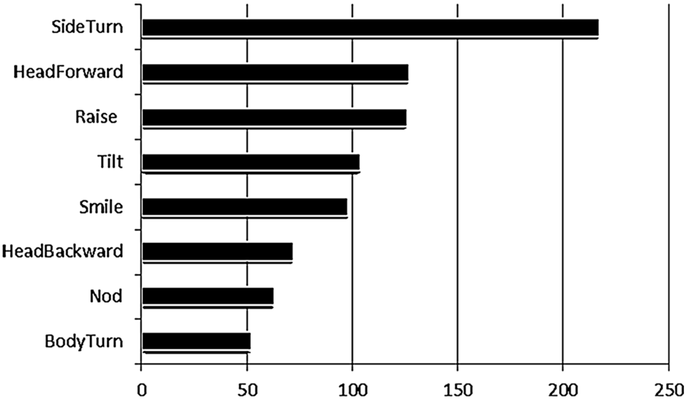


Fig. 11 Most frequently occurring turn management behaviours: absolute counts

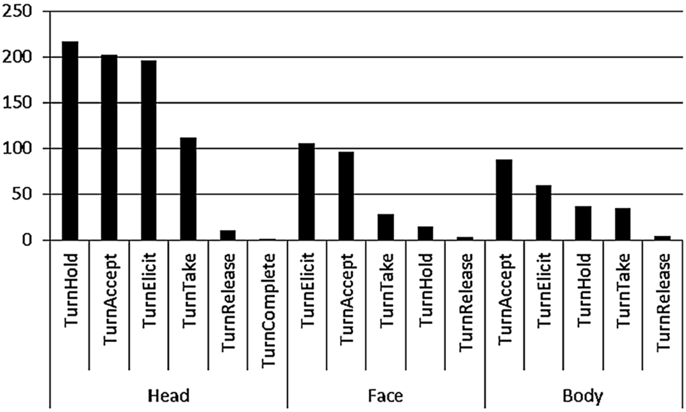


Fig. 12 Turn management related types and body behaviours

The turn management categories we see in the corpus are in accordance with the social activity and conversational setting. The participants meet for the first time and want to make a good impression. They do not discuss controversial issues and do not interrupt each other. Thus, TurnYield is rarely assigned, and there are no occurrences of turn release under pression (Turn Release), while categories such as TurnHold, TurnAccept and TurnElicit are much more common.

Head movements are often related to TurnHold, TurnAccept and TurnElicit, while TurnAccept, TurnElicit and TurnHold are the functions most frequently assigned to body posture. Finally, facial expressions have often the functions of TurnElicit, TurnAccept and TurnTake.

In the qualitative study of multimodal turn management, we investigated how various syntactic, prosodic and gestural features contribute to turn management, inspired by Duncan (1972). Our analysis of turn eliciting cues confirms Duncan's observation that a speaker's completion of a syntactic phrase and a high or low pitch can signal that the speaker wants to relinquish the turn. Hedges and vowel lengthening, on the other hand, occur seldom or not at all in the two analysed dialogues. We also find that the speaker can signal their intention to offer the turn by keeping the head and the body still. This behavior can be seen as parallel to that of finishing off on-going hand gestures, as also noted by Duncan.

4.5 Prediction and validation

A number of machine learning experiments were carried out on the NOMCO annotations. Given the focus of the project on the multimodal expression of communicative functions, the primary objective of the experiments was to determine to what extent these functions can be predicted from the coarse-grained shape annotations of the gestures together with co-occurring speech, and which information contributes mostly to the prediction. A positive outcome would provide insight not only on the relation between gesture shape and its function, but also on the possibility of semi-automatic annotation of similar corpora. We also experimented with prediction of other phenomena in the corpus, or of similar phenomena in different data. In general, an important additional aim was to test whether the annotation scheme distinguishes different behaviours in a consistent and reliable way.

All experiments were run in Weka (Witten and Frank 2005), and ten-fold cross validation was applied in the evaluation. Five classifiers were tested on most tasks: Naive Bayes, KStar, BFTree, logistic regression and support vector machine. The results of a majority classifier were used as lowest baseline, and the results achieved on various datasets were compared. Detail on datasets and results can be found in a number of previously published papers (Paggio and Navarretta 2012; Navarretta and Paggio 2012, 2013a; Navarretta 2011, 2013a, b). Here, we want to reflect in general over the knowledge we have acquired from those studies.

Two studies addressed the prediction of the function of gestures from their shape together with information of co-occurring speech. In all these tests, the classifier which performed best was the support vector classifier.

In the first study, classifiers were trained on the annotations to predict the feedback function of head movements and facial expressions. As seen in Sect. 4.3, feedback is the most frequently occurring function assigned to these gesture types in our data. The best result, with an F-score of 0.76, was obtained with the classifier that combined features of head movements, facial expressions and co-occurring speech. Considering head movements and facial expressions separately, on the other hand, yields an F-score in the range 0.63–0.65. In all cases, the classifiers outperform the majority baseline. These results not only indicate that feedback is annotated in a reliable way, but also confirm the fact that feedback in face-to-face communication is inherently multimodal. Moreover, the accuracy reached by the classifiers is similar to the accuracy shown by the human annotators performing the same task, which in turn indicates that the NOMCO data can be used as training data for the automatic annotation of multimodal feedback.

In the second study, supervised machine learning was applied to the classification of multimodal turn management signals. Features concerning head movements, facial expressions as well as body posture were included. The F-scores obtained by the various classifiers are in the range 0.4–0.46. Although this level of accuracy is higher than what is yielded by the majority baseline, it is considerably lower than the results obtained for feedback classification. A number of factors, however, make the task more difficult. While head movements and facial expressions in our data often have a feedback function, turn related behaviours involving these two modalities are much rarer, a fact that clearly influences classification. In fact, turn management is often expressed via prosodic cues, gaze, hand gestures and, as a qualitative analysis of the data seems to indicate, pausing in gesturing. In future, therefore, we will test whether the classification of turn management behaviours can be improved by taking into account hand gestures and prosodic features.

A phenomenon that is closely related to turn management is speech overlapping. Several studies have shown that overlapping is not infrequent in conversational data (Campbell and Scherer 2010). Indeed, there is overlapping in 90 % of the contributions in our data. Overlapping speech mainly occurs in back-channeling or in cooperative speech, in other words the interlocutor helps the speaker to complete an utterance, e.g. suggesting a word, or repeats part of the speakers utterance. In all these cases, the interlocutor does not interrupt the speaker and there is no turn shift. We conducted a study to investigate to what extent it could be predicted based on features coming from different modalities. A number of classifiers were trained on various feature combinations to predict overlapping speech tokens in two ways—either by considering speech and gestural features in the overlapping segment, or by using multimodal features of the contexts preceding and following the overlaps.

When we only consider features of the overlapping segments, we see that facial expressions in combination with the speech tokens are useful for the classification. Head movements and body postures, on the other hand, are not: they simply do not often occur at the same time as overlapping speech. If we take the context into consideration, however, the picture changes. Here, the most accurate results, with an F-score of 0.83, are obtained by a Naive Bayes classifier trained on features concerning three speech tokens before and after the overlap, together with the co-occurring behaviours from all three gestural modalities.

If machine learning experiments on the corpus shed light on the way in which multimodal behaviour is realised in first acquaintance dialogues, trying to apply the models built on the NOMCO data to other domains helps understand how general the models are. Therefore, in the fourth study we want to mention here, we tested how a classifier trained on the NOMCO data could predict feeback in another conversational corpus, and viceversa. The other corpus used in these experiments was the DK-CLARIN corpus, which consists of dyadic and triadic spontaneous and naturally occurring conversations recorded at the participants' private homes. The two corpora are both annotated following the MUMIN annotation model, but the granularity and the amount of features used differ slightly.

The results show that although the classifiers do better than the majority baseline, their accuracy is significantly lower than is the case when training and test data come from the same corpus. In other words, although some knowledge can be transferred across corpora, the outcome of the experiment also shows that different communicative situations and settings have an impact on the type of multimodal behaviour produced.

In spite of the fact that the NOMCO corpus is not large, and the shape annotations of the gestures relatively coarse-grained, the results of our machine learning experiments show that the annotations can be used to model and predict a number of phenomena with levels of accuracy in some cases well above the majority baseline. Therefore, models trained on the NOMCO corpus could be used for semi-automatic annotation of the functions of gestural behaviour in new data from a similar communicative situation, provided that relevant shape features are available.

In general, the lesson learnt from the studies summarised above is that the automatic analysis of face-to-face conversations cannot abstract away from the important role played by gestural behaviour, and that the type of annotation we have adopted in the NOMCO project provides a representation of the communicative function of multimodal behaviour that is adequate for automatic analysis.

5 Conclusions and future directions

We hope in this article to have given a thorough description of the way in which the NOMCO corpus was collected and annotated, the information it contains, and the way it has been used to analyse how speech and gestural behaviour interact in the expression of communicative phenomena such as focusing, feedback, and turn management. The size of the corpus is limited if we compare with the numbers we are used to from language corpora. However, even though research on multimodal behaviour goes back many decades, the area of multimodal corpus development is still relatively new (the first LREC workshop on multimodal corpora was held in 2000), and consequently annotation methods are far from being standardised. Furthermore, all the annotation in NOMCO was produced manually, and was therefore time-consuming. In future, automatic methods based on image processing are likely to make the annotation of gestures, at least for what concerns gesture shape and dynamics, easier, and the insight gained in NOMCO on how form and function of gestures interact will hopefully contribute to faster annotation of the functional level, too.

In spite of the fact that the NOMCO data were analysed from many different perspectives, there is still room for future development. First of all, hand gestures, movements of limbs and gaze are still to be annotated. Secondly, deeper levels of prosodic and linguistic analyses can be added to allow for other types of analysis, e.g. how content is expressed in speech and representational gestures.

Acknowledgments The NOMCO project was funded by NOS-HS NORDCORP. We would like to acknowledge our partners from the Universities of Gothenburg and Helsinki, the annotators of the Danish data Sara Andersen, Josephine B. Arrild, Anette Studsgård and Bjørn N. Wesseltolvig. We would also like to thank the two anonymous reviewers for their helpful comments.

Appendix

See Table 20.

Table 20 Gesture counts

Gesture	Sum
Nod	746
Tilt	496
SideTurn	437
HeadForward	357
Shake	337
HeadBackward	264
HeadOther	212
Jerk	180
Waggle	88
Total head movements	3117
Smile	667
Laughter	217
FaceOther	92
Scowl	5
Total face	981
Raise	471
Frown	117
BrowsOther	4
Total brows	592
BodyDirectionOther	222
BodyTurn	158
BodyBackward	145
BodyForward	132
BodySide	132
BodyUp	77
BodyDown	22
Total body	888
Shrug	107
ShouldersOther	49
Total shoulders	156

Table 20 displays sums of the various gesture types in the corpus. Note that the total number of facial expressions is in fact 1448: to the 981 expressions that are annotated with one of the general facial features, must be added 467 expressions that are only annotated with a feature related to the eyebrows. Conversely, there 856 facial expressions with no eyebrow annotation. Similarly for body posture, there are 982 behaviours in total: to the 888 movements annotated with a body posture feature must be added 94 shoulder movements with not body posture annotation, while there are 826 body posture annotations not associated with a shoulder movement.

References

Alahverdzhieva, K., Lascarides, A. (2010). Analysing speech and co-speech gesture in constraint-based grammars. In S. Müller (Ed.), *Proceedings of the HPSG10 conference* (pp. 6–26). Stanford: CSLI Publications.

Allwood, J. (2002). Bodily communication dimensions of expression and content. In B. Granström, D. House, & I. Karlsson (Eds.), *Multimodality in language and speech systems* (pp. 7–26). Dordrecht: Springer. doi: 10.1007/978-94-017-2367-1_2.

Allwood, J. (2008). Dimensions of embodied communication—Towards a typology of embodied communication. In I Wachsmuth, M. Lenzen & G. Knoblich (Eds.), *Embodied communication in humans and machines*. Oxford: Oxford University Press.

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. In Martin JC, Paggio P, Kuehnlein P, Stiefelhagen R, Pianesi F (Eds.), *Multimodal corpora for modelling human multimodal behaviour, special issue of the international journal of language resources and evaluation* (Vol. 41, pp. 273–287). Berlin: Springer.

Allwood, J., Lanzini, S., & Ahlsén, E. (2014). Contributions of different modalities to the attribution of affective-epistemic states. In P. Paggio & B. N. Wessel-Tolvig (Eds.), *Proceedings from the 1st European symposium on multimodal communication University of Malta* (pp. 1–6). Valletta: Linköping University Electronic Press.

Allwood, J., Nivre, J., & Ahlsén, E. (1993). On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, *9*(1), 1–26.

Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. Cambridge: Cambridge University Press.

Aung, M. S. H., Bianchi-Berthouze, N., Watson, P., & Williams, A. C. D. C. (2014). Automatic recognition of fear-avoidance behaviour in chronic pain physical rehabilitation. In *Proceedings of 8th international conference on pervasive computing tehcologies for healthcare*.

Boersma, P., & Weenink, D. (2009). Praat: Doing phonetics by computer (version 5.1.05) [computer program]. Retrieved May 1, 2009. From http://www.praat.org/.

Bolinger, D. (1986). *Intonation and its parts: Melody in spoken English*. Stanford, CA: Stanford.

Bourbakis, N., Esposito, A., & Kavraki, D. (2011). Extracting and associating meta-features for understanding people's emotional behaviour: Face and speech. *Journal of Cognitive Computation*, *3*, 436–448.

Bunt, H., Alexandersson, J., Choe, J. W., Fang, A. C., Hasida, K., Petukhova, V., et al. (2012). Iso 24617-2: A semantically-based standard for dialogue annotation. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *LREC, Citeseer* (pp. 430–437). European Language Resources Association (ELRA).

Campbell, N., & Scherer, S. (2010). Comparing measures of synchrony and alignment in dialogue speech timing with respect to turn-taking activity. In *Proceedings of Iiterspeech* (pp. 2546–2549).

Cavicchio, F., & Poesio, M. (2009). Multimodal corpora annotation: Validation methods to assess coding scheme reliability. In M. Kipp, J. C. Martin, P. Paggio, & D. Heyen (Eds.), *Multimodal corpora. Lecture notes in computer science* (Vol. 5509). Berlin: Springer.

Cerrato, L. (2007). *Investigating communicative feedback phenomena across languages and modalities*. Ph.D. thesis, School of Speech and Music Communication, Stockholm, KT.

Cienki, A., & Müller, C. (2008). *Metaphor and gesture*. Amsterdam: Benjamins.

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

Dancey, C. P., & Reidy, J. (2004). *Statistics without maths for psychology: Using spss for windows*. Upper Saddle River, NJ: Prentice-Hall Inc.

De Ruiter, J. P. (2000). The production of gesture and speech. In D. McNeill (Ed.), *Language and gesture*. Cambridge: Cambridge University Press.

Duncan Jr., S., & Fiske, D. (1977). *Face-to-face interaction*. Hillsdale, NJ: Erlbaum.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, *23*(2), 283–292.

Duncan, S., Cassell, J., & Levy, E. (2007). *Gesture and the dynamic dimension of language*. Amsterdam: Benjamins.

Ebert, C., Evert, S., & Wilmes, K. (2011). Focus marking via gestures. In I. Reich et al. (Eds.), *Proceedings of Sinn & Bedeutung 15* (pp. 193–208). Saarbrücken, Germany: Universaar-Saarland University Press.

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

Ekman, P., & Friesen, W. (1975). *Unmasking the face: A guide to recognizing emotions from facial clues*. Upper Saddle River: Prentice-Hall.

Ekman, P., & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, *1*(1), 49–98.

Enfield, N. J. (2012). *TThe anatomy of meaning: Speech, gesture, and composite utterances*. Cambridge: Cambridge University Press.

Gibbon, D. (2011). Modelling gesture as speech: A linguistic approach. *Poznań Studies in Contemporary Linguistics*, *47*, 470–508.

Giorgolo, G., & Verstraten, F. A. (2008). Perception of 'speech-and-gesture' integration. In *Proceedings of the international conference on auditory-visual speech processing 2008* (pp. 31–36).

Goodwin, C. (1981). *Conversational organization: Interaction between speakers and hearers*. New York: Academic Press.

Gullberg, M., & de Bot, K. (Eds.). (2010). *Gestures in language development*. Amsterdam: Benjamins.

Hadar, U., Steiner, T., & Rose, F. C. (1984). The timing of shifts of head postures during conversation. *Human Movement Science*, *3*(3), 237–245.

Hadar, U., Steiner, T. J., & Rose, F. C. (1985). Head movement during listening turns in conversation. *Journal of Nonverbal Behavior*, *9*(4), 214–228.

Jongejan, B. (2010). Automatic face tracking in anvil. In M. Kipp, J. C. Martin, P. Paggio, & D. Heylen (Eds.), *Multimodal corpora: Advances in capturing, coding and analyzing multimodality* (pp. 201–208). European Language Resources Association (ELRA), May 18, 2010.

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

Kendon, A. (1978). Differential perception and attentional frame: Two problems for investigation. *Semiotica*, *24*, 305–315.

Kendon, A. (1980). Gesture and speech: Two aspects of the process of utterance. In M. R. Key (Ed.), *Nonverbal communication and language* (pp. 207–227). Mouton.

Kendon, A. (2004). *Gesture*. Cambridge: Cambridge University Press.

Kipp, M. (2004). *Gesture generation by Imitation—From human behavior to computer character animation*. Boca Raton, FL: Dissertation.com.

Kipp, M., & Martin, J. C. (2009). Gesture and emotion: Can basic gestural form features discriminate emotions? In *Proceedings of the international conference on affective computing and intelligent interaction (ACII-09)*. IEEE Press.

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

Kousidis, S., Malisz, Z., Wagner, P., & Schlangen, D. (2013). *2013*. Exploring annotation of head gesture forms in spontaneous human interaction. In *Proceedings of the Tilburg gesture meeting (TiGeR)*.

Leonard, T., & Cummins, F. (2010). The temporal relation between beat gestures and speech. *Language and Cognitive Processes*, *26*(10), 1457–1471.

Levinson, S. (1983). *Pragmmatics*. Cambridge: Cambridge University Press.

Loehr, D. P. (2004). *Gesture and intonation*. Ph.D. thesis, Georgetown University.

Loehr, D. P. (2007). Aspects of rhythm in gesture and speech. *Gesture*, *7*(2), 179–214.

Lucey, P., Cohn, J. F., Prkachin, K. M., Solomon, P. E., Chew. S., & Matthews, I. (2012). Painful monitoring: Automatic pain monitoring using the UNBC-McMaster shoulder pain expression archive database. *Image and Vision Computing, 30*(3), 197–205.

Maynard, S. K. (1987). Interactional functions of a nonverbal sign: Head movement in Japanese dyadic casual conversation. *Journal of Pragmatics*, *11*, 589–606.

McClave, E. Z. (2000). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, *32*(7), 855–878.

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

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

Navarretta, C. (2011). Annotating non-verbal behaviours in informal interactions. In I. A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud, & A. Nijholt (Eds.) *Analysis of verbal and nonverbal communication and enactment: The processing issues, LNCS* (Vol. 6800, pp. 317–324). Berlin: Springer.

Navarretta, C. (2012). Annotating and analyzing emotions in a corpus of first encounters. In IEEE (Ed.) *Proceedings of the 3rd IEEE international conference on cognitive infocommunications* (pp. 433–438), Kosice.

Navarretta, C. (2013a). Predicting speech overlaps from speech tokens and co-occurring body behaviours in dyadic conversations. In *Proceedings of ACM international conference on multimodal interaction (ICMI 2013)* (pp. 157–163). Sidney: ACM.

Navarretta, C. (2013b). Transfer learning in multimodal corpora. In IEEE (Ed.) *Proceedings of the 4th IEEE international conference on cognitive infocommunications (CogInfoCom2013)* (pp. 195–200). Hungary: Budapest.

Navarretta, C. (2014). Predicting emotions in facial expressions from the annotations in naturally occurring first encounters. *Knowledge Based Systems*, *71*, 34–40.

Navarretta, C., Ahlsén, E., Allwood, J., Jokinen, K., & Paggio, P. (2012). Feedback in Nordic first-encounters: A comparative study (pp. 2494–2499). Istanbul: European language resources distribution agency.

Navarretta, C., & Paggio, P. (2012). Verbal and non-verbal feedback in different types of interactions. In *Proceedings of LREC 2012* (pp. 2338–2342). Istanbul.

Navarretta, C., & Paggio, P. (2013a). Classifying multimodal turn management in Danish dyadic first encounters. In *NEALT proceedings of the 19th nordic conference of computational linguistics (Nodalida 2013), Oslo, Linköping electronic conference proceedings* (pp. 133–146).

Navarretta, C., & Paggio, P. (2013b). Multimodal turn management in Danish dyadic first encounters. In *NEALT proceedings. Northern European association for language and technology, Proceedings of the fourth nordic symposium of multimodal communication, Göthenburg, Linköping electronic conference proceedings* (pp. 5–12).

Paggio, P. (2006a). Annotating information structure in a corpus of spoken Danish. In *Proceedings of the 5th international conference on Language Resources and Evaluation LREC2006* (pp. 1606–1609). Italy: Genova.

Paggio, P. (2006b). Information structure and pauses in a corpus of spoken Danish. In *Conference companion of the 11th conference of the European chapter of the association for computational linguistics* (pp. 191–194). Italy: Trento.

Paggio, P. (2016). Coordination of head movements and speech in first encounter dialogues. In E. Gilmartin, L. Cerrato, & N. Campbell (Eds.), *Proceedings from the 3rd European Symposium on Multimodal Communication, Dublin, September* (pp. 69–74). Linköpings universitet: Linköping University Electronic Press.

Paggio, P., Allwood, J., Ahlsén, E., Jokinen, K., & Navarretta, C. (2010). The NOMCO multimodal nordic resource—Goals and characteristics. In *Proceedings of the seventh conference on international language resources and evaluation (LREC'10)*. European Language Resources Association (ELRA), Valletta.

Paggio, P., & Diderichsen, P. (2010). Information structure and communicative functions in spoken and multimodal data. In P.J. Henriksen (Ed.), *Linguistic theory and raw sound, Copenhagen studies in language* (Vol. 49, pp. 149–168). Frederiksberg: Samfundslitteratur.

Paggio, P., & Navarretta, C. (2011). Head Movements, facial expressions and feedback in Danish first encounters interactions: A culture-specific analysis. In *Lecture notes in computer science* (Vol. 6766, pp. 583–590). Springer.

Paggio, P., & Navarretta, C. (2012). Classifying the feedback function of head movements and face expressions. In *LREC 2012 workshop multimodal corpora—How should multimodal corpora deal with the situation?* (pp. 34–37). Istanbul: European language resources distribution agency.

Paggio, P., & Vella, A. (2014). Overlaps in maltese conversational and task oriented dialogues. In P. Paggio & B. N. Wessel-Tolvig (Eds.), *Proceedings from the 1st European symposium on multimodal communication University of Malta* (pp. 55–64). Valletta: Linköping University Electronic Press.

Peirce, C. S. (1931). *Elements of logic. Collected papers of Charles sanders peirce* (Vol. 2). Cambridge: Harvard University Press.

Poggi, I. (2007). *Hands, mind, face and body: A goal and belief view of multimodal communication*. Berlin: Weidler.

Russell, J. A., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, *11*, 273–294.

Savva, N., Scarinzi, A., & Bianchi-Berthouze, N. (2012). Continuous recognition of player's affective body expression as dynamic quality of aesthetic experience. *IEEE Transactions on Computational Intelligence and AI in Games*, *4*(3), 199–212.

Schegloff, E. A. (1984). On some gestures' relation to talk. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action* (pp. 266–298). Cambridge: Cambridge University Press.

Studsgård, A. L., & Navarretta, C. (2013). Annotating attitudes in the Danish NOMCO corpus of first encounters. In *NEALT proceedings. Northern European association for language and technology, 4th Nordic symposium on multimodal communication* (pp. 85–89). Linköping University Electronic Press.

Vallduví, E., & Engdahl, E. (1996). The linguistic realisation of information packaging. *Linguistics*, *34*(3), 459–520.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd edn.). San Francisco: Morgan Kaufmann.