Gefördert vom

bmb+f

Bundesministerium für
Bildung, Wissenschaft,
Forschung und Technologie

Mensch-Technik-Interaktion

SmartKom

Labeling of Gestures in SmartKom – Concept of the Coding System

Silke Steininger

Ludwig–Maximilians–Universität München

Report Nr. 2
März 2001

März 2001

Silke Steininger

Ludwig Maximilans Universität München
Schellingstr. 3
80799 München

Tel.: (089) 2180−5751
FAX: (089) 2180−99−5751

E−Mail: kstein@phonetik.uni−muenchen.de

Dieser Report gehört zu Teilprojekt 1: Modalitätsspezifische Analysatoren

Das diesem Technischen Dokument zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01 IL 905 gefördert. Die Verantwortung für den Inhalt liegt bei den Autoren.

Inhaltsverzeichnis

1 Introduction

This paper will be presented at the Gesture Workshop 2001 in London.

The SmartKom project is concerned with the development of an intelligent computer–user interface that allows a user to communicate almost naturally with an adaptive and self–explanatory dialogue system. Among other things the system will be able to analyze the gestural input of the user. To train a gesture analyzer, data is required, preferably realistic data. One of the tasks of our institute in the project is the collection and annotation of such data. Since the machine does not yet exist the data collection is done with help of so called Wizard of Oz–experiments: The system is simulated by humans (the "wizards") and the subjects are made believe that they interact with an existing machine. We record the subjects (video and audio) as they solve short tasks. The recordings are labeled off–line with respect to the gestures that the subjects used.

This contribution is concerned with our concept for the coding of the gestures. The presented concept is the first step in the development of a gesture coding system specifically designed for the description of communicative and non–communicative gestures that typically show up during a human–machine dialogue. We decided to use a functional level of description (instead of a morphological one) because we think that this level allows us to highlight the most interesting aspects of gestural human–machine interaction. Additional emphasis is put on the practicality of the system. Since it will be used for applied research, it must be relatively fast and easy to use.

After giving a short overview over the SmartKom project, we will describe the practical and theoretical principles that served as the basis for the development process. We will then outline the heuristic analysis of a small corpus of dialogues during which we generated the label definitions. After that the labels will be described in detail. We will conclude with a short overview over the open points and planned future work and a comparison of our concept with the well–known taxonomy for gestures of Ekman (1999).

2 The SmartKom Project[1]

The goal of the SmartKom project (started in January 2000) is the development of an intelligent computer–user interface that allows a computer novice as well as an expert to communicate almost naturally with an adaptive and self–explanatory dialogue system. New possibilities of the interaction between human and machine are investigated: The system does not only allow input in the form of natural speech but also in the form of gestures. Additionally the emotional state of the user is analyzed via his/her facial expression and prosody of speech. The output of the system comprises a graphic user interface and synthesized language. The graphic output is realized as a computer screen that is projected onto a graph tablet.

To explore how users interact with a machine that has far greater communication skills than the machines we are used to at the moment, data is collected in so–called Wizard–of–Oz experiments: The subjects have to solve certain tasks with the help of the system (like planning a trip to the cinema, programming a VCR or sending an e–mail). They are made believe that the system they interact with is already fully functional. Actually many functions are only simulated by two "wizards" that control the system from another room. The different functionalities of the system are developed by different partners of the project. The Institute of Phonetics and Speech Communication in Munich is responsible for the collection and annotation of the multimodal data and the evaluation of the system.

2.1 Collection Of Multimodal Data

In the first phase of the project the collected data is needed for three different main purposes:

1. The training of speech, gesture and emotion recognizers.

1 http://smartkom.dfki.de/index.html

2. The development of user–, language–, dialogue–models etc. and of a speech synthesis module.
3. The general evaluation of the behavior of the subjects in the interaction with the machine.

In each Wizard–of–Oz session spontaneous speech, facial expression and gestures of the subjects are recorded.

For the **audio** recordings we use:

- a microphone array (4)
- a directional microphone and
- (alternating) a headset or a clip–on–microphone.

Video:

- A digital camera captures the *facial expression* of the subjects.
- The gestures are recorded with a second digital camera which captures a *side view of the subject* (hip to head)
- and an infrared sensitive camera (from a *gesture recognizer*: SIVIT/Siemens) which captures the hand gestures (2–dimensional) in the plane of the graphical output.

Other:

- The coordinates of pointing gestures on the work space are recorded (with SIVIT),
- as well as the inputs of a pen on the graph tablet.
- Additionally the output of the beamer is logged into a slow frame video stream.

Each subject is recorded in two sessions of about 4.5 minutes length each. The recordings that are relevant for the gesture coding are: the 2–dimensional black–and white picture from the infrared camera, the sideview of the person and the output of the beamer (view of the display). To facilitate the coding, the beamer output is overlaid with the stream of the infrared sensitive camera (see figure 1). In this way the coders have a sideview, a view from above and information on which area of the display the hand was directed to[2].

Figure 1: The 4–view–videostream for the coding of the gestures.

2 The front view is added too, but mainly for the labeling of the facial expression. For the gesture coding it does only play a

The recordings are done directly on harddisk. They are synchronized manually and copied into a QuickTime frame which is then distributed to the project partners for further analysis. For the gestural coding the rogram Interact[3] is used.

3 Basis of the Development Process

3.1 Practical requirements

The coded gestures in the project are needed for the training of a gesture recognizer, as well as for the development of a model which is able to to predict typical human–machine interactions. To achieve this the behavior of the subjects while interacting with the machine has to be examined. It is of interest which gestures are used in such a context and in which way[4]. The coding system therefore has to provide training material for the gesture analyzer, but it also has to mark gestural episodes that are perhaps too complex to be used as training material (at the moment) but are interesting with respect to the nonverbal communicative behavior in general.

The following requirements are the result of these considerations:

1. **The labels should refer to the functional level, not the morphological level.**

For theoretical reasons we want to use a functional coding system (see below). However, the decision is also made for practical reasons since the coding of the morphological form of gestures ("phenomenal" systems as well as "structural" classification systems) and body movements is exceedingly time consuming. This partly follows from the fact that phenomenal or morphological coding requires a huge amount of labels if one wants to describe the movements exhaustively. Structural coding systems are time consuming as well because they further subdivide each gesture into several phases (Wallbott, 1982). The labelers therefore have to be trained carefully (e.g. Ekman & Friesen, 1978) or direct, objective methods like SELSPOT or OPTOTRAK should be used (e.g. Levelt et al. 1985;). We neither have the required time for the former nor the resources for the latter, so we opted for functional codes[5].

2. **The labels should be selective.**

Functional codes (as indirect measurements) are not as exact as direct methods, therefore exceptional care has to be taken to find labels that are well–defined, easy to observe and unproblematic to discriminate by means of objective (communicable) criteria.

3. **The coding system should be fast and easy to use. It should be easy to understand even for non–experts.**

4. **The resulting label file should facilitate automatic processing (a consistent file structure, consistent coding, non–ambiguous symbols, ASCII, parsability). Preferably it should be easy to read[6].**

The data collection is done primarily for our partners in the project who develop the actual modules of the system. We have to conform to their requirements: they need the data fast and in an easily accessible format.

5. **The main labels and most modifiers should be realized as codes not as annotations in order to heighten consistency.**

marginal role.

3 http://www.mangold.de/

4 The collection of such data is a difficult task in itself, see the demonstration "Evoking Gestures in SmartKom – Design of the Graphic User Interface" (Beringer) at this workshop.

5 "Functional code" or "functional unit" is sometimes defined differently by different people. We use the term in accordance with Faßnacht (1979) for a unit that is defined with regard to its effect or its context.

6 Many of the practical criteria were adopted from the transliteration conventions for speech in SmartKom, see Beringer et al. (2000).

Annotations (free comments and descriptions that don’t follow as strict rule) are more flexible, but codes (predefined labels from a fixed set) increase the conformity between labelers. During the development of the system some parts of the labels will be realized as annotations (to collect data about possible variations), but in the end most of these annotations will be transformed into codes.

3.2 Theoretical considerations

We assume that existing taxonomies are not very well suited for our task of labeling human–*machine* interaction because they were developed with respect to human–*human* interaction (e.g. Efron, 1941; Bales, 1970; McGrew, 1972; Ekman & Friesen, 1972; Wallbott, 1982; McNeill, 1992). Others are too specific for our need of a usable, relatively fast system (for example methods for transcribing American Sign Language – see SignStream[7] or methods for transcribing dance: Laban, 1956). But how should a system look like that is exactly tailored to our context?

Every coding system is a kind of filter: Some aspects of the observable "reality" are highlighted, others are concealed. The image of the filter emphasizes the fact that each coding system includes expectations concerning the events to be labeled. A good coding system is a great help for objective labeling but it is *not* objective in itself: The labels define what will be observed. Therefore it is important to use the right filter for your needs. In order to be able to study the nonverbal aspects of human–machine interaction one needs labels that correspond to units that have meaning in a (human–machine) dialogue. It can be expected that gestures are used differently when the dialogue partner is a machine that communicates with synthesized speech and graphic output – similar to the changes that take place in speech in such a context (Jönsson et al, 1988). For example a lot more deictic gestures can be expected and it is not clear if other illustrative gestures will show up at all.

Because the function or meaning of the gesture within the communication process is the interesting point in our context, we decided to define labels on a functional level[8]. Functional systems emphasize the connection with speech (Wallbott, 1982) which is of special interest to us. Additionally, a morphological classification system (with a great amount of categories) is ill suited to allow statistical analyses and inter–individual comparisons (Ekman & Friesen, 1972). Most importantly, with morphological labels it is not possible to signify the difference between communicative and non–communicative gestures.

The disadvantage of a functional coding system is that more interpretation is required in the coding process than in a morphological system. We don’t consider this restriction problematic because the frame of interpretation for the coding is the same one that has to be used by the gesture analyzer/the system: Which gestural movements are meaningful to me with regard to the interaction with my dialogue partner? The mistakes which are likely to occur during labeling resemble the one’s a human communication partner makes and therefore are more acceptable than other mistakes.

In order to still be able to catch some of the morphological information we decided to include morphological *modifiers*. They roughly describe the form, duration and stroke of the gesture. By including some aspects of structural or "micro" coding systems we hopefully can work around some of the disadvantages of a purely functional system. But since the morphological descriptions are not the main criterion in defining the labels, we benefit from the fact that the labels highlight the point which is most interesting for us: The intent with which the gesture is performed.

4 Definition of the Coding Concept

To match our practical needs and theoretical considerations with the context that we have to deal with (a multimodal human–machine dialogue), we did not start with explicit definitions or an existing taxonomy for gestures, but with a heuristic analysis of a number of recordings of human–machine dialogues. Such a free, unsystematic observation can lead to valuable insights for the

7 http://web.bu.edu/asllrp/SignStream/

8 Well–known functional classifications are from Ekman and Friesen (1972; the newest version being described in Ekman, 1999) and from McNeill (1992). Older systems are from Efron (1941), Krout (1935), Rosenfeld (1966), Mahl (1968), Freedman et al. (1973) and Sainsbury & Wood (1977).

definition of a system that allows systematic observations (Faßnacht, 1979). The recordings that served as the basis for the analysis are described below.

No definite taxonomy was used as a basis for the analysis, but the theoretical frame outlined below. Two observers[9] separately listed every episode that they could identify as a "functional unit". Then they tried to assign a label to the unit with respect to the obvious meaning of the gesture ("meaning" with respect to the communication process). For obvious reasons, it was not attempted to ascertain the true intent with which the gesture was performed but the intent that could be identified by an observer. Identifiable units with no obvious meaning for the communication process were listed as well. The underlying questions for the list of units were:

- Which functional/non–functional units (with respect to the dialogue) can be identified (e.g. "pointing", "no", "back", "non functional" etc.)?
- What is the best way to categorize the observed units? (e.g. are there different "pointing" gestures? Which similar gestures can be combined?)
- What is the best way to describe the units?
- Is there a reference on the display – if so which one (e.g. "Button X", "Region Y", "nothing")?
- Which broad morphological form does the unit have (e.g. "one hand/finger pointing", "one hand circling", "two hand crossing")?
- How can we define beginning and end of a given unit?
- Is there a reference word in the audio channel and if so which one (e.g. "this", "here", "no")?

The observers made their judgments independently. After completing the list, the identified gestures were discussed with the author. Every label that could not be operationalized satisfactorily was removed. Similar labels were combined. Three broad categories emerged, each with several sub–categories, called labels. After this some additional labels were added that had not been observed but were thought probable to appear or had to be added in order to complete the categories. Before giving a detailed description of how the labels have been defined so far, we will describe the data that served as the basis for the analysis.

4.1 The Empirical Basis

The Subjects

The data we used consisted of 70 sessions of about 4.5 min length each. 35 voluntary, naive subjects were recorded and paid a small recompense. 18 subjects were female, 17 male. They were between the age of 19 and 60. Occupations: 24 students (including 2 PhD students, 1 pupil) and 11 employed persons (including 1 retired person).

The Task

A full account of the Wizard–of–Oz procedure cannot be given here, it will be published separately[10]. Instead we give a short overview of the task and the performance of the subjects. The subjects were carefully instructed. They were told that they had to test a new prototype of a dialogue system which could understand spoken language as well as gestures. It was not shown to them what sort of gestures, it was only pointed out that the system understood movements of the hand which were performed on or above the display. They were encouraged to try for themselves what the system could understand and what it could not. A detailed description of the functionality of the system was *not* given. It was emphasized that experimentation was appreciated.

The subjects had to imagine themselves being in Heidelberg (a German town) and using a new information booth for the first time. They were told to "Plan a trip to the cinema this evening". The subjects were told that they should try to solve the task but were encouraged to try different things as well.
After the instruction two sessions were recorded (with a short break in between). Afterwards the

9 One of the observers was trained in observing and judging movements (a dancing master) the other one had no previous experience in the field (a student of German and English).

10 For information on the graphic user interface see the demonstration "Evoking Gestures in SmartKom – Design of the Graphic User Interface" (Beringer) at this workshop.

subjects were led into a different room, where they were interviewed. They were for example asked about problems, what they had liked and disliked about the system, if it seemed more like a computer or a machine to them etc.

No subject reported or showed knowledge of the fact that the system was not real. Most subjects thought that the use of the system was easy and almost everyone reported that it was fun to use the system. Some communication problems arose but in most cases the interaction went smoothly (the wizards sometimes built in errors on purpose). Most subjects stated that they did not have any problems with the test situation because the task was very interesting, some even showed fascination ("Wow – he really understood me!"). About half of the subjects reported that they tried a few things out. About two third thought the system was more like a machine, the rest judged it to be either something in between or rather more human. Even of the latter subjects none suspected it actually was human.

5 The Coding Concept

In the following the concept that emerged from the qualitative analysis of the data and the first refining step will be described. It serves as the basis for the second step of the development process: The actual labeling of a set of dialogues and the calculation of quality measures. It can be expected that after the second step some other labels will be added, existing labels will be refined and some modifiers will be discarded.

5.1 Overview

A *label*, that belongs to one of three superordinate *categories*, is assigned to each identified gesture. The label is complemented by several *modifiers*. Three modifiers refer to aspects of time (*beginning*, *end*, *stroke*), while three refer to aspects of content (*morphology*, *reference word*, *reference object*). If necessary an identified gesture is specified by a *comment*. Beginning and end are marked as points in time, the stroke is marked as a period. Explanatory notes can be added (see below). The reference word is an annotation, i.e. the word in the audio channel that corresponds to the gesture is quoted. Additionally it is marked whether the reference word showed up before, simultaneously or after the gesture. The modifiers "morphology" and "reference object" are planned as codes (assignment to a closed set of descriptions). In the first test of the concept they will probably be realized as annotations (open descriptions).

5.2 Limits of the coding conventions

The coding conventions for SmartKom are especially designed for the labeling of gestures during the human–machine dialogue. The term "gesture" normally denominates segments of the stream of movements with the arms and hands. The segments are sorted into different categories according to the respective underlying description system. The length of the segment that someone identifies as a single gesture and the chosen category are arbitrary in the sense that they don't result from observation, but conform to the theoretical foundation (such a theoretical foundation can also be the common understanding of what a "gesture" is). For a more complete discussion of the problem of the definition of a unit see Faßnacht (1979).

Our units derive from a functional definition with three broad categories: Is the gesture an interaction with the system (an "interactional gesture"), an "interaction with oneself" (a "supporting gesture") or something else (a "residual gesture")? The only (functional and supporting) gestures that are labeled are the ones that enter the so called "cubus", the field of the display and the room above the display where the SIVIT gesture recognizer records data (it corresponds roughly to the border of the display)[11]. Additionally, gestures that are interesting (but take place outside the »cubus«) are coded in the category of residual gestures.

5.3 Definitions

Each gesture is assigned to one of the three following categories: **I**nteractional gesture (I−gesture), s**U**pporting gesture (U−gesture)[12] and **R**esidual gesture (R−gesture). The criterion for this assignment is the *intention* of the subject. However, the goal of the labeling process is not to label the "true" intention of the subject. We think it is quite sufficient to retrieve the intention that gets through to the communication partner (in this case the system). The question that has to be answered is therefore "What does the user want to achieve with the gesture *obviously*?" and *not* "What does the user want to achieve with the gesture *actually*?".

Interactional Gesture

The interactional gesture is constructive, i.e. it is (possibly together with the verbal output) the means of the interaction with the computer. When a subject uses a hand/arm−movement to give a command to the computer this gesture is called interactional. A command is any request to the system. If the computer cannot fulfill the command, the requesting gesture still remains an interactional gesture (functionality is defined with regard to the communication process not with regard to the effectiveness in the use of the system). A second type of interactional gesture is the confirmation of a question from the system.

Supporting Gesture

Like the functional gesture, the supporting gesture is constructive. It occurs in the phase when a request is prepared. It signifies the gestural support of a "solo−action" of the user (like reading or searching), e.g. an action that is not an interaction with the system (interaction in the sense of communication). A supporting gesture does not end with a command. If a supporting gesture is followed by a command, the command is labeled separately as a interactional gesture.

Before a gesture is assigned to the category "supporting gesture" and to distinguish it from an "interactive gesture" at least one (or both) of the following prerequisites has to be confirmed as true:

a) The gesture is accompanied by verbal comments from the user that make the preparational character of the underlying cognitive process obvious (e.g. "hm ... where is... here?... no...")

b) The gesture is accompanied by a facial expression that makes the preparational character of the underlying cognitive process obvious (like lip movements, searching eye movements, frowning).

In order to be able to differentiate a supporting gesture from a residual gesture at least one (or both) of the following two prerequisites have to be given:

a) A linkage between gesture and speech (both streams refer to the same topic).
b) An identifiable focus, i.e. gaze and gesture fall on the same spot.

Residual Gesture

This category subsumes all gestures that take place within the cubus, but do not belong to one of the above categories. The few of the labeled gestures that take place outside of the cubus belong to this category too.
A residual gesture is not constructive. It does not prepare a request (at least not obviously) and is not a request or confirmation. A residual gesture is either an emotional gesture or an unidentifiable gesture.

When an unidentifiable gesture is given, there exists

a) no linkage between gesture and speech

---

11 In reality the "cubus" has the form of a pyramid. For practical purposes of the coding this can be ignored however.

12 We called the supporting gesture U−gesture for reasons of consistency with the originally German name "Unterstützende Geste".

b) no identifiable focus, i.e. gaze and gesture don't fall on the same spot.

Label

The label signifies the exact kind of the gesture, i.e. it refers to the function of the gesture within the communication process. To facilitate the assignment, the name contains one of the prefixes F−, U− and R−, which stand for the three categories.

The following F−labels exist:

- F−circle (+)
- F−circle (−)
- F−point (long +)
- F−point (long −)
- F−point (short +)
- F−point (short −)
- F−free (free)

The U−labels:

- U−circle (read)
- U−circle (search)
- U−circle (count)
- U−circle (ponder)
- U−point (read)
- U−point (ponder)

The R−labels:

- R−emotional (+ cubus)
- R−emotional (− cubus)
- R−unidentifiable (+ cubus)

Morphology

The term "morphology" here signifies the modifier that subdivides the label even more precisely with regard to the manner of performance. A pointing gesture for example can be performed with the left or the right hand, a circling can be done with one finger or with the whole hand. The modifier therefore shortly describes with which body part the movement was performed and in which way. At the moment the modifier is realized as an annotation, but it is planned to develop a list of codes (a list of standard descriptions). The description always refers to the stroke of the movement, i.e. not the whole movement is described, only its most important part.

Reference Word

This modifier signifies the word that is spoken in correspondence to the gesture. It is known that "hand gestures co−occur with their semantically parallel linguistic units" (McNeill, 1992). "Correspondence" therefore means that the word occurs shortly before, simultaneously or shortly after the gesture and gesture and word are linked with regard to the content. Of course the reference word can be "none" (i.e. non existing).

Reference Object

This modifier signifies the object on or the region of the display to which the gesture refers. At the moment the modifier is realized as an annotation, but it is planned to develop a list of codes (standard objects/regions). The modifier can comprise one specific object/region, the "whole display" or "none".

Stroke

This modifier signifies the "culmination point" of the gesture, i.e. the "most energetic" or "most important" part of a gesture that is often aligned with the intonationally most prominent syllable of the accompanying speech segment (Kendon, 1980; McNeill, 1992). Its beginning and end are labeled. If necessary with optional comment ("unclear").

Beginning

The beginning of a gesture is determined with the help of the following criteria:
– Hand enters cubus
– End of previous gesture
– Hand moves after pausing within the cubus for a certain time
A comment is added if the beginning was difficult to specify.

End

The end of a gesture is determined with the help of the following criteria:
– Hand leaves cubus
– Beginning of following gesture
– Hand stops moving within the cubus
A comment is added if the end was difficult to specify.

Comment

Comments about the stroke, beginning, end or peculiarities that cannot be noted anywhere else.

5.4 Descriptions of the labels

1. F−Label

F−circle (+)
A continuous movement with one hand that specifies an object on the display through the circling of this object. The display is touched.

Figure 2: An extended pointing gesture (F−point (long +)). Without context long and short pointing gestures cannot be distinguished.

Operationalization: The whole hand or parts of the hand circumscribe a certain region on the display fully or almost fully. It doesn't have to be a circular movement. It is always a well directed movement. At least during part of the gesture the gaze of the subject is aimed at the chosen region. Possibly the intent to select something becomes apparent from the words that are spoken.

F−circle (−)

Like F−circle (+), but the display is not touched.

F−point (long +)

A selective movement of one hand that specifies an object on the display, with the display being touched. The stroke of the movement (i.e. the pointing of the hand at the object) is of extended duration.

Operationalization: The whole hand or parts of the hand are pointed at a certain spot on the display. The movement is well−directed. At least during part of the gesture the gaze of the subject is aimed at the chosen region. Possibly the intent to select something becomes apparent from the words that are spoken.

F−point (long −)
As F−point (long +), but this time the display is not touched.

F−point (short +)

As F−point (long +), but the stroke of the movement (i.e. the period, during which the hand is pointed at the object) is not extended/is very short.

Figure 3: A pointing gesture during which the display is not touched (F−point (short −)).

F−point (short −)

As F−point (long +), but the stroke of the movement (i.e. the aiming of the hand at the object) is not extended/is very short and the display is not touched.

F−free

A movement of the hand/the hands which takes place above the display and signifies a wish or command of the user (for example to go to another page). This gesture can vary considerably with respect to its morphology. For example a "pageturn" can look like waving or the mimicking of turning a page. A "stop" can be a decisive waving with both hands. The display is not touched.

Operationalization: One or both hands make a movement above the display which can be identified as a unit . The movement obviously has a function in the communication process (the context makes the intent of the user obvious). The movement is always performed emphatically, never casually. The subject‘s gaze is directed towards the display.

2. U–Label

U–circle (read)

A continuous movement of the hand above or on the display. No object is specified. At least during a part of the movement the subject is obviously reading.

Operationalization: The whole hand or parts of the hand move continuously above or on the display, possibly sometimes pausing shortly in–between. At least during part of the gesture the gaze of the subject follows a textline. There is no circling of a region. The movement can be well–directed, but also tentative. The gaze of the subject is at least for a part of the gesture directed towards the text. At least during a part of the movement the subject obviously reads, i.e. the text is read aloud or the lips move.

U–circle (search)

A continuous movement of the hand above or on the display. No object is specified. It is obvious that the subject is looking for a certain object or region.

Operationalization: The whole hand or parts of the hand move continuously above or on the display, possibly sometimes pausing shortly in–between. The movement can be straight or curved. An interactional gesture can follow but this is not mandatory. There is no circling of a region. The movement can be well–directed, but also tentative. At least during part of the gesture the gaze of the subject is aimed at the display (towards the spot where the hand is placed). From the context it is obvious that the subject is looking for something, i.e. through the spoken words and/or a following interactional gesture (selection).

U–circle (count)

A continuous movement of the hand above or on the display. No object is specified (but it is possibly referred to objects). At least during a part of the movement the subject is obviously counting.

Operationalization: The whole hand or parts of the hand move continuously above or on the display, possibly sometimes pausing shortly in–between. The movement does not stop on the counted objects (a stop is counted as an interactional gesture – a selection). A full circling of the region does not take place. The movement can be well–directed, but also tentative. At least during part of the gesture the gaze of the subject is aimed at the spot where the hand is placed. At least during a part of the movement the subject obviously counts, i.e. numbers are spoken or formed with the lips.

U–circle (ponder)

A continuous movement of the hand above or on the display. No object is specified. The subject does not search, count or read, yet the focus of attention/the gaze for the most part follows the movement.

Operationalization: The whole hand or parts of the hand move continuously above or on the display, possibly sometimes pausing shortly in–between. The movement can be straight or curved. An interactional gesture can follow but this is not mandatory. The movement can be well–directed, but also tentatively. At least during part of the gesture the gaze of the subject is aimed at the display (on the spot where the hand is placed). From the context it is obvious that the subject is not searching, counting or reading. If a full circling does take place it is obvious from the context that the circled object is not selected (e.g. from a verbal comment, "should I choose this?").

U−point (read)
A selective movement of the hand, which is aimed at a text (a touch is possible, but not mandatory). The duration is short. It is obvious that the subject does not want to select the text. The subject obviously reads the text.

Operationalization: The whole hand or parts of the hand are aimed at a specific spot on the display (text). The movement is well−directed. The gaze of the subject rests on the text. It is obvious that the subject reads, i.e. the text is read aloud or the lips move. From the context it is obvious that the movement is not a selection, as it is verified by the spoken words of the subject.

U−point (ponder)
A selective movement of the hand to a spot on the display (a touch is possible, but not mandatory). The duration is short. It is obvious that the subject does not want to select the object.
Operationalization: The whole hand or parts of the hand are aimed at a specific part of the display. The movement is well−directed. The gaze of the subject rests on the object/region. The subject does not read. From the context it is obvious that the movement is not a selection, as it is verified by the spoken words of the subject.

*Figure 4: A pointing gesture during which the subject **doesn't** want to make a selection (U−point (ponder)). Without context, it cannot be distinguished from an interactional pointing gesture.*

R−Label R−emotional (cubus +)
A movement of one or both hands within the cubus with an obviously emotional content.

Operationalization: The movement is neither a functional nor a supporting gesture. The movement of the hand/the hands form a unit. The gesture conveys the impression of a certain emotion (anger, irritation, joy, amusement, surprise, puzzlement etc.).

R−emotional (cubus −)

Like R−emotional (cubus +), but outside of the cubus.

R−unidentifiable (cubus +)

A movement of one or both hands within the cubus that cannot be assigned to any of the other labels.

Figure 5: An emotional gesture (puzzlement).

5.5 Problems and future work

As noted the presented concept is the product of the first step in the development process. The next step is the evaluation and refinement of the concept. The following points have to be dealt with specifically:

- **Quality measures**: Quality measures have to be determined.
- **Time:** Up to now, time spans are defined only vaguely ("very short", "extended"). We have to examine which criteria are most useful for a stricter definition. Beyond that we have to evaluate how reliable the borders of the "functional units" can be determined by different coders.
- **Stroke:** It is not clear yet, how reliably a stroke can be identified.
- **Morphology**: A set of standard morphological descriptions has to be defined.
- **Reference object**: A set of standard reference objects has to be defined that is independent from the display used.
- **Reference word**: The definition of which word should be noted as a "reference word" has to be refined.
- **Additional labels**: The gestures that can be observed are heavily influenced by the display used[13]. Therefore additional labels will have to be added. We have to allow for possible adaptations of the system to new observations.
- **Practicality**: The concept described is probably still too complex to fulfill the need for a fast system. At the moment it is optimized for the development process in which great detail is of specific importance. For the practical use it has to be trimmed.

6 Relation to other taxonomies

We have mentioned that we did not want to adapt an existing taxonomy of body movements from the literature. Some systems were too detailed to fulfill our need for a fast and easy to use system or were too general for our specific task of coding gestures in the context of a human–machine dialogue. Additionally these systems are used for human–human interaction and we felt that the interaction with a machine was a very different form of dialogue than a dialogue with a human being (Dahlbäck et al., 1993). Therefore we decided to develop a system from scratch, although it of course has a lot of similarities with existing taxonomies. For reasons of conciseness, I will only point out differences and similarities to one of the best known taxonomies, the one from Ekman

13 See the demonstration "Evoking Gestures in SmartKom – Design of the Graphic User Interface" (Beringer) at this workshop

(1999).

Ekman distinguishes emblems (movements with precise meaning), illustrators (movements that illustrate speech), manipulators (manipulations of the face/body that are on the edge of consciousness like scratching, stroking etc.), regulators (movements that have the sole purpose of regulating the speech–flow) and emotional expressions.

- Emblems: Up to now, no emblems have shown up in the data. Even if they do we will not include a separate label for them. Although emblems and illustrators can be distinguished reliably (Friesen, Ekman & Wallbott, 1979), there are severe difficulties with the not so obvious cases (Wallbott, 1982). The definitions of the two categories partially overlap. This may be necessary for theoretical reasons – for a practical coding system this is fatal. Therefore we will code emblems (if they show up) with the system described above and include the information about the "emblematic meaning" as a modifier (e.g. under "morphology").
- Manipulators: Are not coded, because they are not of interest in our context.
- Regulators: Are not coded, although they would be of interest in our context. Our broad functional approach however is not suited for the detailed analysis that is needed for coding regulators. Additionally, regulators are mostly other nonverbal behaviors, not gestures (Wallbott, 1982; Duncan & Fiske, 1977).
- Emotional Expressions: Emotional gestures are coded in a single label.
- Illustrators: Are the focus of our concept. We use a slightly different classification of illustrators, see below.
    - Batons (emphasize a particular word), rhythmic movements (depict the pacing of an event): are not coded for the same reasons that were given for not coding regulators.
    - Ideographs (sketch a direction of thought): Correspond (widely) to our "supporting gestures". We needed to define ideographs that are specific for the human–machine dialogue situation.
    - Deictic movements (pointing gestures): Correspond (widely) to our "interactional gestures". To us, they are the most interesting movements. Therefore, we defined several deictic movements that are specific to our context.
    - Kinetographs (depict bodily action), spatial movements (depict spatial relationship): Especially the last would be of great interest in our context ("no, no – *this* big"). We would like to include them – but again, a functional approach is not suited very well to code these kind of gestures. Since we have no data with regard to such gestures, the point will remain unsolved for the moment.
    - Pictographs (draw a picture of their referent): Will probably be included in some way in our system if they show up.

Apart from the categories mentioned above, that simply are not coded or coded as "residual gestures" the main difference between the taxonomy of Ekman (1999) and our system lies in the deictic and the kinetographic gestures. They refer widely to our most and second most important categories, namely functional and supporting gestures. The difference is that we included deictic gestures in the category of supporting gestures.

Why was this done? As has already been mentioned one of our main goals is to define meaningful units within a human–machine dialogue. In this situation a deictic gesture can have the meaning of a selection ("ok, I want this") but in some cases people point at something while they are still making up their minds if they really want the thing indicated ("hm, should I go to this cinema") or points to figure something out ("ah – this restaurant is near my place – or is it?"). It is important to know how often these different kinds of deictic gestures show up and if and how they can be distinguished from real selections.

In a human–human dialogue there are gestures (ideographs for example) which sketch a thought. These gestures normally take place in front of the thorax. We assume that the supporting gestures we describe are such "thought sketching" gestures. Since the subjects think about the things presented on the display, they move their hands in this area (and not in front of their body as in a human–human interaction). We suppose that a display encourages the subjects "to think with the hands" (i.e. trace a thought with their hands). This can be compared to the "thinking aloud" one sometimes uses to help make up one's mind. A fatal situation for a gesture analyzer that has to disambiguate these gestures from real selections!

It can be assumed that this "gesture monologue" is partly increased due to the test situation, in which the subjects are a bit nervous. However, during the first encounter with the system on the street the users will be slightly nervous as well. Therefore such "gestural monologue behavior" can be expected to show up in real situations too. It would be interesting to find out if the verbal monologue (called "off–talk") and the "gestural monologue" are connected and how. We plan to look into this question in the future.

7 Conclusion

We have presented a concept for the coding of gestures in the context of human–machine interaction. The concept is the first step in the development process of a fast and easy to use system for the practical needs of applied research. We did not start from an existing taxonomy because we think that gestures that are used during the interaction with a multimodal machine need a description system that is specifically designed for the context. Therefore we analyzed a set of human–machine dialogues that were recorded with Wizard–of–Oz experiments. The emerging categories are good candidates for selective labels and will be tested for their quality criteria in the second step of the development process. The interaction with the system emerged as an interesting discrimination criterion: Does the user want something from the system or does she react to an output from the system (interaction)? Does she prepare such an interaction (supporting gesture)? For a gesture analyzer it is vital to be able to distinguish a deictic gesture that is a command from a deictic gesture that is *not* a command. The coding system we presented shows a way to examine which differences there are between "conversational" and "non conversational" gestures. Another interesting question for future research is, how a human–machine dialogue and a human–human dialogue differ with regard to "interactional" and "non–interactional" gestures. Human–human interaction has so far been the focus for analyzing gestures in the context of speech – we hope that the new context of human–machine interaction will bring new insights for the topic.

8 Acknowledgments

This research is being supported by the German Federal Ministry of Education and Research, grant no. 01 IL 905. We give our thanks to the SmartKom group of the Institute of Phonetics in Munich that provided the Wizard–of–Oz data. From the group we especially thank Bernd Lindemann and Thorsten Paetzold who were significantly involved in the development of the coding system. Thanks to Florian Schiel, Urban Hofstetter and Bernd Lindemann for helpful comments on the document.

9 References

Bales, R. F. (1970). Personality and interpersonal behavior. New York: Holt.

Beringer, N., Oppermann, D., & Burger, S. (2000). Transliteration spontansprachlicher Daten – Lexikon der Transliterationskonventionen. SmartKom Technisches Dokument Nr. 2.

Dahlbäck, N., Jönsson A., & Ahrenberg, L. (1993). Wizard of Oz Studies – Why and How. Knowledge–Based Systems, 6 (4), p. 258–266.

Duncan, S. D., & Fiske, D. W. (1977). Face to face interaction. New York: Wiley.

Efron, D. (1941). Gesture and environment. Morningside Heights, NY: King's Crown Press.

Ekman, P. (1999). Emotional and conversational nonverbal signals. In Messing, L.S., & Campbell, R. (Eds.), Gesture, speech and sign (p. 45–55). New York: Oxford University Press.

Ekman, P., & Friesen, W.V. (1968). Nonverbal behavior in psychotherapy research. In: J.M. Shlien (Ed.), Research in psychotherapy. Vol. 3 (p.179–216). Washington: APA

Ekman, P., & Friesen, W.V. (1972). Hand movements. Journal of communication, 22, p.353–374.

Ekman, P., & Friesen, W.V. (1978). The Facial Action Coding System. Palo Alto, Calif.: Consult. Psychological Press.

Friesen, W.V., Ekman, P., & Wallbott, H.G. (1979). Measuring hand movements. Journal of Nonverbal Behavior, 1, P. 97–112.

Faßnacht, G. (1979). Systematische Verhaltensbeobachtung. München: Reinhardt.

Freedman, N., Blass, T., Rifkin, A. & Quitkin, F. (1973). Body movement and the verbal encoding of aggressive affect. Journal of Personality and Social Psychology, p. 72–83.

Jönsson, A., & Dahlbäck, N. (1988). Talking to a computer is not like talking to your best friend. Proc. of the First Scandinavian Conference on Artificial Intelligence, Tromso, Norway, p. 297– 307.

Kendon, A. (1980). Gesticulation and speech: two aspects of the process. In M. R. Key (ed.), The Relation Between Verbal and Nonverbal Communication. The Hague: Mouton.

Krout, M. H. (1935). Autistic gestures: An experimental study in symbolic movement. Psychological Monographs, 46, p.119–120.

Laban, R. (1956). Principles of Dance and Movement Notation. London.

Levelt, W. J., Richardson, G., la Heij, W. (1985). Pointing and voicing in deictic expressions. Journal of Memory and Language, 24 (2). p. 133–164.

Mahl, G. F. (1977). Body movement, ideation, and verbalization during psycho–analysis. In: N. Freedman & S. Grand (Eds.), Communicative structures and psychic structures. New York: Plenum Press, p. 291–310.

McGrew, W. C. (1972). An ethological study of children's behavior. New York: Academic Press.

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

Rosenfeld, H. M. (1966). Instrumental affiliative functions of facial and gestural expressions. Journal of Personality and Social Psychology, 4, p. 65–72.

Sainsbury, P., & Wood, E. (1977). Measuring gesture: Its cultural and clinical correlates. Psychological Medicine, 7, p. 63–72.

Wallbott, H. G. (1982). Bewegungsstil und Bewegungsqualität. Weinheim, Basel: Beltz.