Annotation of Conversational Gestures using TASX and CoGesT

Ulrike Gut, Jan-Torsten Milde, Bielefeld University

1 Introduction

This paper describes the work of two closely cooperating projects working on the development of a system for the transcription, representation, analysis and synthesization of gestures in conversational settings. The projects, which are located at Bielefeld University, focus on different aspects of the problem space and provide

1. CoGesT: a formal transcription system, which allows to capture the form of hand gestures based on linguistically motivated categories and the
2. TASX-environment: an XML-based toolset, which simplifies and accelerates the creation of multimodal, multilevel annotated corpora

The next section will explain the CoGesT transcription system in more detail, section 3 briefly describes the TASX-annotator and how the system is going to be extended in order to optimise the annotation process; the final section gives a short conclusion.

2 CoGesT: a transcription system for conversational gestures

The CoGesT Conversational Gesture Transcription System is being developed in the DFG-funded project “Theory and design of multimodal lexica”. It is a twofold system for the transcription of gestures produced in conversational speech:

3. it provides a system of linguistically motivated categories for gestures, and
4. it is a practical machine and human readable transcription and annotation scheme with simple and complex symbols for simple and complex categories.

The term gesture as used here refers to the communicatively relevant position or movement of any limbs or parts of the body of a person involved in face-to-face discourse, most prominently conversation. CoGesT is based on the notion that conversation is multimodal and comprises at least the acoustic and the visual modalities (see also Gibbon et al., 2000). Speakers are assumed to express meanings by means of information in the acoustic and the visual modalities, which are both inextricably linked to and codependent on each other. Each modality can be further divided into several submodalities: The acoustic modality, for example, consists of parallel information streams on different linguistic levels such as the segmental (speech sounds) and the prosodic (pitch, speech tempo, voice quality). The visual modality can be divided into several submodalities according to the various body parts which produce gestures during conversation, e.g. hands and eyebrows. In face-to-face conversation, visual information is always present, and a parallel use of the acoustic information is possible.

The purpose of CoGesT is the description of gestures within this multimodal conversational context. The focus does not lie on an exhaustively detailed description of every aspect of gesture as for example in HamNoSys (see Prillwitz et al., 1989) or FORM (see Martell, 2002). CoGesT focuses on linguistically relevant gestural forms motivated by the functions of gestures within multimodal conversations, and appropriate for collating in a multimodal lexicon. The theoretical assumptions underlying the CoGesT system differ from other descriptions of gestures in three important ways.
5. **Linguistically motivated categories**: It is assumed that the patterning of visual gesture is semiotically organised in much the same way as the acoustically transformed articular gestures in speech and that gestures have both morphological and syntactic rules for structural and sequential combinations.

6. **Clear distinction between form and function**: CoGeST is based on the theoretical assumption that a clear distinction between gestural forms and functions is possible. This is in contrast to McNeill (1992), who claims that there are no separately structured systems of form and meaning in gestures (p. 23). The CoGeST system is being developed for the purpose of representing gestures in both a corpus and a multimodal lexicon. The classification of gestures proceeds according to their functional relation with other modalities of the conversation.

7. **Notational system**: CoGeST provides a notational system based on a clear distinction between categories of gestural form and function and by distinguishing between compulsory basic and additional optional categories of description.

In CoGeST, the form of all gestures is categorized physically according to their temporal and spatial characteristics. To fully specify a gesture, a distinction must be made between source and the target of the gesture and the intermediate trajectory between those two points. Those gestures that can be described with regard to all three aspects are dynamic gestures. Static gestures, in contrast, have only a source and a target point, which are identical. Locations are described according to the position of the relevant limb or body part relative to the speaker-centered coordinate system. Whereas for static gestures it will suffice to define them by their location in space, dynamic gestures can be further described according to their trajectory along the axes up-down, backward-forward, and right-left, relative to the speaker. If necessary, information from two axes can be combined. In addition, trajectories can be classified into different shapes such as circles, squares, or simple trajectories such as straight lines (for a more detailed description, see Gut et. al. (2003)).

### 2.1 The CoGeST notational system

Two important criteria for transcription are machine readability and rendering for human readability. We first propose a feature-based vector notation for machine readability. Abbreviations for human readability are under development. A gesture represents a set of values of the gesture attributes in the following order:

8. source: location and hand shape  
9. trajectory: speed directionality, shape, hand shape, modifiers  
10. target: location and hand shape  
11. symmetry of hands

Location is described by using a modified and extended matrix-like version of the HamNoSys 4.0 charts (see Prillwitz et al., 1989). Hand shape symbols are taken from the FORM hand shape inventory (see Martell, 2002). Thus, a gesture can be described in form of a vector with a maximum of values as listed below:

12. location of source: HamNoSys 4.0  
13. hand shape of source: extended FORM inventory  
14. directionality of trajectory: ri (right), le (left), fo (forward), ba (backward), up (up), do (down) and combinations of those, which are joined with a slash (“/”)  
15. shape of trajectory: ci (circle), li (straight line), wl (wavy line), ar (arch), zl (zig-zag line), el (ellipses), sq (square)  
16. hand shape of trajectory: extended FORM inventory  
17. size of trajectory: xs (very small), s (small), m (medium), l (large), xl (very large)
18. repetitions of trajectory: \( r(0), r(1), r(2), \) etc.
19. speed of trajectory: \( sl \) (slow), \( me \) (medium), \( fa \) (fast)
20. location of target: HamNoSys 4.0
21. hand shape of target: extended FORM inventory
22. symmetry: \( rp \) (right part), \( lp \) (left part), \( sy \) (symmetrical), \( pa \) (parallel)

The transcription system is tentative and is still under development. We currently propose that transcription symbols are concatenated in two ways: in order to connect values on different attribute levels in a simple gesture, a “.” is used, in order to connect two gestures forming a more complex gesture, a “;” is used. The following CoGesT string

\[
15m,5A,ri,ci,1B,l,r(0),me,15m,5A,rp
\]

thus describes an unrepeated gesture carried out with medium speed with the right hand tracing a large circle with a pointed index finger, which starts and ends with the hands in a relaxed position on the lap.

3 Corpus creation with TASX

Based on the formal definition of CoGesT a small sample corpus has been created. Three independent annotators annotated a 15 minute video, starting with a roughly pre-segmented annotation file, manually refining the annotation step by step.

Figure 1 illustrates a two-tier annotation in the telling of a story. The top tier gives an orthographic transcription of the words, and in the bottom tier the gestures of the hand are transcribed.

The corpus was created using the TASX-annotator, a tool which is specifically designed for the annotation of time aligned primary data, such as video and audio files (see Milde and Gut, 2001, 2002). The program stores the transcriptions in an XML-based format called TASX: the Time Aligned Signal data eXchange format. A TASX-annotated corpus consists of a set of sessions, each one holding an arbitrary number of descriptive tiers, called layers. Each layer consists of a set of separated events. Each event stores some textual information (e.g. a CoGesT string, a syllable, a comment) and is linked to the primary video data by two time stamps denoting the interval of this event.

3.1 Extending the TASX-annotator and synthesizing gestures

Currently the TASX-annotator is designed to be a general purpose transcription program, which can be used in arbitrary linguistic (and other) fields. The system is configurable only in limited ways. Therefore we are working on the development and integration of CoGesT extensions, which will aid the human annotators by optimising the gesture transcription process. Problems during the annotation process fall into two classes: first, creating a valid CoGesT string is error prone, because currently the notational format is intended to be primarily machine readable. Second, finding the correct CoGesT string is complex, because
the notational format is very flexible, which makes the annotation process sensitive to the subjective judgement of the human annotator. To cope with the first problem class, a number of extensions for the TASX-annotator are going to be developed, which help the user with the creation of the valid CoGesT strings:

23. **Constrained creation of CoGesT strings I**: with the current version of the TASX-annotator, it is already possible to define hierarchies of context sensitive popup menus, which hold predefined values. An appropriate set-up with the most prominent CoGesT strings is being defined.

24. **Constrained creation of CoGesT strings II**: based on the formal specification of the CoGesT notational format, a context dependent syntax check is being implemented. The system will thus guide the annotator when manually entering CoGesT strings.

25. **Constrained creation of CoGesT strings III**: a macro mechanism is being implemented, that allows to create parameterised CoGesT strings. Instead of specifying the complete CoGesT string, only the most relevant parameters have to be provided by the user.

26. **Visually guided annotation**: for the specification of location (relevant for source, target) and shape, visual maps have been developed. These will allow the user to select valid CoGesT parts by simply clicking onto the appropriate location of the map. A first prototype of this extension is already available.

To cope with the second type of annotation problems, it is necessary to provide the annotator with some means of making the annotation process less subjective. We therefore decided to integrate a gesture synthesizer into the TASX-annotator: **Lokutor**. Lokutor is an anthropomorphic agent living a virtual environment (see Milde, 2000 and Milde and Ahlers, 2000). The agent model is based on a H-Anim conformant skeleton. The system implements the inverse kinematics of the simulated human arm, making it relatively simple to specify complex arm trajectories. The arm movements and the hand shapes can be specified independently. As such, it is possible to fixate the hand form of the agents while the arm is moving along a given trajectory.

Gestures and hand shapes are specified in an XML-based format. Complex gestures may be constructed by chaining simple gestures; the duration of each gesture has to be specified. The format thus allows to set up libraries of simple and complex gestures and hand forms, which can be recombined in multiple ways. The animation process integrates the movements of arms and hands and finally displays the movement in real time. Both formats, the CoGesT notation and the XML-based format of Lokutor are compatible to a large degree. Using the XSL-T based transformation engine, which is built into the TASX-annotator, control sequences for Lokutor will be automatically created from the annotated CoGesT strings. Thus, Lokutor will be used to synthesize the annotated gestures. By this mechanism, the human annotator will be provided with visual feedback displaying a standard version of the annotated gesture and is therefore better able to control whether the CoGesT annotation represents the intended gesture.

**4 Conclusion**

CoGesT and TASX provide a powerful combination for the creation and analysis of multimodal, multilevel annotated corpora. The flexible CoGesT notation is able to represent conversational gestures occurring in natural communication in a machine readable form. The extended version of the TASX-annotator supports the user in the creation of valid CoGesT strings. By integrating the artificial agent Lokutor gestures can be synthesized in a standard way, yielding visual quality feedback to the annotator.
5 Acknowledgment
CoGesT and TASX are jointly developed by Dafydd Gibbon, Ulrike Gut, Benjamin Hell, Jan-Torsten Milde, Karin Looks, Alexandra Thies and Thorsten Trippel.

Bibliography