Originally published in: Lino, Maria Teresa/Xavier, Maria Francisca/Ferreira, Fátima/Costa, Rute/Silva, Raquel (Eds.):
Proceedings of the fourth international conference on language resources and evaluation (LREC'04). 26 May - 28 May 2004, Lisbon, Portugal.
– Paris: European Language Resources Association (ELRA). Pp. 2215-2218.

CoGesT: a formal transcription system for conversational gesture

Thorsten Trippel*, Dafydd Gibbon*, Alexandra Thies*, Jan-Torsten Milde‡, Karin Looks*, Benjamin Hell*, Ulrike Gut†

* Universität Bielefeld, Germany {ttrippel,thies,klooks,ben,gibbon}@spectrum.uni-bielefeld.de
† Albert-Ludwigs-Universität Freiburg i.Br., ulrike.gut@anglistik.uni-freiburg.de
‡Fachhochschule Fulda, milde@coli.uni-bielefeld.de

Abstract

In order to create reusable and sustainable multimodal resources a transcription model for hand and arm gestures in conversation is needed. We argue that transcription systems so far developed for sign language transcription and psychological analysis are not suitable for the linguistic analysis of conversational gesture. Such a model must adhere to a strict form-function distinction and be both computationally explicit and compatible with descriptive notations such as feature structures in other areas of computational and descriptive linguistics. We describe the development and evaluation of a suitable formal model using a feature-based transcription system, concentrating as a firs step on arm gestures within the context of the development of an annotated video resource and gesture lexicon.

1. Objectives

The transcription of gestures in conversation is based on the observation that, in most human conversational contexts, communication is multimodal, involving speech as well as gestures. Meaning is conveyed in different modalities, where the semantic content of the signs in each of them overlap and add to the complex meaning of the communicative context. Examples of gesture types and their semantic content are:

- the identificatio of items in pointing or deictic gestures (McNeill, 1992),
- the use of gestures as word-like units, sometimes called *iconic gestures* (McNeill, 1992) or even highly conventional *emblematic* signs,
- prosody-related movements (called *beats* by McNeill, 1992),
- shape specificatio in descriptions of concrete objects of *metadeictic* type (Gibbon, 1983).

The purpose of the development of the *Conversational Gesture Transcription system* (CoGesT) is to provide a transcription system for the linguistic analysis as well as automatic processing of such gestures.

2. Requirements specificatio

We consider requirements for a standard gesture transcription system which will permit a quality of gesture analysis comparable to that of phonetic analysis, taking the complexities of simultaneous and sequential gesture patterning into account and using the following criteria:

1. Comparability with linguistic notations and their underlying categories in order to permit semantic and "phonetic" interpretation of the transcriptions.
2. Human and machine readability of the transcription and annotation scheme, taking both ergonomic requirements of trained annotators and the need for a well-defined uniquely parsable token stream into account. For this purpose the main target is not to specify the detailed positions, angles and corresponding measurements in absolute values, but to describe relative positions, in order to enable a person familiar with the transcription system to produce an equivalent gesture by interpreting the description and using their knowledge about gesture production. This permits a certain degree of underspecificatio .
3. Clear distinction between form and function in gesture transcription, differing from many previous approaches which rely heavily on intuitive functional categories without regard for form-function distinctions or issues such as gestural homonymy, synonymy, idiomaticity and compositionality, and thus making automatic analysis impossible.

Our criteria are related to language independence, form description and underspecif cation as described for a sign language transcription system by Kennaway, 2003.

In the medium term the gesture descriptions are intended to allow automatic segmentation of gestures, as well as automatic classificatio based on an ontology of lexical gesture patterns. The gesture patterns are integrated into a gestural lexicon, in order to permit the comparison of gestural information with information on other linguistic levels of description and the creation of richly annotated corpora.

3. Design

3.1. Previous approaches

Gesture researchers tend to develop individual transcription systems based on their specifi research questions and development goals. Many classificatio systems have been proposed, mainly function-oriented and comprising a considerable degree of highly personal functional interpretation. Consequently, they often appear fairly speculative, and comparison of them ranges from hard to impossible. We term approaches of this type *functional glossing*, based on the functions of gestures, but without detailed description of forms. The CoGesT system approaches the problem from the other side, providing detailed descriptions of

forms instead of intuitive descriptions of functions. For a thorough overview, see Thies, 2003.

We make a clear distinction between form and function, familiar from linguistics, and initially describe gestures from a perspective comparable with phonetics in the acoustic modality. We term this approach *analytic transcription*. That is, the form of a gestural movement is described initially, as perceived via the visual modality. The form category is only then assigned a functional gloss and other interpretations.

One well-known transcription system (McNeill, 1992) does not explicitly distinguish between form and function when referring to the phases of gestures as *preparation*, *stroke*, *hold*, *retraction*. McNeill's system is fundamentally semantic with regard to categories such as *iconics*, *metaphors*, *deictics*, and *beats*. Form parameters are added in recent work (McNeill et al., 2001, McNeill, 2002) where a detailed approach to temporal synchronisation with other modalities is also presented.

Other systems such as that used by the SmartKom project (Steininger, 2001) concentrate on gestures in human-machine communication whose focus is explicitly limited to the functional rather than the formal level of description.

The most well-known form-oriented systems are HamNoSys (Prillwitz et al., 1989), developed originally for German Sign Language, and FORM (Martell, 2002), developed for conversational gestures and general body motions, both aiming at facilitating a thorough description of gesture. The annotation of comprehensive video corpora with FORM, however, is very time-consuming, owing to the fin granularity of physical description for every picture/frame. The descriptive scope of HamNoSys is, on the one hand, not sufficientl detailed since it is restricted to the gesture space of Sign Languages, but, on the other hand, its overly fin granularity with regard to positions appears to be unnecessary (and too time consuming) for the transcription of conversational gestures. Several systems developed for robotic purposes use exact numerical coordinates, which are useless for human transcribers.

3.2. Gesture objects

The CoGesT system provides a firs approximation towards a gesture transcription scheme that meets our requirements specific tions. The application domain of CoGesT is currently arm and hand gestures, and the fundamental object described by CoGesT transcriptions is the *Simplex Gesture*.

The Simplex Gesture has an obligatory source specifi cation (the location of the hand or arm in space) and an optional route specificatio (the movement of the hand or arm in space). Gestures which consist only of a source are static gestures such as postures and held movements (or holds). Gestures which have a source and a route are *dynamic gestures*, and include a movement. The route consists of a trajectory and a target. A dynamic gesture is consequently fully specifie by its source (the starting point), the target (the end point) and the trajectory between these two points. In McNeill's terminology the starting point would be somewhere in the preparation phase, the movement would be the process of transition between preparation phase via stroke back to the retraction phrase, of which the fina position would be the target.

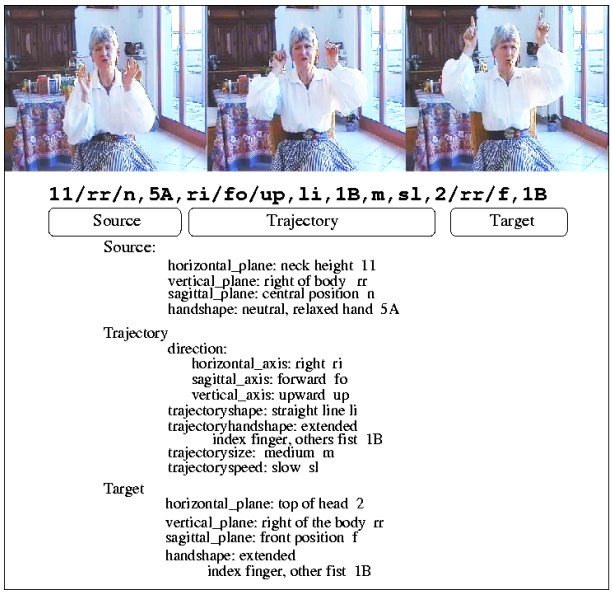

Figure 1: CoGesT vector decomposed

The resulting gesture description is a feature vector, which becomes rather complex, so that an annotation support tool is needed.

3.3. Data structure for Simplex Gestures

The basic data structure used to transcribe the Simplex Gesture is a feature vector. The visual semantics of Simplex Gestures define postures or movements which are carried out with one body part or limb only. Simplex Gestures are only compositional with respect to their internal 'gestalt', somewhat like complex phonetic units such as stop consonants. The notion of Simplex Gesture abstracts away from functional categories, from concatenations of gesture sequences, and from associations of simultaneous movements of different limbs.

Figure 1 illustrates the annotation of the right hand part of a gesture that is performed with both hands, starting with relaxed hands at neck height and moving upwards with a pointing hand. This gesture is taken from a corpus where at the same time growing ears of a donkey are described on the spoken tier.

3.4. Treatment of limbs

Gestures involving different limbs and functional interpretations or temporally related verbal utterance components are annotated on separate tiers, and relations between them are induced a posteriori by distributional analysis of temporal precedence and overlap relations. Examples of induced gestural relations are parallel and mirrored gestures by the arms, or coordinated gestures involving the arm and other limbs (Gibbon et al., 2003).

4. Implementation

A CoGesT transcription vector consists of:

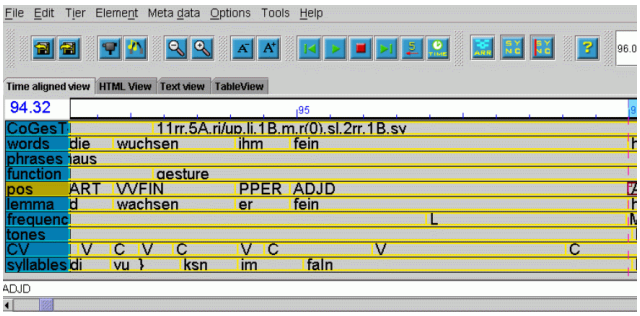

Figure 2: Multi-tier annotation using the TASX-annotator

- The location specificatio for gestures, which refers to a virtual grid over the space in which a body is located. This grid is not meant to be absolute but relative to one's perception, specifying a perceived location in respect to horizontal (19 horizontal divisions), vertical and sagittal (5 divisions each) planes.
- The shape of the hand, which is currently described by codes for the 48 different prototypes that correspond to the handforms used by Prillwitz et al., 1989 and Martell, 2002.
- The movement (if any), which is described in terms of
  - the direction of a movement, which is given in a vector for all three axis relative to the previous location,
  - the shape of the movement, which is described in 7 elementary time functions; for more complex movements the shape of the movement is expressed as an iterative time function with iterations referred to as *microgestures*,
  - the shape of the hand during the movement,
  - a description of the size of a gesture and the speed of the movement,
  - the target location.

However, for practical applications the fuzziness of this method is accepted in order to allow the integration into a multi-tier score with all sorts of other annotation levels, such as prosodic or orthographic annotation or glossing. Figure 2 illustrates such a multidimensional annotation.

4.1. Gesture annotation in corpus creation

A corpus of three narratives in two different languages (German and Anyi (Côte d'Ivoire, West Africa)) containing different modalities was created, including standard morphosyntactic, phonemic, phonetic, prosodic, and orthographic annotations. Gestures were annoated in the CoGesT system for all three narratives.

For the annotation the TASX-Annotator was used, a tool designed specificall for annotating time aligned primary data such as video and audio file (Milde and Gut, 2002). The TASX-Annotator stores the transcriptions in the XML-based *Time Aligned Signal data eXchange* (TASX) format and supports the CoGesT annotation process by incremental monitoring and validating of CoGesT string input. The input tool includes:

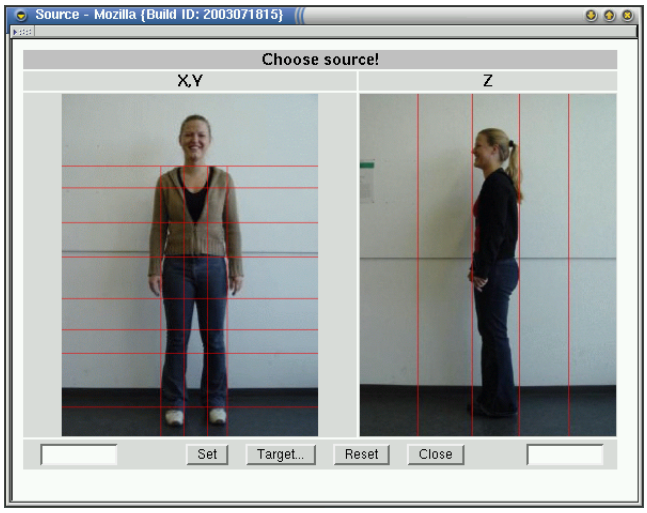

Figure 3: Clickable picture with virtual grid for location specificatio

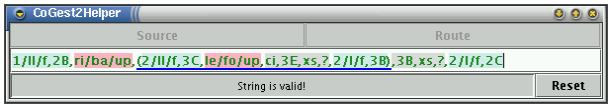

Figure 4: CoGesT string generator with syntax highlighting

- Hierarchies of context sensitive pop-up menus with predefine values, organising the restricted strings in a more usable way. Figure 3 is an example of these applications, showing the virtual grid to specify the location of the hand.
- Context dependent CoGesT syntax check, with syntax highlighting facilities for structuring the gesture and syntactic markup. Figure 4 illustrates this application.
- Macro mechanism for the creation of parametrised CoGesT strings.

5. Evaluation

5.1. Strategies

The original CoGesT transcription system was developed and qualitatively evaluated on a video of practised story narration in German. The video is of 15 minutes duration and contains roughly 60 gesture sequences interrupted by immobile postures.

It is not immediately clear how best to evaluate transcription systems, though a considerable amount of work has been done on this, particularly in regard to prosodic transcription. The problems of gesture transcription are in many ways related to those of prosodic transcription, and it was decided to adopt three procedures which have been used in that context:

Usability: Continual feedback from transcribers was used in order to develop ergonomically optimal usability; the main point made was that 'macro' symbols for common vectors (like phonetic symbols) would improve consistency. This is a matter for further research.

Transcriber Consistency: An evaluation of the CoGesT annotations was conducted with three independent raters. Reliability was measured using a consistency test, areas of divergence were identified and the transcription system was revised accordingly.

Resynthesis and visual inspection: An increasingly accepted evaluation technique, established long ago for spoken language, is resynthesis: the parameters extracted from a signal are used for re-synthesis of the signal and compared with the original. This 'diff' operation is not without problems, but can give an indication of verisimilitude and pointers towards a better analysis. Operational evaluation of the CoGesT-TASX mapping is performed by re-synthesising with an avatar synthesiser, Lokutor (Milde, 2000). The avatar is driven by a gesture lexicon, with interpretation of the abstract feature vector in terms of a realistic coordinate system. Direct visual inspection of the avatar gesture is made in comparison with the original gesture. Currently this is done for a restricted set of gestures only.

For further discussion of relevant criteria of evaluation see Gibbon et al., 2002, Gibbon et al., 1997. For more detail see Gibbon et al., 2003.

6. Outlook and further work

The development of CoGesT continues on different levels. One is the classificatio of gestures by comparing existing annotations. First attempts at gesture type classifica tion based on similarity have been used. Another strategy is to use distance measures to compare the different CoGesT strings for further investigation, as in text classification

On the level of tools, based on the classificatio of gestures, a macro function is being developed for CoGesT. This function simplifie the annotation process by using these macros instead of complex CoGesT vectors.

The annotation is supposed to be used by an interactive gesture resynthesis progam for immediate annotator feedback. For this program a robust animation synthesizer function which creates fully specifie gesture descriptions for avatar input from the underspecifie CoGesT strings is being defined Presently this is only possible for a limited number of gestures in a non-interactive, independent avatar.

The description of the functional categories of gestures, a gesture semantics, remains an open issue. A formal mapping of form to function based on predefine semantic categories is the prerequisite for this fina step. Part of this work will be consideration of synchronisation issues with other levels of language performance (Thies, 2003, McNeill et al., 2001, McNeill, 2002). Model-theoretic work by Carson-Berndsen on the autosegmental modelling of spoken language gestures for automatic speech recognition with event logic and finit state transducers provides a promising platform for formalisation and computational operationalisation.

7. References

Gibbon, Dafydd, 1983. Intonation in context. an essay on metalocutionary deixis. In G. Rauh (ed.), *Essays on Deixis*. Tübingen: Narr.

Gibbon, Dafydd, Ulrike Gut, Benjamin Hell, Karin Looks, Alexandra Thies, and Thorsten Trippel, 2003. A computational model of arm gestures in conversation. In *Proceedings of Eurospeech 2003*. Geneva.

Gibbon, Dafydd, Inge Mertins, and Roger Moore (eds.), 2002. *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation*. New York etc.: Kluwer Academic Publishers.

Gibbon, Dafydd, Roger Moore, and Richard Winski (eds.), 1997. *Handbook of Standards and Resources for Spoken Language Systems*. Berlin: Mouton de Gruyter.

Kennaway, J. Richard, 2003. Experience with and requirements for a gesture description language for synthetic animation. In *Proceedings of the 5th International Workshop on Gesture and Sign Language Based Human-Computer Interaction*. Genova, Italy.

Martell, C., 2002. Form: An extensible, kinematically-based gesture annotation scheme. In *LREC Proceedings*.

McNeill, D., F. Quek, K-E. McCullough, S. Duncan, N. Furuyama, R. Bryll, X-F. Ma, and R. R. Ansari, 2001. Catchments, prosody, and discourse. *Gesture*, 1:9–33.

McNeill, David, 1992. *Hand and Mind: What Gestures Reveal about Thought*. Chicago and London: University of Chicago Press.

McNeill, David, 2002. Gesture and language dialectic. *Acta Linguistica Hafniesia*.

Milde, Jan-Torsten, 2000. The instructable agent lokutor. In A. Nijholt (ed.), *International Twente Workshop on Language Technology L̈earning to Behave(TWLT17. Interacting Agents)*.

Milde, Jan-Torsten and Ulrike Gut, 2002. The TASX-environment: an XML-based toolset for time aligned speech corpora. In *Proceedings of LREC 2002*. Las Palmas.

Prillwitz, Siegmund, Regina Leven, Heiko Zienert, Thomas Hanke, and Janothers Henning, 1989. *HamNoSys. Version 2.0; Hamburg Notation System for Sign Languages. An Introductory Guide*. Hamburg: Signum.

Steininger, Silke, 2001. Labeling gestures in SmartKom - concept of the coding system. Technical report, LMU Munich, Munich.

Thies, Alexandra, 2003. *First the Hand, then the Word: On Gestural Displacement in Non-Native English Speech*. Bielefeld: SII thesis, Universität Bielefeld.

Acknowledgment

The work presented in this paper was funded mainly by the German Research Council grant to the project *Theory and Design of Multimodal Lexica*, Research Group *Text Technological Information Modelling.* Too many colleagues and students have helped with critical feedback following guest lectures and conference presentations to be named here; we are grateful to them all.