A Computational Model of Arm Gestures in Conversation

Dafydd Gibbon, Ulrike Gut, Benjamin Hell,
Karin Looks, Alexandra Thies, Thorsten Trippel

Department of Linguistics and Literary Studies
Bielefeld University, Germany
{gibbon,gut,ben,klooks,thies,ttrippel}@spectrum.uni-bielefeld.de

Abstract

Currently no standardised gesture annotation systems are available. As a contribution towards solving this problem, CoGesT, a machine processable and human usable computational model for the annotation of a subset of conversational gestures is presented, its empirical and formal properties are detailed, and application areas are discussed.

1. A gesture representation strategy

There is as yet no standard generic transcription scheme for gestures, and researchers in the field tend to develop their own systems, based on rather specific research and development goals. Various classification schemes have been proposed (cf. [1, 2, 3, 4, 5, 6]), but none of these is particularly suitable for generalisation into a standard. We provide an initial solution to this problem by selecting a subset of gestures as a starting point, and by a cyclical development procedure with specifications for the gesture transcription system (CoGesT - Conversational Gesture Transcription), formalisation of the specifications, detailed annotation of video recordings of conversational gestures with the TASX Annotator[1], and evaluation of the annotations. Hand gestures are included by adopting an existing taxonomy [7] but are not modelled.

The initial CoGesT requirements specification is:

- the transcription provides a practical *machine and human readable transcription and annotation scheme* with simple and complex symbols for simple and complex categories.
- the transcription system is based on *linguistically motivated distinctions and levels of analysis*, and extends these to the description of gestures,
- the transcription system is formalised in order to facilitate both automatic processing and the creation of interfaces with computational linguistic models of verbal communication.

Unlike HamNoSys, for sign language [8], and FORM [7], CoGesT focuses on linguistically relevant gestural forms motivated by the functions of gestures within multimodal conversations, and appropriate for collating in a multimodal lexicon and for use in multimodal systems [9].

Development proceeds in three phases: first, a multimodal digital video corpus is designed, recorded and processed; second, a full specification of conditions on lexicon structure is developed; third, a lexicon is induced from the corpus (see e.g. [10]), which in this case is a multilingual, multimodal lexicon.

[1] Download from http://tasxforce.lili.uni-bielefeld.de/

In the corpus processing phase the basic, informal gesture transcription annotation system was developed [11]. With this annotation system the initial annotation of the corpus was performed. The present paper is concerned with the second part of the corpus processing phase, and reports on the first formalisation steps for the CoGesT transcription and annotation system, specifically with the development of a basic BNF syntax for machine processing of the selected gesture domain, and as a basis for an XML DTD for the systematic archiving of gesture resources and an attribute-value representation system for integrating. The event-based semantics of the notation will be dealt with at a later stage. On this basis, an attribute-value based microstructure definition for gestures in a comprehensive formal multimodal lexicon is being developed.

2. The CoGesT transcription system

The idea underlying the CoGesT gesture transcription system is to represent gestures in terms of their component features, specifically: *source* and *target* location, *form of the body part*, the *trajectory* between source and target, described in terms of *direction*, *form of movement*, *change of form* of body part during the movement, and *modifiers* for *speed* and *size*. These features are represented in a vector:

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

This vector describes a gesture which starts with the hand on the lap (position coordinate code relative to the body: 15m), in a relaxed position (coordinate code: 5A). The direction of the trajectory is to the speaker's right (ri) and has the form of a circle (ci), with the handform "extended index finger" (1B). The size of the movement is large (l), and the movement is carried out with zero repetitions (r(0)) and medium speed (me). The target position is the lap again (coordinate code: 15m) with a relaxed handform (5A). This vector describes the movement of the hand as the extremity of the right arm, indicated by *rp*.

3. Compositional aspects of CoGesT

The first annotation of the corpus suggests that it is useful to distinguish between the following gesture types in the formalisation:

Simplex Gesture: The "normal" gesture level, described in terms of a 9-position vector.

Simplex gestures are the kind of gestures that have a source position and possibly a movement towards a target. Two categories of simplex gesture were developed:

static (2-place vector): The *static* gestures include *hold* and *posture* cases, e.g.: [*15m, 0B*]

dynamic (9-place vector): All other gestures involving Source, Trajectory and Target have a structure like [*11r, 5C, ri/fo/do, ar, 0B, m, me, 15m, 5A*]. The third position (for direction) is itself a 3-place vector.

For practical purposes, no movement in a dimension may be underspecified, and the vectors are extended by macros (*sy* for symmetric gestures, *pa* for parallel gestures) and indicators for the side of the body part (*rp* for right part and *lp* for the left) for paired body parts. The Simplex Gesture level may correspond to something like a fundamental gestural lexical item.

Microgestures: The components of an iterative simplex gesture, each described in terms of a feature vector embedded into the trajectory position of a simplex gesture.

Gesture Pair: A pair of gestures performed by both members of a limb pair such as the hands additionally requires specification of relations between the members of the pair. This would entail the specification of the Gesture Pair as a Gesture Triple:
$< relation, member_1, member_2 >$

Complex Gesture: A concatenation of gesture pairs with cohesive function in dialogue. Concatenation is represented by "^".

4. Evaluation of formal properties of the CoGesT notation

The CoGesT transcription notation is a first approximation to a manual gesture transcription system which both meets linguistic and computational lexicographic requirements for automatic processing.

The formalisation of CoGesT was developed in three stages: First, the properties of the CoGesT notation are discussed at a pre-formal level, in a first step introducing the level of *simplex gesture* and two levels of *compound gesture*: *simultaneous compound gestures* and *sequential compound gestures*. Second, a formal grammar is proposed. Third, the development of a formal semantics for the system is envisaged, but will not be dealt with in the present contribution. The formal semantics is concerned with gestural event relations of simultaneity (overlap), and precedence expressed in terms of event logic.

5. Evaluation of informal CoGesT vectors

As already noted, the development of CoGesT involved first of all the development of an informal system in close rapport with detailed transcription and annotation of video recordings. An analysis of the informal transcriptions revealed a number of problems which needed to be solved in order to meet the overall specifications. The intended substantive content of the transcriptions is not directly affected by these technical problems, though its application of course presupposes a solution.

5.1. Inconsistencies of vector structure

The flexibility of vector structures using macros and defaults leads to inconsistencies of length and ordering in the initial informal version of CoGesT. For example the vector position denoting direction constitutes a 1-, 2- or 3-place subvector due to underspecification in one or two dimensions. Another inconsistency here is the ordering of the vector. These inconsistencies (which incidentally did not provide practical obstacles for transcribers) were removed in the formalisation.

5.2. Homogeneity of vector value types

In two cases of the annotation scheme CoGesT, feature values are not concerned with properties of gestures but with whole gestures:

1. The final position with the values *sy*, *pa*, *rp*, *lp* does not contain values of the same type as the other positions, but they define macros to be resolved:
   (a) The values *rp*, *lp* designate the body members which share the properties described by the other values in the vector. Since the internal vector order of simplex gestures is fixed, the values can be omitted.
   (b) The elements *pa* and *sy* stand for relations between concurrent gestures of left and right body parts (generally hands): *pa* expresses identity (except for spatial offset) of the two gestures, and *sy* expresses a mirroring of the gesture on the vertical axis.

   For example the following compound gesture with the operator *sy*
   [*13rr, 5A, le/ba/do, li, 0A, l, fa, 15m, 5A, sy*]
   can be expanded to
   [*13rr, 5A. . . 15m, 5A; 13ll, 5A . . .*]
   with 9 elements in each member of the pair.
2. The values *r(n)* refer to repetitions (iterations) of smaller component *microgestures* (cf. 3) which themselves — presumably — have the same structure as the simplex gestures of which they are parts but which are nested into simplex gestures into trajectory position. These values treat repetitions as modifying properties, but they have internal structure. The string
   [*13r, 5C, up, ci, 5C, xs, r(3), fa, 13r, 5C, sy*]
   can be expanded using microgestures into
   [. . . *up, (*$A^{\cap}A^{\cap}$ *A), 5C, xs. . .*]
   with the microgesture:
   $A := 13r, 5C, up, ci, 5C, xs, fa, 13r, 5C$

This step necessitates distinguishing between two hierarchical gesture ranks, the microgesture and what was referred to above as a simplex gesture, and allowing for iterated microgestures to be embedded as whole gestures into the trajectory position of simplex gesture vectors. It is not yet clear what the consequences of this "strict layering" hypothesis would be. This gesture is shown in Figure 1.

5.3. Spelling out defaults

In two cases, abbreviatory defaults were introduced in the informal version, in order to shorten the vector:

1. the distinction between static and dynamic gestures, by which the static variant is reduced significantly,
2. the distinction between the trajectory direction subvectors with 1, 2 or 3 positions.

Strictly speaking, the static gesture vector can be spelled out into a vector of the same length as the dynamic vector, and, as already noted, the alternate orderings and length variants of the trajectory vector element can be spelled out into a full direction vector of length 3, and the order of component elements normalised.

It is important to note that the trajectory subvector is a vector of *movements*, i.e. of *position translation functions* relative

Figure 1: Illustration of symmetrical arm gesture in an oral narrative.

to a coordinate system, such as *up*, *down*, etc., and not a vector of *positional coordinates*.

5.4. Grouping of related vector values into a hierarchical structure

The flat 9-place CoGesT vector for dynamic simplex gestures contains elements which are related to different degrees, and which can therefore be grouped into a hierarchy of subvectors according to these degrees of syntagmatic relatedness. For example, the elements of a sample simplex gesture are:

[*12rr,5A,le/ba/do,ar,5A,m,me,15m,5A*]

can be grouped informally into a tentative attribute-value hierarchy as follows:

$$\begin{bmatrix} \text{Source:} & \begin{bmatrix} \text{Location} & 12\text{rr} \\ \text{Form} & 5\text{A} \end{bmatrix} \\ \text{Dynamic:} & \begin{bmatrix} \text{Trajectory:} & \begin{bmatrix} \text{Directionality:} & \begin{bmatrix} \text{lateral:} & \text{le} \\ \text{saggital:} & \text{ba} \\ \text{vertical:} & \text{do} \end{bmatrix} \\ \text{Shape:} & \text{ar} \\ \text{Form:} & 5\text{A} \\ \text{Size:} & \text{m} \\ \text{Speed:} & \text{me} \end{bmatrix} \end{bmatrix} \\ \text{Target:} & \begin{bmatrix} \text{Location:} & 15\text{m} \\ \text{Form:} & 5\text{A} \end{bmatrix} \end{bmatrix}$$

5.5. Redundancies and gaps

One case of redundancy was already noted: the "right" and "left" designations of objects in the CoGesT vector is redundant if the gesture pair is always represented.

Likewise, the representation of both source and target coordinates and a direction vector is redundant. The target location and target hand shape of a gesture and the source location and source hand shape of the following gesture are identical. Thus, the target labels of the gestures could be omitted for reasons of efficiency in annotation. Nevertheless as manual annotation is error-prone the duplication is useful, enabling automatic plausibility tests.

Therefore the spelling out of defaults is necessary for machine processability and using the macros is essential for efficient manual annotation.

A major gap in the system is that the source and target coordinates do not contain values for the saggital (front-back) dimension. In order to be able to cope with touching versus various distances from the reference points on the torso, a *distance* dimension needs to be added for the discription of arbitrary body parts involved in a gesture, for example, hand-clapping, hands resting on the lap, etc., versus non-contact gestures involving two limbs.

6. Syntax design for CoGesT

The CoGesT notation consists of gesture feature vectors of length 2 or 9 or in case of inclusion of macros of finite length 3 or 11. These vectors occur in pairs separated by a semicolon meaning concurrence, and these pairs may be concatenated with " $\cap$ ". The grouping of vectors proposed in Section 2 yields a tree with a finite depth limit. The microgestures proposed in Section 2 are embedded into the trajectory position of simplex gestures, yielding a grouped or tree structure with finite additional depth limit. These structures are paired, and the pairs are concatenated.

6.1. CoGesT syntax

The description of a tree language with a finite depth limit suggests that CoGesT can be formalised initially as a regular language. For ease of semantic interpretation, the tree structure will be represented by a context-free grammar in BNF notation, which is illustrated in Table 1.

Table 1: BNF Grammar for the CoGesT system

<cogest>	→ <complexgesture>
<complexgesture>	→ <gesturepair>[<complexgesture>]
<gesturepair>	→ <simplexgesture><simplexgesture>
<simplexgesture>	→ <source>[<route>]
<source>	→ <location><handshape>
<route>	→ <direction> (<trajectoryshape> | <microgesture>) <trajectoryhandshape> <trajectorysize> <trajectoryspeed><target>
<microgesture>	→ <source><route>[<microgesture>]
<direction>	→ <lateral><saggital><vertical>
<lateral>	→ ri | le | NULL | ?
<saggital>	→ fo | ba | NULL | ?
<vertical>	→ up | do | NULL | ?
<trajectoryshape>	→ ci | li | wl | ar | zl | el | sq | ?
<trajectoryhandshape>	→ <handshape>
<trajectorysize>	→ xs | s | m | l | xl | ?
<trajectoryspeed>	→ sl | fa | me | ?
<target>	→ <location><handshape>
<location>	→ <height><verticalpos>
<height>	→ 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | ?
<verticalpos>	→ ll | l | m | r | rr | ?
<handshape>	→ 0A | 1A | 2A | 3A | 4A | 5A | 6A | 0B | 1B | 2B | 3B | 5B | 6B | 0C | 1C | 2C | 3C | 5C | 6C | 0D | 1D | 2D | 3D | 5D | 6D | 0E | 1E | 2E | 3E | 5E | 6E | 0F | 1F | 2F | 3F | 5F | 6F | 1G | 2G | 5G | 6G | 5H | 6H | 2I | 5I | 6I | 2J | 2K | 7A | ?

7. Conclusion and Outlook

The present report develops a systematisation and formalisation of the syntax of the CoGesT gesture transcription notation, resulting in the simplified and hierarchically structured human-readable and machine-readable CoGesT notation.

The syntax formulations are intended to be used as follows:

Regular expressions: corpus pre-processing (probably to be implemented in Perl).

BNF: semantic interpretation, lexical induction (probably to be implemented in Prolog) and lexical inference (probably to be implemented in DATR).

XML: archiving and interchange (probably to be implemented with XML tools, and imported into the TAMINO XML database).

An attribute-value representation for CoGesT syntax for integration into a general grammatical description and for the description of CoGesT semantics will be developed later.

The requirements defined at the outset of the report have been fulfilled: provision of a formal basis for a lexical representation of gestures, contribution towards optimising the CoGesT transcription system, checking of the consistency of the system, both in definition and in practical use, mapping of the system on the formalised language CoGesT. Specifically, the following requirements were fulfilled:

- a formally defined machine-readable gesture annotation system is given,
- a BNF syntax description is available for machine processing, and
- a mapping from the attribute-value-based lexicon microstructure definition exists into an XML DTD for the archiving of multimodal lexica.

For practical purposes a subset of this grammar is used, which does not include microgestures and compound gestures, resulting in a context-free grammar which can be implemented using XML syntax and standard XML tools. This has been done for two major purposes:

1. evaluation of annotations: as manual annotation is error-prone the annotations need to be checked automatically. Initial tests accounted for about 1000 labels in 99 gesture segments with 60 syntactic errors resulting in 50% of incorrect CoGesT strings.
2. with Lokutor[2] there is an existing avatar that uses XML specification files for gesture generation. Though the format is different, it can be used for gesture generation in a multimodal lexicon system.

A number of open points remain:

1. The "semantics" of the notation are complex, and involve mapping to formal model structures in the following domains:
   (a) a "meaning" in the conventional sense, e.g. of an emblem gesture meaning "farewell", "the bill please", etc.;
   (b) a mapping to the motor and visual properties of the gesture;
   (c) a mapping to the concurrent linguistic feature vectors, amounting to an autosegmental subvector of segmental phonetic features, as well as vectors defining categories at morphological, phrasal, textual and discourse levels.

   Consequently the "semantics" of a notation such as CoGesT can be defined as a mapping from the CoGesT vectors to (at least) a triple of possible worlds.
2. The gestures, as discovered empirically by the annotation of a video corpus, need to be mapped into a class hierarchy or type subsumption hierarchy of similar gestures in preparation for inducing a set of significant lexical gesture objects from the corpus.
3. The implementation of a transformation function for the XML representation of CoGesT annotations into the input format for the avatar *Lokutor* is needed. This is to be done for immediate feedback for the annotator and for gesture generation in a multimodal lexicon system.

[2] Download from http://coli.lili.uni-bielefeld.de/lokutor/ .

These issues will be dealt with at a later stage of the project *Theory and Design of Multimodal Lexica* in the project *Texttechnologische Informationsmodellierung*(FOR 437) funded by the Deutsche Forschungsgemeinschaft (DFG).

8. References

[1] M. W. Knudsen et al., "Survey of multimodal annotation schemes and best practice." ISLE/IMDI WG Deliverable D9.1, University of Southern Denmark etc., Tech. Rep., 2002.

[2] W. Wundt, *The Language of Gestures*. The Hague and others: Mouton, 1973.

[3] D. Efron, *Gesture, Race and Culture*, 2nd ed. The Hague: Mouton, 1972.

[4] P. Ekman and W. Friesen, "The repertoire of non-verbal behavioral categories: Origins, usage and coding," *Semiotica*, vol. 1, pp. 49–98, 1969.

[5] D. McNeill, *Hand and Mind: What Gestures Reveal about Thought*. Chicago, London: University of Chicago Press, 1992.

[6] B. Butterworth and G. Beattie, "Gestures and silence as indicators of planning in speech," in *Recent Advances in the Psychology of Language. Formal and Experimental Approaches*, ser. NATO Conference Series III: 4B, R. N. Campbell and P. T. Smith, Eds. New York and London: Plenum Press, 1978.

[7] C. Martell, "Form: An extensible, kinematically-based gesture annotation scheme," in *LREC Proceedings*, 2002, pp. 183–187.

[8] S. Prillwitz, R. Leven, H. Zienert, T. Hanke, and J. Henning, *HamNoSys. Version 2.0; Hamburg Notation System for Sign Languages. An Introductory Guide*. Hamburg: Signum, 1989.

[9] D. Gibbon, I. Mertins, and R. Moore, Eds., *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation*. Dordrecht, New York: Kluwer Academic Publishers, 2000.

[10] W. Daelemans, G. Durieux, and A. van den Bosch, "Toward inductive lexicons: a case study," in *Proceedings of the LREC workshop on adapting lexical and corpus resources to sublanguages and applications*, Granada, 1998, pp. 29–35.

[11] U. Gut, K. Looks, A. Thies, and D. Gibbon, "Cogest: Conversational gesture transcription system version 1.0," Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld, ModeLex Tech. Rep. 1, 2002.