From Human Gesture to Synthetic Action

Michael Kipp
University of the Saarland (DFKI)
Stuhlsatzenhausweg 4
66123 Saarbrücken
kipp@dfki.de

ABSTRACT

Embodied Agents are still looking for their own body language. Our efforts aim at analyzing human speakers to extract individual repertoires of gesture. We argue that TV personalities or actors are better suited for this purpose than ordinary subjects. For analysis, we propose an annotation scheme and present the tool *Anvil* for transcoding gesture and posture. For transfer to synthetic agents, we suggest to think of gestures as categorizable in equivalence classes, a subset of which can make up an agent's nonverbal repertoire. Results of investigating different levels of gesture organisation and posture shifts are presented. The concept of G-Groups is introduced which is hypothesized to correspond to rhetorical devices. Also, we give a brief sketch on how these results are planned to be integrated in a multimodal generation system for presentation teams.

Keywords

Nonverbal Behavior, Embodied Agents, Lifelike Characters

1. INTRODUCTION

Nonverbal behavior like gesture and posture plays a crucial role in making communication smoother, facilitating understanding and interacting in a social way. With the rise of synthetic actors in the human-computer interface this area has experienced a surge of interest in an effort to make agents more "life-like". At DFKI, we pursue the idea of interfacing with the user via *teams* of agents that interact with each other in what can be considered a *performance* [2]. Information is conveyed indirectly through observation of dialogues. This offers possibilities of more clearly structured, more socially interesting, more subtle and more entertaining presentation [1]. First informal user studies of this approach have been promising but also revealed a need for more sophisticated generation of nonverbal behavior. As opposed to single agent scenarios [8, 24] we must deal with the additional challenge of making the agents distinguishable individuals. A basic requirement for this are *individual repertoires* of nonverbal behavior.

Our general research agenda has three phases. First, by looking at empirical data, we search for consistent gestures that could be taken as equivalence classes in terms of form and function, and assemble them to individual repertoires. Second, we aim to integrate these results in a running application by modeling instances of these equivalence classes as animation clips. Third, a system evaluation will assess its nonverbal effectiveness.

Looking at empirical data has been done before for other systems [8, 24]. Such studies have a long history in psychological, anthropological and semiotic research [25, 15, 9]. The majority of the literature has been concerned with general properties of "spontaneous gesture" [22]. Subjects were ordinary people chosen without regard to quality of gesture. Also, it has usually not been undertaken to collect individual repertoires [23]. For effective use in synthetic characters, though, the time has definitely come to analyze the kind of speaker we aim to model: a rhetorically proficient speaker who uses his/her nonverbal repertoire to maximum effect, thus giving flesh to the motivation of using a body at all. For such speakers we need to identify individual repertoires, patterns of rhetorically effective delivery and all the different sources that inform the selection of nonverbal signals. We envision a state of the art where interchangeable *gesture profiles* can be used and reused to give a synthetic character its very special human touch.

But can there be any such thing as a *repertoire*? Is the space of personal gesticulation not infinite? As has been argued before [17] this strongly depends on the *topic* of the talk. Describing spatial relations (e.g. giving directions to a foreigner on the street) or actions (e.g. recounting cartoon animations) results in a myriad of different, often singular gestures, most of them probably made up on the spot for the specific purpose at hand. Most such gestures would be pointless to model. On the other hand, "the more abstract and metaphorical the content the gesture pertains to, the more likely we are to observe consistencies in the gestural forms employed" [17]. We understand these "consistencies" as equivalence classes of gesture that can be found, modeled and used for agents involved in "abstract" talk. Abstract topics can be as diverse as sales dialogues, weather forecasts, news reports or literary discussions.

This paper describes work in progress, covering phase one (empirical investigation) and aspects of phase two (integration) in our research agenda. We report on the annotation of video material, the coding scheme involved and coding

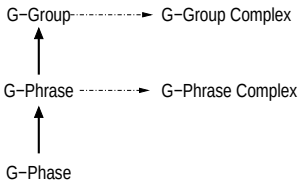

Figure 1: Kinesic hierarchy.

reliability. The tool Anvil[1] [18], especially developed for this project, is presented. Results about repertoires of high frequency gestures, about gesture timing and functions are laid out. Finally, we outline a sample application where the found results can be integrated.

2. ANNOTATION

Our data was taken from the German TV show "Literary Quartet", where four people present and discuss recent book releases. The show is ideal for our purposes for two reasons. First, there is a good mix of abstract talk and interaction. Second, the two speakers chosen for this research, called MRR and HK, have active, proficient gesticulation and a strongly perceived personality. "Proficient" is an intuitive concept here, not to be fully discussed. We only point out that criteria like clear gesture segmentation (in analogy to clear articulation in speech), clear gestural forms, variation in form and tempo could all contribute to the perception of proficiency.

For analysis, we transcribed speech using PRAAT[2] and encoded text structure according to *Rhetorical Structure Theory* (RST) [20] with the RSTtool[3]. Both data were imported in Anvil where posture and gesture were encoded as described below.

2.1 Gesture Structure

Gesture is a fleeting concept. In order to be able to transcribe gesture and talk about timing and function, a structural framework is indispensable.

Gestures can be dissected into more elementary movement phases [15] or clustered together, then constituting a higher unit with a function of its own. This is the essence of the hierarchical view on gesture structure as proposed by [15, 22] and depicted in Fig. 1 (we replaced the G-Unit layer by the G-Group layer and added *complexes*, see below).

The G-Phase layer comprises the movement phases of a gesture. Since this is not the focus of this paper, all we need to know about this layer is that the *stroke* is the most energetic part of a gesture, usually the part that is synchronized with speech (cf. [19, 15, 22] for further reading). Gestures themselves are located on the G-Phrase layer and consist of one or more phases. Many classification systems for gesture have been proposed [23, 10, 22, 9]. We settled on a compromise between semiotic and functional views [16], using the following categories: emblems, adaptors, deictics, iconics, metaphorics and beats. Our priority was clearly that the categories are easy to identify (see Section 2.2 for subcategorization).

[1]Anvil is freely available for research purposes under http://www.dfki.de/~kipp/anvil

[2]by Boersma/Weenik: http://www.fon.hum.uva.nl/praat

[3]visit http://www.sil.org/linguistics/rst/micktool.htm

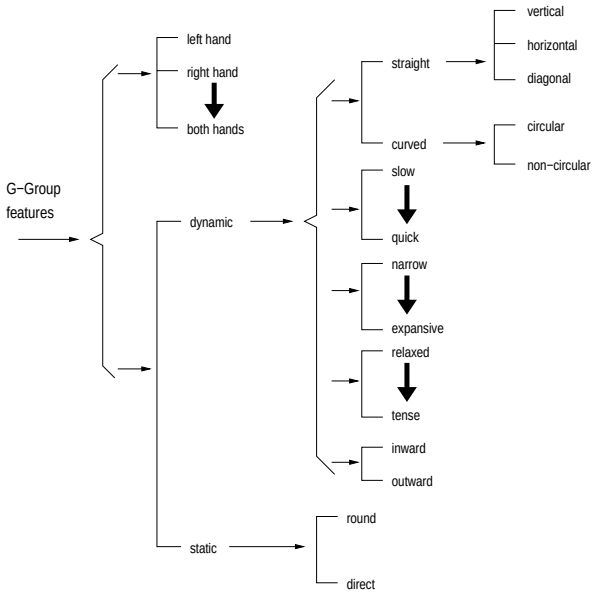

Figure 2: G-Group features with intensity arrows.

On the next higher layer, we define *G-Groups* as sequences of contiguous gestures that share the cohesive properties shown in Fig. 2 (inspired by [21]). E.g., a stretch of two-handed gestures forms a G-Group in opposition to a following group of left-handed gestures. A G-Group has no type, its sole function being the grouping together of gesture phrases. We divert from the literature [15, 22] where the next higher layer is usually the *G-Unit* defined as subsequent gestures between two *rest positions* because we did not find any obvious correlation to speech. Kendon called gesture sequences with shared handedness "groupings or parts" in [15] and has thus considered aspects of G-Groups without putting them into the hierarchy.

On top of the three-layered hierarchy, we call special patterns of G-Phrases and G-Groups *complexes* (see [13] for an analogous language concept). Repetition of gesture is such a pattern on the G-Phrase layer. Likewise, gestures that are frequently used together by an speaker form a complex.

2.2 Coding Scheme

Gestures were transcoded on three levels: G-Phase, G-Phrase and G-Group. We focus on G-Phrases whose subcategories make up our notion of gesture *equivalence classes*.

Gestures are first classified into the six major categories: *Emblems* are gestures with conventionalized meaning that can usually be replaced by a verbalization [5]. *Adaptors* are self-touches or object-touches and serve to disperse excess energy (nervosity) and/or to focus on cognitive processes [12]. *Deictics* are pointing gestures. *Iconics* are gestures that correspond to a speech concept by some direct resemblance relation (e.g. writing into the air while spelling out a name), whereas *metaphorics* relate to speech in a more abstract way (e.g. holding your palm like a cup when asking a question as if ready to "receive" material answer, cf. [22]). Finally, we decided to treat *beats* as a rest class on the G-Phrase layer because we see a functional separation between G-Phrase and G-Phase layer:

Rhythmical beats, that accompanied most of the gestures we encountered, are seen as the repetition of strokes on the

G-Phase layer. Almost every gesture can have such rhythmical beats which serve certain functions (e.g. highlighting new information [22]). On the G-Phrase layer, however, a different level of meaning is added through the specific *form* of a gesture. What are then the traditional beat or baton gestures found in the literature [11][22]? In our opinion, they are gestures whose form does *not* carry any meaning, only their rhythm being important. Therefore, on the G-Phrase layer, such gestures can be considered a rest class. This is a convenient way of dealing with the alternative view that beats are superimposed on other gestures which would complicate annotation.

For the emblems, iconics and metaphorics there is a range of 67 *subcategories* which were found during annotation and documented in a manual with video stills. Entries for emblems specify form, function, concomitant speech, verbalization and similar gesture categories (to avoid confusion). Entries for iconics and metaphorics specify form and function. Deictics and adaptors are coded with one parameter each: aim of deixis (self, addressee, space) and touch region (head, shoulder, table) for the adaptor.

These subcategories are what we hypothesize to be equivalence classes, i.e. instances of one subcategory are taken to be exchangeable by any other instance in this group.

Note that *qualitative* parameters [4] are not coded on the G-Phrase layer. Instead, they are handled on the G-Group layer. Parameters like handedness, speed or direction of movement are seen not as inherent parameters of single gestures but as cohesive glue grouping sequences of gestures together to G-Groups. Subsequent gestures belong to the same G-Group iff they share all the same properties of Fig. 2.

For *posture*, we distinguished between two positions (upright, relaxed), and four kinds of motion (legs-cross, legs-uncross, upper body, and whole body).

2.3 Annotation Tool

For transcoding gesture and posture the generic annotation tool Anvil [18] (**An**notation of **Vi**deo and **L**anguage) was developed (Fig. 3). It allows annotation of video files (e.g. Quicktime or AVI) on multiple layers, so-called *tracks*. In *primary* tracks the user can add *elements* by marking begin and end time. *Secondary* tracks contain elements that point to elements of another, *reference* track. To insert an element, the user specifies first and last element in the reference track. All track elements contain attribute-value pairs. The attributes' names and types are specified by the user for each track. Attributes types are `ValueSet`, `String`, `Boolean`, `Number` and `MultiLink`. For ValueSets the user defines a set of possible values that will reappear in the GUI as an option menu. String typed attributes offer a string input field, Booleans a check box and Number types a slider. MultiLink types allow the selection of arbitrary track elements in the annotation, thereby permitting within-level and cross-level linkage.

For the scheme outlined in Section 2.2, G-Phases were encoded in a primary track, G-Phrases in a secondary track whose elements point to G-Phase elements. Likewise, G-Groups were coded in a secondary track pointing to the G-Phrase track (see Fig. 3).

Anvil stands out in comparison to similar tools in that it is the only non-commercial, fully implemented, XML-based and platform-independent tool specifically designed for the annotation of video material (cf. [18] for details on related work). Especially cross-level linkage is a feature rarely found in other tools.

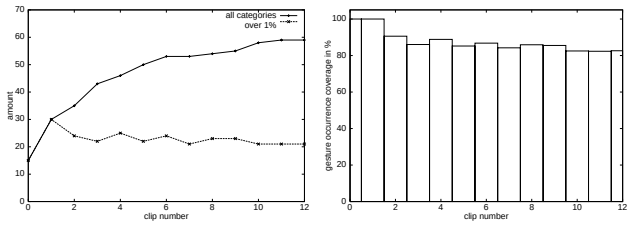

Figure 4: Increase of subcategories with increasing material (left) and percentage of gesture occurrences covered by most frequent categories (right).

2.4 Reliability

Inter-coder reliability was analyzed in two different studies on the *G-Phrase* level to check consistency of the 67 gesture subcategories. In each study, two coders independently annotated 2:54 minutes (study 1) and 2:26 minutes (study 2) of material containing pre-annotated G-Phases of speaker MRR.

Segmentation into G-Phrases (gestures) was near-perfect:

	segmentation	*subcat.*	*subcat. kappa*
study 1	93.0%	64.0%	0.60
study 2	100.0%	79.4%	0.78

In subcategory agreement, the first study yielded critical results. The second study, building on insights of the first study and the grown experience of the coders, resulted in a satisfying 79% agreement and a Kappa value of 0.78 which is very close to "good agreement" [7].

3. RESULTS

Our results yielded evidence that our equivalence class approach is feasible in terms of extraction (see above) and modeling (by taking high frequency gestures, see below). We will also present insights on timing and function.

In all, we annotated 40:09 minutes of data of the two speakers MRR and HK. Tab. 1 shows the amount of phases, phrases and groups found per speaker. We identified 67 different gesture subcategories for both. Naturally, the more material we annotated the more subcategories we found. Fig. 4 shows the increase of gesture subcategories with increasing material for the speaker MRR (left diagram, upper line). When we consider only those subcategories whose instances make up more then 1% of all occurrences we obtain a surprisingly constant amount of 21–24 subcategories (left diagram, lower line). These high frequency subcategories seem natural candidates for being modeled as animations. But how much of the original occurrences of gesture does this set of subcategories cover? Fig. 4 (right) shows the coverage of the MRR's 21 subcategories over 1%. Although it is slightly declining with growing data, in the final corpus of 604 occurrences the "over 1%" subcategories still covers about 83%.

For speaker HK we found 19 "over 1%" subcategories. Comparing the two sets of HK and MRR, there was an overlap of 12 subcategories. Individuality is thus found already in the different repertoire, furthermore in different frequency distribution and usage in function and timing as treated in the next section.

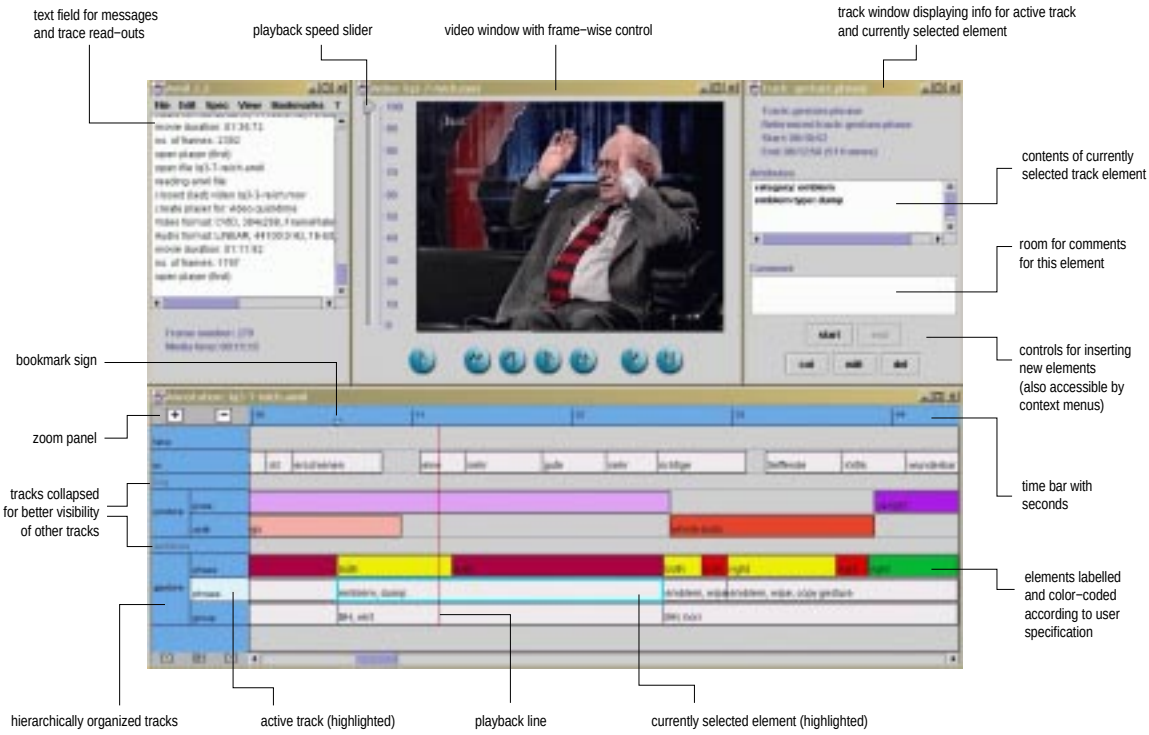

Figure 3: Anvil tool for multi-layered annotation of audio-visual data

	MRR	*HK*	*total*
G-Phases	1,422	663	2,085
G-Phrases	604	265	869
G-Groups	286	103	389

Table 1: Number of encoded entities

3.1 Gesture Analysis

Gestures are timed to co-occur with related speech concepts. Kendon introduced the notion of *idea units* as the common origin of speech and gesture production [15]. These would be manifest in intonation unit (tone units) on one hand and G-Phrases on the other. We tried to find relationships to *syntactic* entities as they would be more readily available in a speech generation architecture. The following timing patterns occurred:

1. *Direct*: Gestures synchronize their stroke directly with a word (noun, verb, adverb...) or a noun/prepositional phrase (NP, PP).

2. *Indirect*: The stroke does not cover the correlated noun, verb etc. itself, but closely linked information, e.g. the noun's modifier or the verb's transitive object or the verb's negation particle.

3. *Init*: Gestures that refer to concepts deeper in a clause or phrase but only cover the first 1–2 words; this is especially encountered with gestures that illustrate a process and correlate to a verb. They occur at the beginning of a verbal phrase (VP) where the verb is in end position, alleviating the fact that in German the verb is often located at the back of the sentence.

4. *Span*: Gesture stroke(s) (plus hold) cover(s) a whole or almost all of the clause.

What follows are some insights on function and timing we gathered for each gesture type, for G-Phrase complexes, G-Groups and G-Group complexes.

Emblems should be most easily identified since their meaning is conventionalized. Functions identified are: speech act, display emotion, display attitude (certainty) and semantic illustration. Some emblems always work with the same timing pattern, e.g., MRR's most frequent speech act emblem "attention" consistently with span timing. Emblems like this also convey a strong high *status* message (according to [14]) that may partly explain MRR's reputation as being highly dominant. A point to be further investigated in the future.

Adaptors are considered to possibly co-occur with "significant points of the speech flow" [25]. In our data, adaptors occurred while the speaker searching for a word (speech failure) or with the beginning of an RST segment, usually covering a short pause. According to Johnstone, all adaptors convey low status [14].

Deictic gestures directed at the addressee function as (1) reference to the addressee by direct sync with pronouns ("you", "your"), (2) reference to the addressee's utterance, timed with the span or init pattern and (3) turn-taking signals [26]: HK exclusively signaled hold-turn, MRR only yield-turn (MRR chairs the discussion). Pointing directly at somebody was also found capable of expressing aggression.

Iconics illustrate objects, processes and qualities [15]. Object gestures are usually in direct sync with a noun. Process gestures sync with a verb (direct), the verb's transitive object (indirect) or the beginning of the VP (init). Quality gestures occur in direct sync with an NP.

Metaphorics are similar to iconics in usage, only their meaning is more generic. The most common *conduit* ges-

ture [22], e.g., can co-occur with almost any object or abstract notion in speech. Some process gestures were found to co-occur with connectives or clause borders, metaphorically indicating motion to the next part of the discussion.

G-Phrase g_1	*G-Phrase* g_2	$\tilde{P}(g_2\|g_1)$
emblem.dismiss	**emblem.wipe**	**0.21**
metaphoric.conduit-fling	metaphoric.conduit-fling	0.20
emblem.dismiss	emblem.dismiss	0.20
emblem.attention	emblem.attention	0.14
emblem.dismiss	emblem.dismiss	0.14
metaphoric.conduit	**iconic.strength**	**0.13**
deictic.space	deictic.space	0.13
metaphoric.conduit	metaphoric.conduit	0.10

Table 2: Most frequent bigrams of MRR's gestures (G-Phrases) with estimated conditional probability.

G-Phrase Complexes are formed, e.g., by repetition. Repetition adds cohesion to the discourse and can bind parts of a sentence together that are separated by embedded structures. But are all gestures suitable for repetition? Statistical analysis results in the patterns in Tab. 2 for speaker MRR (only the most frequent, together with estimated conditional probability). More importantly, if the same two gestures occur together again and again, they can be interpreted as an idiosyncratic compound gesture that extends the original repertoire, thus adding to personal style. We found MRR's frequent emblem sequence "dismiss–wipe" highly characteristic for the speaker.

G-Groups are formed by gestures that share the same properties from Fig. 2. Most G-Groups correspond to one of the two rhetoric devices stated by [3]: List of Three and Contrast. These devices are non-verbally marked in public speaking to coordinate audience reaction (laughter, applause). More generally, List of Three and Contrast can be taken as RST relations and G-Groups as markers for rhetorical segments.

G-Group Complexes. A two-handed gesture is more visible than a one-handed one, an expansive gesture more than a narrow one. To capture a grading of intensity, we introduce an intensity gradation for some dichotomies in Fig. 2. Intensity rises in direction of the arrow. Now, G-Group A is more intense than G-Group B if A has at least one more intense property. If there are more and less intense properties there is no relation. Speaker MRR organizes his G-Groups such that intensity rises (crescendo) and then suddenly drops to give weight to the last part of speech. A simpler identified pattern is contrast: two contrasting segments are covered by G-Groups differing in intensity. A third usage of G-Groups is the marking of embedded clauses to make the surrounding utterance more cohesive.

3.2 Posture Analysis

	total	segment		interaction		*other*
		begin	*end*	*engage*	*disen.*	
cross	24	15	4	—	2	4
uncross	24	4	14	3	—	3

Table 3: Posture shifts of speaker MRR

Posture was hypothesized by Scheflen to correspond to a topic or theme [27, 6]. Consequently, posture *shifts* would signal moving from one topic to another. We investigated for speaker MRR when and how he moved (HK hardly changed posture at all). Specifically, we found that his crossing and uncrossing legs followed a pattern shown in Tab. 3. Most posture shifts considered occurred at the boundary of two RST segments A and B (77%), with a speech pause of 0.2–0.5 seconds between A and B. Moreover, depending on whether the shift occurred at the end of A or at the beginning of B, MRR preferably uncrossed or crossed his legs.

A second correlation, posture shift and interaction, is only slightly indicated by the data (there was too little interaction): MRR uncrossed his legs for engaging in (getting ready for) an interaction and crossed his legs after having disengaged from an interaction.

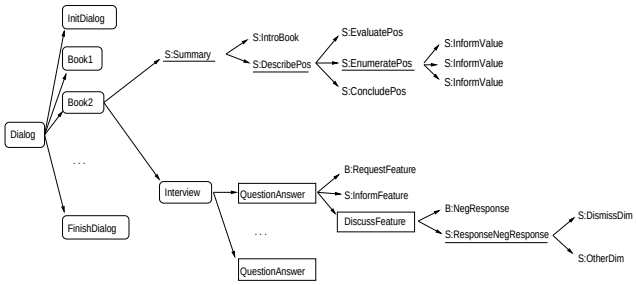

Figure 5: Discourse specification tree

4. APPLICATION SKETCH

As this is work in progress, we only briefly outline how our results will be integrated in an application of presentation teams by pointing at critical decision points for gesture/posture selection:

Gesture repertoires (of subcategories) are modeled as animation clips using the *3D Studio MAX* animation software. Each gesture must be provided with (1) various in/out positions, (2) different postures and (3) different handedness to allow variability and posture shifts. Posture shifts must be modeled separately. Gestures are then classified and stored according to gesture category, function and timing pattern.

As a sample application an existing presentation generation system for selling cars [2] was adapted to a book sales scenario called "Book Jam". The presentation team consists of a book selling agent (S) who presents a number of books to two buying agents (B). The ensuing dialogue, containing monologic passages of the seller and critical questions from the buyers, indirectly presents different facts and aspects of the book to an observing user in an entertaining fashion.

To generate such a dialogue, a plan-based system computes a presentation plan at run-time (a sample is shown in Fig. 5) making use of parameters that model personality and status of the agents. This plan specifies the whole agent-agent interaction, each node representing a plan operator that fired during processing. We now distinguish four node types: *Surface nodes* are the leaves where text and gesture actually appear in their final form. *Rhetorical nodes* (underlined) are those nodes that only contain child nodes of a single agent. The minimal nodes that comprise more than one agent are *interactive nodes* (boxed). All others are called *topical* (boxed, round corners) as they structure larger portions of the dialogue.

To integrate gesture/posture generation into the plan operators we can say what decisions need to be taken in what

node type. Topical nodes decide about posture shifts and adaptors. Interactive nodes on adaptors and turn-taking signals (e.g., deictics), speech act emblems and posture shifts. In rhetorical nodes, G-Groups and G-Phrase complexes are planned (determining handedness and gesture qualities) and also discourse structuring gestures (metaphorics, emblems). Finally, on the leaves, gestures of all types can be selected according to the semantics of concomitant speech. The challenge at surface nodes is the integration of all decisions made here and in higher nodes and the coordination of gestures and posture shifts in accordance with given timing patterns.

At the time of writing, all this belongs to future work. Furthermore, we will think about ways of evaluating the system with regard to the impact of nonverbal behavior, thus tying together empirical results and application.

5. DISCUSSION

We have identified and documented two individual repertoires of gesture that can be used in a generation system for presentation teams. As opposed to related research, we (1) have proposed the analysis of *rhetorically proficient* speakers and (2) have embarked on the effort to collect and analyze *individual repertoires* instead of generalizing over subjects. The main research tool, Anvil, was presented, as well as an annotation scheme and some results on timing, function and frequencies. A new structural layer, the G-Group, was introduced and hypothesized to coincide with rhetorical devices, though more work needs to be done on a conceptual level. Posture shifts were found at borders of larger RST segments. Their timing and form (crossing/uncrossing legs) were identified for one speaker. A sample application was outlined with first ideas how to integrate the results.

For the future, more work on fast, easy and reliable annotation will be done. A coding manual using decision trees for form and function is being considered. Emblem definitions will be made more formal using, e.g., Bitti and Poggi's scheme for classifying holophrastic emblems by specifying (1) arguments, (2) predicate, (3) tense, and (4) attitude [5]. On the generation side we are working on architecture, 3D gesture modeling and multimodal fusion on surface nodes. Attribution experiments are planned to investigate whether some emblems have strong connotations in terms of status (high, low), attitude (positive, negative) and emotion.

6. ACKNOWLEDGMENTS

This research is supported by a doctoral scholarship from the German Science Foundation (DFG). Many thanks to Ralf Engel and Martin Klesen for gesture annotation, and to Elisabeth André for discussion and advice.

7. REFERENCES

[1] E. André and T. Rist. Presenting Through Performing: On the Use of Multiple Lifelike Characters in Knowledge-Based Presentation Systems. In *Proceedings of the Second International Conference on Intelligent User Interfaces (IUI 2000)*, pages 1–8, 2000.

[2] E. André, T. Rist, S. van Mulken, M. Klesen, and S. Baldes. The Automated Design of Believable Dialogues for Animated Presentation Teams. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors, *Embodied Conversational Agents*, pages 220–255. MIT Press, Cambridge, MA, 2000.

[3] M. Atkinson. *Our Masters' Voices. The language and body language of politics.* Methuen, London and New York, 1984.

[4] N. I. Badler, D. Chi, and S. Chopra. Virtual human animation based on movement observation and cognitive behavior models. *Computer Animation Conference*, 1999.

[5] P. E. R. Bitti and I. Poggi. Symbolic nonverbal behavior: Talking through gestures. In R. S. Feldman and B. Rimé, editors, *Fundamentals of Nonverbal Behavior*, pages 433–457. Cambridge University Press, New York, 1991.

[6] P. E. Bull. *Posture and Gesture.* Pergamon Press, Oxford, 1987.

[7] J. Carletta. Assessing Agreement on Classification Task: The Kappa Statistics. *Computational Linguistics*, 22(2):249–254, 1996.

[8] J. Cassell, T. Bickmore, L. Campbell, and H. Vilhjálmsson. Human Conversation as a System Framework: Designing Embodied Conversational Agents. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors, *Embodied Conversational Agents*, pages 29–63. MIT Press, Cambridge, MA, 2000.

[9] D. Efron. *Gesture and Environment.* King's Crown Press, New York, 1941.

[10] P. Ekman and W. V. Friesen. The Repertoire of Nonverbal Behavior: Categories, Origins, and Coding. *Semiotica*, 1:49–98, 1969.

[11] P. Ekman and W. V. Friesen. Hand movements. *Journal of Communication*, 22:353–374, 1972.

[12] N. Freedman and S. P. Hoffman. Kinetic behavior in altered clinical states: Approach to objective analysis of motor behavior during clinical interviews. *Perceptual and Motor Skills*, 24:527–539, 1967.

[13] M. A. K. Halliday. *Language as social semiotic.* Edward Arnold, London, 1978.

[14] K. Johnstone. *Impro for Storytellers.* Routledge/Theatre Arts Books, New York, 1999.

[15] A. Kendon. Gesticulation and speech: Two aspects of the process of utterance. In M. R. Key, editor, *Nonverbal Communication and Language*, pages 207–227. Mouton, The Hague, 1980.

[16] A. Kendon. Gesture and Speech. how They Interact. In J. M. Wiemann and R. P. Harrison, editors, *Nonverbal Interaction*, pages 13–45. Sage, London, 1983.

[17] A. Kendon. An Agenda for Gesture Studies. *The Semiotic Review of Books*, 7(3):8–12, 1996.

[18] M. Kipp. Anvil – a Generic Annotation Tool for Multimodal Dialogue. In *Proceedings of Eurospeech 2001*, 2001. (submitted).

[19] S. Kita, I. van Gijn, and H. van der Hulst. Movement phases in signs and co-speech gestures, and their transcription by human coders. In I. Wachsmuth and M. Fröhlich, editors, *Gesture and Sign Language in Human-Computer Interaction*, pages 23–35. Springer, 1998.

[20] W. C. Mann and S. A. Thompson. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[21] R. Martinec. Types of process in action. *Semiotica*, 130(3/4):243–268, 2000.

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

[23] C. Müller. *Redebegleitende Gesten: Kulturgeschichte, Theorie, Sprachvergleich.* Berlin Verlag, Berlin, 1998.

[24] I. Poggi, C. Pelachaud, and F. de Rosis. Eye Communication in a Conversational 3d Synthetic Agent. *AI Communications*, 2000.

[25] B. Rimé and L. Schiaratura. Gesture and speech. In R. S. Feldman and B. Rimé, editors, *Fundamentals of Nonverbal Behavior*, pp. 239–281. Cambridge University Press, N. Y., 1991.

[26] H. Sacks, E. A. Schlegloff, and G. Jefferson. A simplest systematics for the organization of turn taking for conversation. *Language*, 50:696–735, 1974.

[27] A. E. Scheflen. The Significance of Posture in Communication Systems. *Psychiatry*, 26:316–331, 1964.