Systematicity and Idiosyncrasy in Iconic Gesture Use: Empirical Analysis and Computational Modeling

Kirsten Bergmann and Stefan Kopp

SFB 673 Alignment in Communication, Bielefeld University
Sociable Agents Group, CITEC, Bielefeld University
P.O. Box 100 131, D-33615 Bielefeld, Germany
{kbergman,skopp}@techfak.uni-bielefeld.de

Abstract. Why an iconic gesture takes its particular form is a largely open question, given the variations one finds across both situations and speakers. We present results of an empirical study that analyzes correlations between contextual factors (referent features, discourse) and gesture features, and tests whether they are *systematic* (shared among speakers) or *idiosyncratic* (inter-individually different). Based on this, a computational model of gesture formation is presented that combines data-based, probabilistic and model-based decision making.

Key words: Iconic gesture, meaning-form mapping, systematicity, idiosyncrasy

1 Introduction

The use of speech-accompanying iconic gestures is ubiquitous in human-human communication, especially when spatial information is expressed. Current literature on gesture research states that the question *"why different gestures take the particular physical form they do is one of the most important yet largely unaddressed questions in gesture research"* [2, p. 499]. This holds especially for iconic gestures, for which information is mapped from some mental image into (at least partly) resembling gestural form. Although their physical form, hence, corresponds to object or event features like shape or spatial properties, empirical studies have revealed that similarity with the referent cannot fully account for all occurrences of iconic gestures [19]. Rather, recent findings indicate that a gesture's form is influenced by specific contextual constraints or the use of more general gestural representation techniques such as shaping, drawing, or placing [17, 4]. In addition to those *systematic* patterns in gesture use, human speakers are of course unique and inter-subjective differences in gesturing also hold (cf. [12]). For example, while some people rarely make use of their hands while speaking, others gesture almost without interruption. Similarly, individual variations are seen in preferences for particular representation techniques or low-level morphological features such as handshape [4]. Such inter-subjective differences in gesture behaviour are common and reflect an *idiosyncrasy* of iconic gestures – gestures are created locally by speakers while speaking, without adhering to any conventionalized standards of good form. McNeill & Duncan [25, p. 143] conclude that, by this idiosyncracy, *"gestures open a 'window' onto thinking that is otherwise curtained"*.

In this paper, we look at how systematic and idiosyncratic aspects appear and interrelate with each other in iconic gesture production. We start with empirically analyzing their influence in the formation of gestures, given certain visuo-spatial features of the referent and an overall discourse context. Section 2 introduces the experimental setting and the data coding methodology, Section 3 presents results from the corpus analysis. Based on these findings, we describe in Section 4 a computational modeling account that goes beyond previous systems, which either rely on generalized rule-based models that disregard idiosyncrasy in gesture use [6, 18], or employ data-based methods that approximate single speakers but have difficulties with extracting systematicities of gesture use. These data-based approaches are typically (and successfully) employed to generate gesturing behavior which has no particular meaning-carrying function, e.g., discourse gestures [27] or beat gestures (Theune & Brandhorst, this volume). We propose to combine probabilistic and rule-based decision-making. Embedded into an integrated production architecture, this approach allows for generic, yet speaker-attunded gesture production, which is driven by iconicity as well as the overall discourse context. We conclude with modeling results from a prototype implementation.

2 Empirical study

We aim at identifying systematic and idiosyncratic patterns in the formation of gestures. In our experimental setup, two interlocutors engage in a spatial communication task of direction-giving and sight description, in which they are to convey the shape of objects and spatial relations between them (Fig. 1).

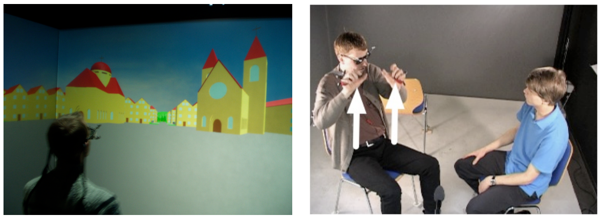

Fig. 1. Experiment setting: VR stimulus presentation (left) and dialog phase with the speaker uttering “the left church has two towers” (right).

2.1 Data Coding

We collected a dialog corpus of 25 dyads (∼5000 gestures). In the scope of the work reported here, we concentrate on descriptions of four different landmarks from 5 dyads (489 noun phrases, 290 gestures). Multimodal annotation has been carried out on several different levels as described in the following (see Table 1).

Gesture Form All coverbal gestures have been segmented and coded for their *representation techniques* for transforming perceived object information into a gesture. Representation techniques capture the aspect that iconic gesturing does not seem to entirely follow the goal to maximize similarity with the referent model, but also brings into play conventions, experiences, and habituations, which people seem to have acquired and apply in their multimodal deliveries (cf. [26, 15, 32]) According to our focus on object descriptions we distinguish the following five categories: (1) Indexing: pointing to a position within the gesture space; (2) Placing: an object is placed or set down within gesture space; (3) Shaping: an object's shape is contoured or sculpted in the air; (4) Drawing: the hands trace the outline of an object's shape; and (5) Posturing: the hands form a static configuration to stand as a model for the object itself.

In addition, each gesture has been coded for its morphology in terms of handedness, handshape, hand position, palm and finger orientation, and movement features (cf. Rieser, this volume). For the scope of this paper, however, we will consider only one of these features, namely *handedness*: each gesture is either performed with the right hand (rh), with the left hand (lh) or with both hands (2h).

Table 1. Coding scheme for gestures and their discourse context.

	Variable	**Annotation Primitives**	**Agreement Coefficient**
Gesture	Representation Technique	indexing, placing, shaping, drawing, posturing	$AC_1 = .784$[1]
	Handedness	rh, lh, 2h	$\kappa = .924$
Discourse Context	Thematization	theme, rheme	$\kappa = .917$
	Information State	private, shared	$\kappa = .802$
	Communicative Goal	lm, lmDescrProp, lmDescrConstr, lmDescrPos	$\kappa = .847$
Referent Features	Subparts	1 or more, none	$\kappa = .91$
	Symmetry	sym, none	
	Main Axis	x-axis, y-axis, z-axis, none	
	Position	3D vector (left, middle, right)	

Discourse Context The transcription of the interlocutor's words is enriched with further information about the overall discourse context. For this purpose, the utterance is broken down into clauses, each of which holding to represent a proposition. For each clause we annotated its communicative goal. Denis [8] developed several categories of communicative goals that can be distinguished in route directions. As we were mainly interested in object descriptions, we revised and refined these for this case into four categories: (1) Landmark (*lm*): a landmark is mentioned without further exploration, e.g., 'there is a chapel'; (2) Landmark property description (*lmDescrProp*): the properties of an object are described as in 'the town hall is u-shaped'; (3) Landmark construction description (*lmDescrConstr*): an object's construction is described, e.g., 'the church has two towers'; and (4) Landmark Position Description (*lmDescrPos*): the description localizes the object as in 'there is a tree in front of the building'.

Clauses are further divided into two smaller units of thematization partitioning of the content of a sentence according to its relation to the discourse context. The structuring

[1] We employ the *first order agreement coefficient AC* since the gestural representation techniques are data of type II according to [10].

of utterances into a topic part and a comment part is a pervasive phenomenon in human language and there are numerous theoretical approaches describing thematization and its semantics (cf. [21]). Following Halliday [11] we distinguish between thematization in terms of *theme* and *rheme* on the one hand, and information focus in terms of *given* and *new* on the other hand. According to the former, a sentence's theme is what the sentence is about. The rheme is defined as what is being said about the theme. For example, in the utterance "the church has two towers" the first noun phrase ("the church") is the theme and the second noun phrase is the rheme. Focussing on noun phrases and their accompanying gestures, to which we restrict our annotation, we annotate information focus following Stone et al. [31] in using the terms 'information state' and distinguish straight-forward between 'private' and 'shared' knowledge: a referent (or referent feature) already mentioned in the previous discourse is 'shared' between interlocutors, whereas a discourse referent which lacks an antecedent in the previous discourse, is not part of the discourse situation is assumed to be 'private'. For instance, in the utterance "the church has a dome-shaped roof" the first noun phrase ("the church") is shared since the must haven been introduced into the discourse before (use of definite article). The second noun phrase ("a dome-shaped roof"), on the contrary, is private because the object (feature) is discourse-new. As suggested in [28], thematization and information focus are annotated independently as different dimensions of information structure, assuming no prior relation between them. In particular, rhematic information is not always private as for instance when content is repeated for better comprehension or in reply to interposed questions.

Referent Features All gestures used in the object descriptions have further been coded for their referent and some of its spatio-geometrical properties. These object features are drawn from an imagistic representation built for the VR stimulus of the study (e.g., houses, trees, streets). This hierarchical representation is called *Imagistic Description Trees* (IDT) [29], and is designed to cover all decisive visuo-spatial features of objects one finds in iconic gestures. Each node in an IDT contains an Imagistic Description (IMD) which holds a schema representing the shape of an object or object part. Object schemas contain up to three axes representing spatial extents in terms of a numerical measure and an assignment value like 'max' or 'sub', classifying this axis' extent relative to the other axes. Each axis can cover up to three spatial dimensions to account for rotation symmetries (becoming a so-called 'integrated axis'). The boundary of an object can be defined by a profile vector that states symmetry, size, and edge properties for each object axis or pair of axes. Links in the tree structure represent spatial relations between parts and wholes, and are defined by transformation matrices. Fig. 2 illustrates how imagery can be operationalized with the IDT model. The spatio-geometrical features entered in the corpus are drawn from these visuo-spatial representations (see Table 1).

3 Results—systematicities and idiosyncracies

As reported in [4] individuals differ significantly in the surface level of their gestural behavior, i.e., in their gesture rate and their preferences for particular representation techniques or morphological gesture features. Our corpus analysis here investigates

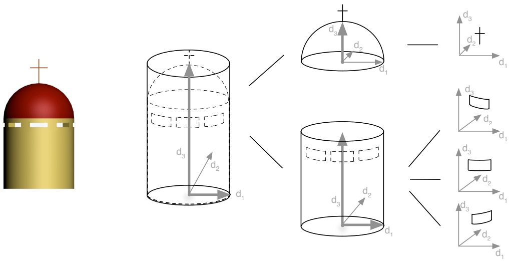

Fig. 2. A church tower (from the VR stimulus) and IDT representation of its shape.

whether such aspects of iconic gesture production seem to be rather systematic, i.e. common among speakers in the same situation, or idiosyncratic, i.e. perculiar to a certain individual. We focus on the following two generation decisions: (1) whether or not a gesture will be produced, and (2) which hand(s) will be used. Since we are dealing with data measured on a nominal scale, we employ Pearson's chi-square test to to judge whether paired observations on two variables, expressed in a contingency table, are independent of each other.

One important question arising is whether the individual differences are due to the fact that different speakers follow different subsets of (possibly shared) dependencies between contextual factors and gestural features? Or do they rather select different features for the same factors, i.e. do individuals diverge from the general tendencies found in the data? We employ two different kinds of measure to investigate these questions. First, we assessed for each individual if the particular (significant) correlation found in the whole data is also significant for the individual. Second, as a measure for individual divergence from the significant common correlations, we assessed for each speaker if her/his joint distributions coincide with the general distribution. Notably, we only considered those cells in the contingency tables in which there is at least a significant difference between observed and expected number of occurrences ($p < .05$).

3.1 To gesture or not to gesture?

The question whether or not a gesture is produced for a particular object seems to be highly idiosyncratic. In the whole corpus (N=25) gesture rates differ from a minimum of 2.34 to a maximum of 32.83 gestures per minute. The mean gesture rate is 15.64 gestures per minute (SD = 7.06). For the five dyads which are analyzed in detail here, the gesture rates vary between 12.5 to 25.0 gestures per minute. The a priori probability for gesture occurrence during a noun phrase in speech is 58.0%. This distribution varies inter-subjectively between a minimum of 44.4% and a maximum of 74.5%.

The choice to produce a gesture is decisively influenced by several other variables, as displayed in Table 2. For the discourse context we found the thematization to be decisive insofar as rhematic information is significantly more likely to be accompanied by a gesture ($\chi^2 = 66.396, df = 1, p < 0.001$). Individuals share this relationship: although the relation is not significant for one of the speakers, all five speakers agree on the

Table 2. Interrelation of gesture occurrence and influencing variables. Parenthetical values are expected occurrences (*p<.05, **p<.01, ***p<.001).

		Gesture (y/n)		Individuals	
		no gesture	**gesture**	**significance**	**similar distr.**
Thematization	**rheme**	103 (142.8) ***	248 (208.2) **	4/5	5/5
	theme	96 (56.2) ***	42 (81.8) ***		
InfoState	**private**	83 (102.1)	168 (148.9)	2/5	5/5
	shared	116 (96.9)	122 (141.1)		
CommGoal	**lm**	17 (10.6) *	9 (15.4)	2/5	5/5
	lmDescrProp	67 (65.1)	93 (94.9)		
	lmDescrConstr	54 (49.6)	68 (72.4)		
	lmDescrPos	61 (73.7)	120 (107.3)		
MainAxis	**none**	44 (47.2)	72 (68.8)	3/5	4/5
	width	44 (43.5)	63 (63.5)		
	height	100 (87.9)	116 (128.1)		
	depth	11 (20.3) *	39 (29.7)		
Subparts	**none**	68 (92.8)**	160 (135.2)*	2/5	5/5
	1 or more	131 (106.2) *	130 (154.8) *		
SymAxes	**none**	97 (74.5) **	86 (108.5) *	2/5	5/5
	sym	102 (124.5) *	204 (181.5)		

distribution, i.e. they tend to use gestures for rhematic information whereas for thematic information gestures are less likely to occur. Regarding the information state, people are more likely to produce gestures for entities whose information state is private ($\chi^2 = 12.432, df = 1, p < 0.001$). This is in line with the view that new information is introduced into the discourse by gesture [24]. Again, all individuals share the same distribution, although the relation is only significant for two of them. So it seems as if this link between information state and gesture occurrence is not as strong as the link between thematization and gesture occurrence. Moreover, the communicative goal has an impact on the question whether or not to gesture ($\chi^2 = 10.970, df = 3, p = 0.012$). When a landmark is just mentioned (lm), this utterance is not very likely to be accompanied by a gesture. This dependence between variables, however, is only significant for two individuals, although all five agree on the distribution by trend. That is, they use less gestures than expected for landmarks which are just mentioned without further elaboration of any kind.

As concerns the influence of referent features, three features appear to be decisive. First, there is a significant relationship between the choice to gesture and the referent's main axis: if from the speaker's point of view the main axis of an object is its depth (e.g. a tunnel into which one is looking) a gesture is more likely to be produced than in other cases ($\chi^2 = 10.424, df = 3, p = 0.015$). For three of the five speakers this relation is significant, and only one speaker does not share the trend. Moreover, the complexity of the object (part) is influential. Utterances which refer to leaf nodes of the IDT representation are more often accompanied by gestures than utterances referring to inner nodes of the tree representation ($\chi^2 = 20.916, df = 1, p < 0.001$). All individuals share this kind of distribution, however, it is only significant for two of them. Furthermore, for objects which have at least one symmetry axis, gestures are more likely to occur than for objects which do not have any symmetry ($\chi^2 = 18.363, df = 1, p < 0.001$). Again, all speakers share this kind of distribution, but it is only significant for two of them.

In summary, the decision[2] to gesture is influenced by two kinds of variables, the discourse context and referent features. As concerns the former, gestures are predominantly produced for rhematic and private information. Regarding the latter, gestures occur more often for less complex (parts of) objects in the sense that they are symmetrical to some degree and have no subparts in the IDT representation. These two findings are rather systematic, i.e., uncontroversial among the five speakers we looked at. However, a significance of the very correlation is not given for all individuals. In other words, speakers vary particularly in how strong the link between particular variables is.

3.2 Which hand to use?

A further analysis concerned another fundamental choice in the generation of gestures: the question which hand(s) to use when referring to an entity. The general distribution of handedness in our data (in which all speaker describe exactly the same spatial scenes) is as follows: with 56.6% the majority of gestures is performed two-handed, while right-handed gestures occur in 28.6% of the cases and left-handed gestures in 14.8%. Again, this distribution is not shared by all speakers. To illustrate this we contrast two particular speakers, P7 and P15. P7 prefers two-handed gestures (65.8%). Accordingly, the number of right-handed gestures (20.5%) and left-handed gestures (13.7%) is reduced. In contrast, P7 has a strong preference for one-handed gestures: 45.1% of this speakers' gestures are performed with the right hand and 25.5% are performed with the left hand. The number of two-handed gestures is accordingly low (29.4%).

Again, we found several correlations in the data constraining this decision (see Table 3). First, there is a significant relationship between the gestural representation technique and the handedness ($\chi^2 = 50.476, df = 8, p < 0.001$). This positive correlation is due to the fact that indexing gestures are hardly ever performed with both hands. On the contrary, shaping gestures are more likely to be performed two-handed. This distribution is shared among all five speakers, whereas significance is only given in four speakers.

Second, there is a significant relationship between the referent's main axis and the gesture's handedness ($\chi^2 = 54.645, df = 6, p < 0.001$): for objects whose major extent is oriented horizontally, two-handed gestures are likely to occur, whereas left- and right-handed gestures occur less often than expected. On the contrary, objects with a main vertical axis are predominantly accompanied by left- or right-handed gestures. The number of two-handed gestures is decreased in these cases. Here we have four speakers sharing this kind of distribution in a significant way.

And finally, as one would expect, the referent's position is influential for handedness ($\chi^2 = 50.893, df = 4, p < .001$). Referents which are on the right from the speaker's point of view are more likely to be accompanied by right-handed gestures and referents which are on the speaker's left tend to be referred to by left-handed gestures. Additionally, for referents which are centered, the number of two-handed gestures is increased compared to expectation. Four of the five individuals agree on the distribution which is significant in three speakers.

In conclusion, we found that a gesture's handedness is not independent from the representation technique used as well as the referent's main axis and position. Individual

[2] The term 'decision' is not meant to imply a conscious process here.

differences, as we measured them, are not very strong in these relations. As we have already seen in the previous section, individuals tend to agree on the general distribution, but may differ in how strong the links are.

Table 3. Interrelation of handedness and influencing variables. Parenthetical values are expected occurrences (*p<.05, **p<.01, ***p<.001).

		Handedness			Individuals	
		2H	**LH**	**RH**	**significance**	**similar distr.**
Technique	**indexing**	3 (20.4) ***	12 (5.3) **	21 (10.3) ***	4/5	5/5
	placing	41 (38.5)	12 (10.1)	15 (19.5)		
	shaping	81 (65.0) *	13 (17.1)	21 (32.9) *		
	drawing	27 (25.4)	2 (6.7)	16 (12.9)		
	posturing	12 (14.7)	4 (3.9)	10 (7.4)		
MainAxis	**none**	47 (40.7)	8 (10.7)	17 (20.6)	4/5	4/5
	width	56 (43.6) ***	1 (9.3) **	6 (18.0) **		
	height	39 (65.6) ***	27 (17.2) *	50 (33.2) **		
	depth	22 (22.1)	7 (5.8)	10 (11.2)		
Position	**left**	42 (44.1)	16 (11.6)	20 (22.3)	3/5	4/5
	right	32 (56.0) ***	21 (14.7)	46 (28.3)***		
	middle	90 (63.9) ***	6 (16.8) **	17 (32.3) **		

3.3 Which representation technique?

Another interesting question in gesture generation is the choice of representation techniques. The distribution of representation techniques in our data is as follows: shaping (39.7%), placing (23.4%), drawing (15.5%), indexing (12.4%) and posturing (9.0%). As described in [4], the choice of representation technique is influenced by the communicative goal of the utterance: descriptions (lmDescrProp) come along with significantly more depicting gestures (shaping, drawing, posturing), while the spatial arrangement of entities is accompanied by indexing and placing gestures in the majority of cases. Moreover, complex objects (without symmetry, or with further subparts) are likely to be positioned gesturally, while less complex objects are more likely to be depicted by gesture. Individuals tend to agree on this general distribution for the most part, but differ in how strong the relations between correlated variables are. Fore a more detailed analysis of the interrelation between use of gestural representation techniques and correlated variables see [4].

4 Computational Modeling

In the previous section we have shown that decisions in the generation of iconic gestures are influenced by a number of variables. Most of the correlations are shared among individuals, whereas for some there is considerable variance among individuals. For a computational model of speech and iconic gesture production three major conclusions can be drawn: first, iconic gestures are not solely implied by the object they are referring to. Rather, ‘thematization’ as a variable characterizing (part of) the linguistic context into which a gesture is embedded is decisive. This correlation goes with empirical findings that speech and gesture production mutually influence each other. On the one hand, information packaging for iconic gestures parallels linguistic packaging (cf. [16, 9, 3]); on the other hand, representational gesturing can have a significant impact on conceptualization as well as lexical access for language (cf. [1, 13]). An adequate model for speech and gesture production, therefore, should allow for a close interaction between content planning and formulation of speech and gesture. Second, generation decisions are influenced by the overall discourse context. For a computational model this means that a discourse record is necessary to distinguish between private and shared knowledge about a referent. Such a discourse record is indispensable for speech formulation to, e.g., decide for the adequate type of determiner (definite or indefinite), and it must be accessible for gesture formulation too. Third, an adequate model of how speakers produce iconic gestures must account for both types of influential patterns, general and individual ones. Previous modeling attempts either ignored idiosyncrasy coming up with generalized model-based approaches [18], or they employ statistical data-driven techniques which have problems with identifying and explicating systematicities from corpora of managable size [27].

We have proposed and described elsewhere [17, 4] a production architecture that is inspired by psycholinguistic models [16, 7] and accounts for our first two requirements. As outlined in Fig. 3 (right) it consists of interacting, modality-specific modules at each of three stages: (1) an *Image Generator* and a *Preverbal Message Generator* are concerned with content planning; (2) a *Speech Formulator* and a *Gesture Formulator* compose and specify, on-the-fly, natural language sentences and gesture morphologies; (3) *Motor Control* and *Phonation* turn them into synchronized speech and gesture animations. This production model adopts a dual-coding approach to multimodal content representation in that an imagistic description (IDT), propositional knowledge, and an interfacing representation out of so-called multimodal concepts are composed and maintained simultaneously. These semantic representations are also utilized for a multimodal discourse model that is available to both speech- and gesture-specific content planning modules, thus meeting the second requirement.

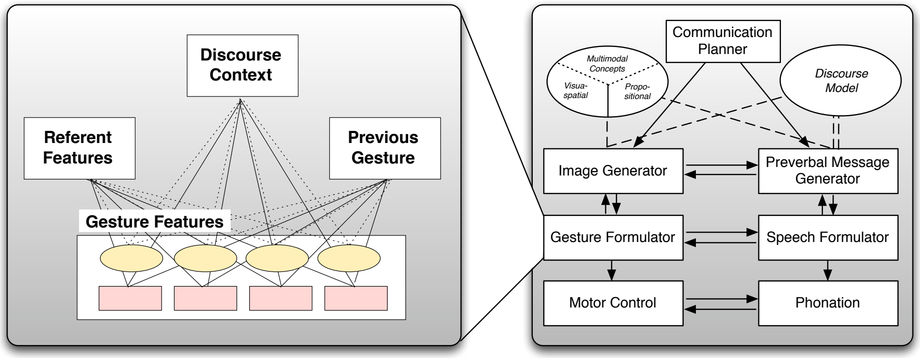

Fig. 3. Overview of the speech and gesture generation model (right), and a zoom in onto the Bayesian decision network for gesture formulation (left).

Accounting for the third requirement, the challenge of considering general and individual patterns in gesture formulation, we employ Bayesian decision networks (BDN) which supplement standard Bayesian networks by decision nodes [14]. This formalism suitably provides a representation of a finite sequential decision problem, combining probabilistic and rule-based decision-making. Each decision to be made in the formation of an iconic gesture, e.g., whether or not to gesture or which representation technique to use, is represented in the network either as a decision node or as a 'chance node with a probability distribution. All factors which potentially contribute to these choices are also entered in the model.

Bayesian networks can be built from the corpus data, both for the whole data corpus, or, for each individual speaker seperately. In either case, the structure of the Bayesian network is learned using the Necessary Path Condition (NPC) algorithm [30]. The NPC algorithm is a constraint-based structure learning algorithm that identifies the structure of the underlying graph by performing a set of statistical tests for pairwise independence between each pair of variables. That is, the independence of any pair of variables given any subset of other variables is tested. The result of structure learning is a network containing all links between variables that are significant for the modeled individual. Once the structure of the network has been found, its maximum likelihood estimates of parameters are computed employing the EM (Estimation-Maximization) algorithm [22].However, not all variables of a complete gesture specification can be learned from the data. This is due to the large set of values some of the variables have. Variables specifying a gesture's morphology, e.g., values for palm and finger orientation, are combined out of six basic values which can moreover be concatenated into sequences to describe dynamic gestures. It is therefore necessary to formulate additional rules and constraints in decision nodes of the network to specify these values adequately.

A resulting decision network is illustrated in the left of Fig. 3. Influences of three types of variables manifest themselves in dependencies (edges) between the large groups of respective chance nodes (drawn as ovals): (1) referent features, (2) discourse context, and (3) the previously performed gesture. Although not in the scope of the current paper, some generation decisions are related to the previous gesture context, i.e., whether the hands have been in a rest position before, and, if already gesturing, which gesture features were found in that gesture (handedness, representation technique). The network is supplemented with decision nodes (drawn as rectangles) which are defined universally, i.e., they do not vary in the individual networks. Nevertheless, each decision node has chance nodes as predecessors so that these rule-based decisions depend on chance variables whose (individual) values have been determined previously. BDNs are suitable for gesture formation since they provide a way to combine probabilistic (data-driven) and model-based decision-making. Moreover, two sources of individual differences are explicated: first, individual 'local preferences for certain aspects are reflected in the respective conditional probability distributions. Second, individuals that do not share significant correlations between variables have a different link structure in their respective networks .

4.1 Modeling Results

A prototype of the previously described generation model has been realized using a multi-agent system toolkit, a Prolog implementation of SPUD [31], the Hugin toolkit for Bayesian inference [23], and the ACE realization engine [20]. In this prototype implementation a virtual agent explains the same virtual reality buildings that we already used in the previously described empirical study. Being equipped with proper knowledge sources, i.e., communicative plans, lexicon, grammar, propositional and imagistic knowledge about the world, the agent randomly picks a landmark and a certain spatial perspective towards it, and then creates his explanations autonomously. Currently, the system has the ability to simulate five different speakers by switching between the respective decision networks built as described above. The resulting gesturing behavior for a particular referent in a respective discourse context varies in dependence of the decision network which is used for gesture formulation. In Figure 4, examples are given from five different simulations, each of which based on exactly the same initial situation. All gestures are referring to the same round window of a church and are generated in exactly the same discourse context ('lmDescrConstr', 'rheme', 'private'). The resulting gesture hence varies depending only on the employed decision network.

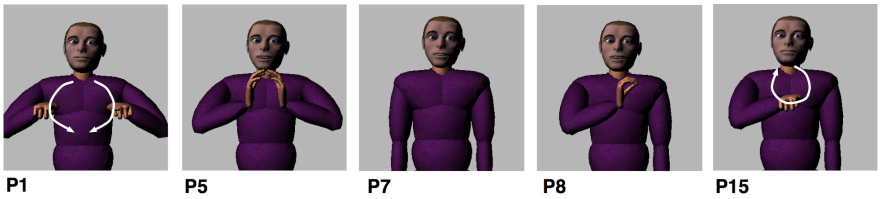

Fig. 4. Example gestures simulating different speakers, each of which produced for the same referent (a round window of a church) in the same initial situation.

In a first evaluation, we measured the model's prediction accuracy by computing how often the models assessment agreed with the empirically observed gesturing behavior. To evaluate the decisions for those four variables we currently assess as chance nodes, we divided the corpus into training data (80%) and test data (20%) and used the training set for structure learning and parameter estimation of the decision networks. In total, we achieved a mean of 62.4% (SD=11.0) highly accurate predictions with the individual networks. The mean accuracy for the rule-based choices made in the network's decision nodes is 57.8% (SD=15.5) (see [5] for details). A perception-based evaluation study is underway to investigate how the generated behavior is judged by human observers.

5 Conclusion

Our objective in this paper was to shed light on the question how systematic (inter-individual) and idiosyncratic patterns interrelate with each other in iconic gesture production. Our empirical corpus analysis has shown that major decisions in the formation of speech-accompanying gestures are influenced by a number of variables, either referent-features or variables characterizing the overall discourse context. Some of the correlations are found among individuals, suggesting systematicity, whereas for others there is considerable variance among individuals, hence suggesting a more idiosyncratic nature. Note, however, that these latter patterns may well be the result of the very dialog situation (including, e.g., the recipient), rather than being hard-wired in the individual speaker. Nevertheless, the reason that different speakers overlap in some features of iconic gestures while they tend to differ in others, suggests that the use of iconic gestures is governed by a number of rather stable systematicities and, at the same time, allows for flexible attunements that may result from a personal or context-sensitive use of iconic gesture by the speaker. A computational model has been developed and the results from a prototype implementation are promising, so that we are confident that our approach is a step forward towards a comprehensive account of iconic gesture generation. Future work in this direction will need to look at a larger spectrum of factors: from individual cognitive skills (as suggested in [12]) to features of the current state between interlocutors. Furthermore, the limitation to five individuals and a set of 290 gestures has to be lifted. Nevertheless, this data suffices to get a first impression of the interrelation of systematic and idiosyncratic patterns and provided a proof of concept for our modeling approach.

Acknowledgements

This research is partially supported by the Deutsche Forschungsgemeinschaft (DFG) in the SFB 673 “Alignment in Communication” and the Center of Excellence in “Cognitive Interaction Technology” (CITEC).

References

1. M. Alibali, S. Kita, and A. Young. Gesture and the process of speech production: We think, therefore we gesture. *Language and cognitive processes*, 15:593–613, 2000.
2. J. Bavelas, J. Gerwing, C. Sutton, and D. Prevost. Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, 58:495–520, 2008.
3. J. Bavelas, C. Kenwood, T. Johnson, and B. Philips. An experimental study of when and how speakers use gestures to communicate. *Gesture*, 2:1:1–17, 2002.
4. K. Bergmann and S. Kopp. Increasing expressiveness for virtual agents–Autonomous generation of speech and gesture for spatial description tasks. In K. Decker, J. Sichman, C. Sierra, and C. Castelfranchi, editors, *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems*, pages 361–368, 2009.
5. K. Bergmann and S. Kopp. Gnetic – using bayesian decision networks for iconic gnetic—using bayesian decision networks for iconic gesture generation. In *Proceedings of Intelligent Virtual Agents (IVA 2009)*, submitted.
6. J. Cassell, M. Stone, and H. Yan. Coordination and context-dependence in the generation of embodied conversation. In *Proceedings of the First International Conference on Natural Language Generation*, 2000.
7. J. de Ruiter. The production of gesture and speech. In D. McNeill, editor, *Language and gesture*. Cambridge University Press, 2000.
8. M. Denis. The description of routes: A cognitive approach to the production of spatial discourse. *Current Psychology of Cognition*, 16:409–458, 1997.

9. M. .Gullberg, H. Hendriks, and M. Hickmann. Learning to talk and gesture about motion in french. *First Language*, 28(2):200–236, 2008.
10. K. Gwet. *Handbook of Inter-Rater Reliability*. STATAXIS Publishing Company, 2001.
11. M. Halliday. Notes on transitivity and theme in english (part 2). *Journal of Linguistics*, 3:199–247, 1967.
12. A. Hostetter and M. Alibali. Raise your hand if you're spatial–Relations between verbal and spatial skills and gesture production. *Gesture*, 7(1):73–95, 2007.
13. A. Hostetter, M. Alibali, and S. Kita. I see it in my hands' eye: Representational gestures reflect conceptual demands. *Language and Cogn. Processes*, 22:313–336, 2007.
14. R. Howard and J. Matheson. Influence diagrams. *Decision Analysis*, 2(3):127–143, 2005.
15. A. Kendon. *Gesture–Visible Action as Utterance*. Cambridge University Press, 2004.
16. S. Kita and A. Özyürek. What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, 48:16–32, 2003.
17. S. Kopp, K. Bergmann, and I. Wachsmuth. Multimodal communication from multimodal thinking–Towards an integrated model of speech and gesture production. *International Journal of Semantic Computing*, 2(1):115–136, 2008.
18. S. Kopp, P. Tepper, and J. Cassell. Towards integrated microplanning of language and iconic gesture for multimodal output. In *Proc. Intern. Conf. on Multimodal Interfaces*, pages 97–104, New York, NY, USA, 2004. ACM Press.
19. S. Kopp, P. Tepper, K. Ferriman, K. Striegnitz, and J. Cassell. Trading spaces: How humans and humanoids use speech and gesture to give directions. In T. Nishida, editor, *Conversational Informatics*, pages 133–160. John Wiley, New York, 2007.
20. S. Kopp and I. Wachsmuth. Synthesizing multimodal utterances for conversational agents. *Computer Animation and Virtual Worlds*, 15(1):39–52, 2004.
21. I. Kruijff-Korbayova and M. Steedman. Discourse and information structure. *Journal of Logic, Language and Information*, 12:249–259, 2003.
22. S. L. Lauritzen. The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis*, 19:191–201, 1995.
23. A. Madsen, F. Jensen, U. Kjærulff, and M. Lang. HUGIN–The tool for bayesian networks and influence diagrams. *International Journal of Artificial Intelligence Tools*, 14(3):507–543, 2005.
24. D. McNeill. *Hand and Mind - What Gestures Reveal about Thought*. University of Chicago Press, Chicago, 1992.
25. D. McNeill and S. Duncan. Growth points in thinking-for-speaking. In *Language and gesture*. Cambridge Univ. Press, 2000.
26. C. Müller. *Redebegleitende Gesten: Kulturgeschichte–Theorie–Sprachvergleich*. Berlin Verlag, Berlin, 1998.
27. M. Neff, M. Kipp, I. Albrecht, and H.-P. Seidel. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics*, 27(1):1–24, 2008.
28. J. Ritz, S. Dipper, and M. Götze. Annotation of information structure: An evaluation across different types of texts. In *Proceedings of the 6th LREC conference*, pages 2137–2142, 2008.
29. T. Sowa and I. Wachsmuth. A model for the representation and processing of shape in coverbal iconic gestures. In *Proc. KogWis05*, pages 183–188, 2005.
30. H. Steck and V. Tresp. Bayesian belief networks for data mining. In *Proceedings of the 2nd Workshop on Data Mining and Data Warehousing*, 1999.
31. M. Stone, C. Doran, B. Webber, T. Bleam, and M. Palmer. Microplanning with Communicative Intentions: The SPUD System. *Comput. Intelligence*, 19(4):311–381, 2003.
32. J. Streeck. Depicting by gesture. *Gesture*, 8(3):285–301, 2008.