J Multimodal User Interfaces
DOI 10.1007/s12193-012-0106-8

ORIGINAL PAPER

Data-based analysis of speech and gesture: the Bielefeld Speech and Gesture Alignment corpus (SaGA) and its applications

Andy Lücking · Kirsten Bergman · Florian Hahn · Stefan Kopp · Hannes Rieser

Received: 29 February 2012 / Accepted: 16 July 2012

Abstract Communicating face-to-face, interlocutors frequently produce multimodal meaning packages consisting of speech and accompanying gestures. We discuss a systematically annotated speech and gesture corpus consisting of 25 route-and-landmark-description dialogues, the Bielefeld Speech and Gesture Alignment corpus (SaGA), collected in experimental face-to-face settings. We first describe the primary and secondary data of the corpus and its reliability assessment. Then we go into some of the projects carried out using SaGA demonstrating the wide range of its usability: on the empirical side, there is work on gesture typology, individual and contextual parameters influencing gesture production and gestures' functions for dialogue structure. Speech-gesture interfaces have been established extending unification-based grammars. In addition, the development of a computational model of speech-gesture alignment and its implementation constitutes a research line we focus on.

Keywords Speech-and-gesture alignment · Multimodal data · Multimodal simulation · Multimodal dialogue · Iconic gesture

A. Lücking (✉)
Goethe-University Frankfurt am Main,
Frankfurt am Main, Germany
e-mail: Luecking@em.uni-frankfurt.de

K. Bergmann · F. Hahn · S. Kopp · H. Rieser
CRC 673 "Alignment in Communication", Bielefeld University,
Bielefeld, Germany
e-mail: Kirsten.Bergmann@uni-bielefeld.de

F. Hahn
e-mail: Fhahn2@uni-bielefeld.de

S. Kopp
e-mail: Stefan.Kopp@uni-bielefeld.de

H. Rieser
e-mail: Hannes.Rieser@uni-bielefeld.de

1 Introduction

In face to face conversation, interlocutors co-produce language and gestures. By 'gesture', we refer to gesticulations and speech-framed gestures according to Kendon's continuum [34,48], namely co-verbal hand and arm movements which contribute to the conversational participants' contributions. Both, words and gesture, are coupled by means of synchrony as well as semantics. They are packaged into bimodal information units [33,48]. To put it in psycholinguistic terms, speech and gesture are aligned [54]. However, to date there is no systematic account of how speech and gestures are used in concert: under which circumstances do speakers make use of co-speech gesture? What motivates the physical form (e.g., handshape, movement trajectory) of a gesture? How is the division of labour between verbal and non-verbal means for their cooperative constitution of an encompassing meaning? We address these topics in an interdisciplinary way, viewing it from a linguistic and a computer science perspective. Theoretical linguistic reconstructions, on the one hand, allow for a formally explicit as well as precise modelling of the interface between speech and gesture. The implementation of theoretical models with computational means, on the other hand, enables to simulate multimodal communicative behaviour in virtual agents or robots. Both research lines, as pursued here, draw on a rich empirical basis in the form of a detailed and systematically annotated speech-and-gesture corpus, called *SaGA*, the Bielefeld *Speech-and-Gesture Alignment* corpus (cf. [45]).

The main focus of our study so far has been on iconic gestures, that are gestures "which exhibit[s] a similarity or analogy to the subject of discourse" ([30] CP 1.369), whereas the "subject of discourse" is introduced verbally. Accordingly, the functioning of iconic gestures can be conceived of as an isomorphic mapping from discourse referents onto

Published online: 12 August 2012

gestural properties realized in the gesture space. In order to reconstruct those iconic mappings, a detailed representation of both gestures and objects[1] is necessary. A detailed gesture representation is implemented by the three-part gesture annotation grid of SaGA that rests on three inhouse annotation manuals (one for gesture classification, one for gesture morphology, and one for discourse gestures). The objects depicted and talked about are entities of a virtual environment. Being already represented in a VR modelling language, the SaGA corpus allows for a rigid empirical investigation of the iconic function of gestures. However, there are good reasons that there is not just one iconic mapping, but rather a variety of them [21]. Regarding gestures, this insight has been elaborated in the work of Müller [52], Kendon [36], and Streeck [70] who distinguish a couple of different gestural depiction functions. For instance, a gesture can depict just the outline of an object, or represent its form three-dimensionally. That is, iconic functions can be distinguished by their domains. Their input can consist of various kinds of properties of the thing to be depicted. We account for different iconic mappings via the specification of a so-called representation technique for each iconic gesture occurrence. Actually, we found that the form of iconic gestures *qua* representation technique is not only determined by the properties of the objects in their domain, but also influenced by aspects of the dialogue context they occur in. The respective data analysis is summarized in Sect. 3.1.

We also pursue an account to iconicity that presupposes just one iconic function. Abstracting away from different depiction methods, the domain of iconic gestures comes out as populated not only with dimensionally closed kinds of entities, but also with entities of mixed dimensionality. The underlying gesture typology is described in Sect. 3.2, providing the basis for a computational simulation approach with virtual agents (Sect. 3.3). How the typology can be used to set up an interface with word meaning within the framework of unification-based grammar is elaborated on in Sect. 4.

Section 5, finally, highlights aspects of multimodal dialogue. Here, it is exemplified how gesture interacts with dialogue structure, how interactive gestures take part in the grounding process, and how such aspects of multimodal dialogue can be incorporated into the multimodal behaviour of a virtual avatar.

At first, however, the SaGA corpus is introduced in Sect. 2. Data collection, data annotation and its evaluation in terms of interrater reliability are described.

2 Experimental setting

A corpus consists of two kinds of data, viz. primary and secondary data. The primary data is the collected empirical material, for instance newspaper articles, video recordings or audio files. The primary data is filed according to metadata, enhanced with annotations or transcriptions. The added information makes up a data set of its own, the secondary data. A corpus is the gathering of both kinds of data, primary and secondary ones. Accordingly, in the following we describe the collection of the primary data underlying the SaGA corpus which has been gathered in an experimental study. Subsequently, the preparation of the secondary annotation data is introduced. Since secondary data usually involves interpreted data produced by a human interpretation process, one has to ask whether the secondary data fulfils the scientific requirement of reproducibility. For this purpose, the secondary data has been evaluated in a reliability study. The rationale of assessing reliability and the respective results for our data completes this section.

2.1 Data and data annotation

The primary data of the Bielefeld Speech-and-Gesture Alignment (SaGA) corpus is built around a virtual reality (VR) town called SaGA town (see Fig. 1). The SaGA town contains five sights worth seeing in addition to a park: a sculpture, a town hall, a church square with two churches, a chapel, and a fountain. The SaGA town was used as the stimulus in our experimental study. Using a VR stimulus enables us to neutralize two confounding variables: First, it allows the researcher to gain complete control over the stimulus. Second, it ensures that each participant gets exactly the same input. The participants, unknown to each other, were grouped into pairs and received a different role each, namely *Route-Giver* and *Follower*. The Route-Giver of each participant dyad was sent on a virtual bus ride along the sights and

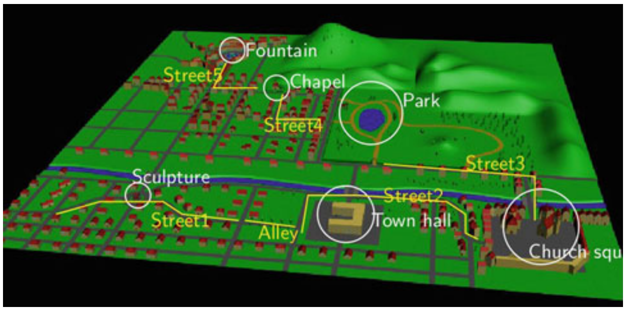

Fig. 1 The SaGA town contains five sights (sculpture, town hall, church square with two squares, chapel, fountain) and a park. The participants are sent on a virtual bus ride along the indicated streets

[1] Assuming a restriction of discourse referents to concrete, but virtual objects, as is the case in many object descriptions. For some notes on an extension to events see Sect. 5.

through the park. Subsequently, the Route-Giver described the route and the passed sights to the Follower.

There are 25 dyads engaged in the spatial communication task combining direction-giving and sight description. The scenario invoked participants to communicate information about the shape of objects and spatial relations between them, a setting that is known to elicit rich gestural behaviour in company to speech [2]. We collected audio and video data from the Route-Giver. For the videotape, three synchronised camera views were recorded. In total, the SaGA corpus consists of 280 min video material containing 4,961 iconic/deictic gestures, approximately 1,000 discourse gestures and 39,435 words.

The data has been completely and systematically annotated based on annotation manuals that have been developed according to theoretical considerations and refined in pilot annotation sessions. Annotation layers divide naturally into two different partitions, the one relating to speech the other relating to gestures. Speech transcription has been carried out using Praat.[2] The utterances of Route-Giver and Follower have been transcribed orthographically on the level of words, with some extensions in order to account for pecularites of spoken language, for instance, the frequent use of interjections ("Ã'h", "mmh") or the melting of finite auxiliars with irreflexive pronouns into one spoken word ["wars" (*was it*), "isses" (*it is*)]. In order to make an empirical study of the iconic functioning of gestures (cf. Sect. 1 above) feasible in the first place, gesture annotation is implemented at a greater degree of detail. The gesture annotation has to provide an analysis of gestural physical components, since "[a]ll of these gestural components may convey meaning." [15, p. 104]. Thus, the annnotation scheme described below differs in focus and content from psychological ones (e.g. [48]) as well as from interactional ones like MUMIN [3] and the thereof derived scheme(s) from the OTIM project[3] [12,13] which aim at the role of gesture in turn management and in which "[i]nternal gesture segmentation is not considered." [3, p. 277]. The SaGA scheme also differs from "affiliation-based" schemes that aim at the temporal relationship between gestural locations or movements and their verbal affiliates [37] in that SaGA is not "restrict[ed …] to hand/arm movement" (p. 326), but also recognizes the significance of, for example, palm orientation or finger configuration. In order to get at the gesture features that potentially make up the value of an iconic function, we have to pursue a kinematical approach to gesture annotation (cf. also [35]). The first systematic kinematic decomposistion of gestural components has, to our knowledge, been explored in the work of Calbris [15]. A related factorization of gestures has been implemented in the CoGEST scheme [22,23]. A systematic extension that is based on the anatomically joints of arm and hands that bring about movement has been implemented in the FORM scheme [47]. The same basic approach has also been pursued by Kopp et al. [39]. The kinematic—or "morphologic", as we call it—part of the SaGA scheme is closely related to the latter two schemes. However, it is the last one we build upon wihin SaGA.

At first, all gestures have been segmented in order to specify the stroke phase [33]. The gestures (i.e. strokes) are then typed for belonging to a certain kind, namely *deictic*, *iconic* or it discourse. 'Iconic', a term coined by McNeill [48] alluding to a Peircean triad. Müller [52] sets up a more fine-grained classification of gestures according to what the hands do. According to our domain of application, we adopt or modify the sets of representation techniques proposed in the literature (see [36,52,70]). The classification of gestures within SaGA now distinguishes the following eight representation techniques:

- *Indexing* pointing to a position within gesture space;
- *Placing* as if an object is placed or set down within gesture space;
- *Shaping* as if an object's shape is contoured or sculptured in the air;
- *Drawing* as if the hands trace the outline of an object's shape;
- *Posturing* the hand(s) form(s) a static configuration to stand as a model or as a proxy for the object itself;
- *Sizing* as if hands or fingers indicate a certain distance or size;
- *Counting* fingers are used to enumerate things by means of what can be construed as an iconic representation of a tally sheet;
- *Hedging* an indication of uncertainty (typically by a wiggling or shrugging movement).

In addition, each gesture has been coded for its so-called morphology which comprises a specification of handshape, wrist position, palm and back of hand (BoH, for short) orientation. Movement within any of these dimensions is coded in terms of concatenations of orientation predicates. The morphology annotation makes use of the fact that the gesture space of the speakers naturally embodies a spatial reference frame. The gesture space is oriented along the anatomical planes, i.e. the sagittal plane, the transversal plane and the frontal plane. The respective orienting axes provide reference points for determining perspective predicates like 'left', 'ahead', 'above' etc. In detail, gesture morphology annotation consists in the specification of the following kinematic or physical gesture properties:

[2] http://www.praat.org.

[3] http://aune.lpl.univ-aix.fr/~otim/index.html, accessed May 9, June.

- The handshape used in the performance of a gesture stroke is coded in terms of the vocabulary given by a modified American Sign Language lexicon [67].
- Palm orientation is specified in terms of the direction of an axis orthogonal to the palm, whereby the following six speaker-centric half-axes were used [31]: 'forward', 'backward', 'left', 'right', 'up' and 'down'. Up to three of these basic values are combined into "slash categories" to encode diagonal or mixed directions, for instance 'up/right' or 'up/right/forward'.
- The orientation of back of hand is treated like palm orientation.
- We use *Wrist Position* for anchoring a gesture within regions of gesture space like "right of body, at the height of shoulder" or "straight ahead of chest". In addition, the extension of a gesture is specified via its distance to the gesturer's body. We distinguish five distance predicate, ranging from contact of hand with body to outstretched arm.
- For dynamic gestures, that are strokes that involve motion, the gestural movement is captured by means of sequences of the annotation predicates (including slash categories) described above. Sequences of annotation predicates are built by means of the temporal '>'-concatenator. For example, if a hand is first moved upwards and then moved downwards, the wrist position would be specified by the dynamic value 'up>down'.
- To further classify the type of movement trajectory, we distinguish between linear and curved movements. Assume, for instance, the sequence of directions 'up>right>down>left'. If it is performed linearly, the resulting trajectory would be a square whereas it would be a circle if the same sequence would be performed in a curved way.

Note that this annotation scheme along with the inventory for decomposing the physical performance of a gesture go beyond previous schemes in various respects. First, anatomically systematic dimensions of palm and BoH orientation are not recognized within the trajectory-oriented CoGesT scheme [22,23]. The FORM scheme [47] is not rooted in the McNeillean notion of gesture space and does not account for functional gesture types or representation techniques. The scheme of Calbris is only concerned with "straight-line gestures in space […]" [15, p. 104] but not with gestural movements of any kind as we are here. The present annotation schemes builds on the scheme used in Kopp et al. [39] to capture two-handed gestures and the manifold configurations they can manifest.

The direction dialogues are enriched with further information about the overall discourse context. For this purpose, an utterance of a participant is broken down into clauses. A clause is the minimal syntactic unit that expresses a proposition. Each clause is annotated for its associated communicative goal. Denis [20] developed several categories of communicative goals that can be distinguished in route directions. We revised these for our purposes into four categories, namely:

- Naming a landmark;
- landmark property description;
- landmark construction description; or
- landmark position description.

Following Halliday [29] we distinguish the thematization structuring of clauses in terms of *theme* and *rheme*. The theme is the topic of an utterance, of which the rheme is predicated. Since the topic does not always coincide with an already introduced discourse referent [40, p. 265] we additionally distinguish the information foci *given* and *new*. Borrowing the terminology of Stone et al. [69], information foci are classified according to the information states *private* or *shared*.

Recall that the SaGA town is a VR town which is built from uniquely named, mereologically organized constructors (like windows, walls, houses, etc.). The fixed referential domain makes it at least in principle possible to specify for any gesture used in an object description the constructor it depicts. This referential interpretation of gestures has been done for a subset of seven dialogues (where a dialogue corresponds to a complete video) so far. In addition, some spatio-geometrical properties of the referent are coded. These object features are drawn from an imagistic representation built for the VR stimulus of the study. Note that this kind of information is hardly available for field data.

The gesture annotation within SaGA is realized within Elan.[4] The multimodal corpus data are stored, retrieved and transformed within the Ariadne system [51].

2.2 Reliability of annotation

The focus of assessing the reliability of annotations is reproducibility (see [41]) for different evaluation foci). Several (at least two) annotators rate the same set of data. The degree of agreement between their ratings provides an index for the reliability of the annotations. A rating is a measurement procedure where a "two-legged meter" ([68, p. 194], quoted from Cohen [18]) classifies some target data according to a set of response categories. The human annotator functions as an interpretive switchpoint bringing about the measurement. Here lies the starting point for a qualitative distinction, namely the distinction between Type I versus Type II ratings [25]. Type I measurements are those where the human inter-

[4] http://www.lat-mpi.eu/tools/elan.

pretation effort leading to a rating is well-understood and, hence, the outcome easily evaluable, whereas this is not the case for measurements of Type II. An annotation of Type II can be understood as "rating under epistemic uncertainty" [66, p. 29]. The uncertainty involved in Type II ratings is a source for random annotations. Ratings based on a random decision, i.e. a decision driven neither by the traits of the target datum in question nor by the coding instructions, are unreproducible and therefore do not provide a base for assessing reliability. This difference in the qualitative status has to be accounted for in evaluations of respective annotations: Type II annotations have to be adjusted for chance-based agreements (cf. [16,18]). The gesture annotation introduced in Sect. 2.1 comprises both types of annotation data, Type I and Type II. The classification of gestures in terms of representation techniques is prone to uncertainty. It is not well-defined which observable features a gesture has to exhibit in order to be classified as, say, shaping. There is some intuitive understanding of representation techniques, to be sure. However, the regarding measuring is highly interpretive and can not be made fully transparent (e.g., completely reduced to perceptible gesture features). One serious reason for this is that the understanding of a gesture depends on the linguistic environment of that gesture.

The second set of gesture annotations, describing the so-called gesture morphology, make up data of Type I. For instance, specifying the orientation of a hand movement by directions ('left', 'right', 'front', ...) is well-understood and has clear-cut application conditions. According to the Type I/Type II data distinction, we employ different methods in order to evaluate annotations of representation techniques and annotations of gesture morphology. As a chance-corrected coefficient determining the level of agreement to be found in Type II data, we calculate the first order agreement coefficient AC1 developed by Gwet [25]. In order to assess the extent of association between annotations of the Type I gesture morphology, we follow a strategy employed by Bergmann and Kopp [6]: annotation predicates for the orientation of the hands within gesture space are translated into angle measures. (Dis-)agreement can then be calculated in terms of angular deviations, but keep their ordinal data type.

The size of the sample of gestures that is large enough to reasonably test for agreement has been calculated for the following values, set in the run-up to the reliability study: we assumed to test for a reasonable agreement level of 70 % with an α-error of 0.05 and a β-error of 0.85. The resulting n of 477 gestures (i.e., segmented movements) has been drawn from the morphology as well as from the technique annotations. The Type I morphology sample has been classified by four expert annotators, the Type II technique sample by three expert coders. The reliability of gesture segmentation has been addressed separately, see Sect. 2.2.4.

Table 1 Overview of Type II data reliability evaluation

Technique	Referent	InfoStruc	InfoState	Goal
0.784	0.91	0.95	0.86	0.88

Values denote AC1 coefficients

2.2.1 Type II annotations

The resulting first-order agreement coefficient AC1 for gesture technique rating is 0.784. It's confidence interval is (0.758, 0.81), so that the proportion of agreement on gestures' representation modes given that the agreement is not due to chance is significantly greater than 75 %, which complies with our initially demanded reliability level. This also holds for the speech-related Type II annotations, namely the coding of information structure, information state, communicative goal and the VR referent. The respective coefficients are collected in Table 1.

2.2.2 Type I annotations

The annotations that make up the Type I data of the SaGA corpus transcribe the movement of a gesture within gesture space—cf. the annotation description from Sect. 2.1. The gesture space is a three-dimensional region which extends along the sagittal, transversal, and frontal axis. The respective annotation predicates thus have a clear spatial interpretation. Nevertheless, annotators may map an observed movement onto different category labels or simply err. However, the disagreement between, say, "movement to the right" and "movement to the right and slightly down", is less than that between "movement to the right" and "movement to the left". Comparing just for sameness of annotation labels would not capture the degree of spatial difference between them. In other words: treating movement annotations as nominal data will miss their ordinal scale information.[5] We address this problem by translating the annotation labels into angular measures which can be analysed in terms of numeric differences. The smallest angular deviation is 2.36° for the direction of hand shapes, the biggest one is 46.16° for BoH direction. On average, the angular difference for gesture morphology as a whole is 27° (with average standard deviation SD = 45). Given that the annotation categories resolve gesture space into "slices" of 45° each, the average difference comes close to the theoretically undecidable mean value of 22.5° (45/2°). Table 2 provides an overview of the angular deviations between annotators.

[5] Since the movement annotation categories are coarse-grained in the sense that they map a range of positions within gesture space onto just one category, they are ordinal rather than interval or ratio scaled.

Table 2 Overview of Type I data reliability evalution

BoH orient	BoH dir	Palm orient	Palm dir	HandShape dir	Wrist dir	HandShape
20.66° (2.47)	46.14° (13.64)	19.14° (1.92)	36.86° (20.33)	2.36° (1.11)	37.08° (6.5)	83 % (AC1 = 0.9)

Values denote mean angular deviation between annotations. The respective standard deviation is given in parenthesis. 'BoH' stands for "back of hand"; 'orient' and 'dir' abbreviate "orientation" and "direction of movement", respectively. For the sake of completeness the table also lists the percentage of accordant handshape classifications—for details, please consult the text

2.2.3 Hand shapes

Evaluating the annotation of hand shapes requires a special treatment, since the categories developed to classify the hand shape observed comprise both Type I and Type II shares. In the first instance, there is a set of basic shapes derived from the *American Sign Language* (ASL) lexicon. These Type I labels are then enhanced by Type II modifiers such as "loose" or "spread". The strategy we pursue is to map all modified hand shapes onto their basic type and treat them as Type I data. As a result, we found that the four annotators agreed on 83 %. For the sake of comparison, we also calculated the chance-corrected agreement coefficient for the hand shapes within the sample of gestures drawn for reliability assessment. The resulting AC1 value was 0.9.

2.2.4 Gesture segmentation

The segmentation of the gestures phases preparation, stroke and retraction (plus eventually some hold phases) poses a reliability assessment problem on its own, since it does not fall in the "assignment of category-"setting. The usual annotation task consists in assigning a given item to one of a set of several response categories, usually of nominal data type.

To the contrary, the segmentation of movements into gestures is an instance of a "marking of items"-setting. Thus, gesture segmentation in the first place provides the items that are the objects of classification in the "assignment of category"-setting. The generic problems for an account of assessing agreement of segmentations are summarised by Thomann [73, p. 340] as follows:

- Each observer produces a different number of markings.
- There is a free "choice" of marking points on the time axis.
- The markings vary in length.
- There is a multiple reciprocal overlapping of the markings.

The first problem in particular makes it impossible to account for agreement of segmentations simply by looking for (temporal or video frame-based extents of) overlaps between the items identified by different observers. Hence, we follow the reliability assessment strategy worked out by Thomann [73] and calculate agreement in the "marking of items"-setting in terms of clusters of markings. The procedure and its implementation is described in Lücking et al. [46]. Roughly, the rationale is as follows: picture the markings of various observers as segments on a time line. If all markings are laid on top of each other, the regions in which all or at least most observers identify items appear as accumulations of segments, i.e. as clusters. The higher the "nearness" of clustering of markings, the higher the degree of agreement. Normalizing clusters against their random baseline results in the respective segmentation agreement coefficient, called *degree of organization* (DoH), which can take values within the interval $(-1, 1)$. The DoH values for the main gesture phases are given in Table 3. We calculated the DoH for each phase separately, since they are relevant to semantic (stroke) and timing (preparation, retraction) aspects of speech-and-gesture alignment (cf., e.g., the synchrony rules of McNeill [48]). With a mean value of 0.7548 they are substantially better than what would be expected by accidental coincidence of segmentations.

In sum, the evaluation of the secondary data of the SaGA corpus reveals a satisfactory degree of reliability. Chance-corrected agreement on Type II data surpasses the self-set threshold of 70 %. Observed interrater agreement on Type I data results in angular values which, by and large, reveal rather harmless dissent between annotators. Agreement for gesture segmentation also reveals quite a large degree of shared understanding of "gesture" among the annotators. Hence, the SaGA corpus provides a reproducible data base which can be exploited for empirically driven research.

3 Putting SaGA to use

3.1 Empirical analysis: what shapes iconic gesture use?

According to the predominant Peircean view, iconic gestures communicate through iconicity, i.e., their physical form is

Table 3 The degrees of organization for gesture segmentation

	Preparation	Stroke	Retraction
Left hand	0.75780	0.64062	0.91494
Right hand	0.68033	0.64865	0.88657

said to correspond with object features such as shape or spatial properties (cf. [48] and Sect. 1). In that respect, they contrast to language or other gesture types such as emblems, whose meaningfulness is grounded in a conventionalized form-meaning mapping. However, just like concrete utterances are not fully determined by their conventional underpinning [there are, amongst other, a wealth of pragmatic (e.g. [24]), psychological (e.g. [64]) and statistical (e.g. [38]) influences on language use], the particular form of an iconic gesture token is not exclusively shaped by similarity with a referent *qua* representation technique (cf. Sect. 1 above). Rather, iconic gestures are also influenced by their linguistic context and by the speakers' individual gesture style. In the following, we will review the major results of our SaGA corpus-based studies.

Contextual factors Especially the choices[6] whether a gesture is produced or not and which representation technique to use are subject to linguistic and discourse-contextual factors [8,14]. As concerns the former, we found gestures to be predominantly produced for rhematic and private information. Further, the linguistic context is influential: particular syntactic noun phrases are much more likely to be accompanied by gestures than others. And finally, it has an impact whether a speaker performed a gesture beforehand, or whether the hands were in a rest position. Concerning the use of representation techniques, our corpus analysis revealed that depicting gestures (shaping, drawing, posturing) are preferred in descriptive utterances, while the spatial arrangement of entities is typically accompanied by localizing gestures (indexing and placing). Moreover, different gestural representation techniques co-occur with certain noun phrase patterns in a significant way, and individual speakers tend to stay in the same technique.

Gesture forms in different representation techniques To investigate how different gesture form features are used and combined, we explored the SaGA data separately for each representation technique [4]. This investigation revealed novel and corpus-based insights into the structure of gesture techniques. On the one hand, we found that techniques are characterized by different *technique-specific patterns*. For instance, drawing gestures—in contrast to gestures of other representation techniques—were found to be distinctive as they were performed predominantly with one hand only, with the pointing handshape ASL-G and with downwards oriented palms. On the other hand, the selectivity of representation techniques with regard to the iconic representation domain is also found empirically. In indexing gestures, for instance, handedness is sensitive to the position of the gesture's referent (in accordance with Pfeiffer [53, p. 141]) while other form features have technique-specific characteristics. In shaping or drawing gestures, by contrast, shape features of the referent were found to be decisive for the trajectory of wrist movement. In sum, each technique was found to be characterized by particular conventional aspects as well as iconic aspects.

Inter-individual differences Analyzing the individual differences in the use of representation techniques revealed that individuals differ significantly in the way they gesture about the same thing, and these differences concern multiple decision levels involved in the process of gesture formation. First, individual speakers in the SaGA corpus differ obviously in how much they gestured, ranging from 2 gestures per minute up to 30 gestures per minute. Second, speakers differed considerably in their preference for particular techniques of representation. Finally, inter-individual differences were also found with regard to particular gesture form features. Especially handedness and handshape choice are subject to inter-individual differences.

3.2 Gesture typology work using SaGA

The data assembled in SaGA show that there are recurrent patterns in one subject's gesture. These patterns generalize to the gesture behaviour of other agents. So there seem to be "gesture dialects". However, there is considerable variation across gesturers—see Sect. 3.1 above. Variation ranges from the frequency of gesture use observed with agents or agents' preferences for certain representation techniques (like drawing, shaping or modelling) to the extent of gestures. As to extent, we have large versus small gestures, lap-oriented versus torso-oriented ones, different scalings etc. Recurrent gesture information was captured in the manual multi-modal annotation of the data. It was assembled in types, represented as typed feature structures and coded in AVMs (see [57]). Types are extracted manually considering which features and information packages enter larger informational structures and are used in different gesture contexts. Feature bundles using a combination of hand-shape, palm, BoH or wrist information are good candidates for a type as are feature bundles associated with "depicting gesture Gestalts" such as lines, flat regions or three-dimensional entities.

Comparing pointing gestures, line gestures and three-dimensional "box" gestures we observe that types can be ordered along dimensions and complexity: points exhibit zero dimension, lines are one-dimensional, signs for flat surfaces two-dimensional, gestures for containers three-dimensional etc. "Line gestures" are lines drawn in gesture space using the extended index finger; frequently these "depict" routes, directions or edges of three dimensional objects; "box gestures" are formed with both hands indicating a container shaped like a box. In addition, there are all sorts of composite cases, for example lines touching a cylinder, two bent lines forming a type of twisted angle and so on. See Fig. 2 below for an illustration also indicating the importance of handedness and mixed cases. Since there exists an

[6] The term 'choice' is not meant to imply a conscious process here.

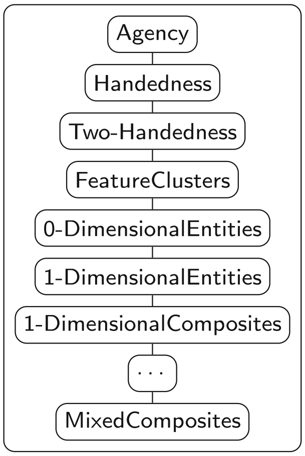

Fig. 2 Structure of the typology matrix

ordering along a scale of complexity and since there is a finite source of shapes or features for all gesturing, an inheritance hierarchy can be established for the gestures going into a full-fledged typology: the more basic types pass their information "further down" to other types. Looking at the range of gestures observed in SaGA, we found out that some are fairly well known and researched into, for example beats, emblems, pointing gestures and iconic gestures (cf. [48, p. 76] for these types). In addition, we have "quantificational" ones such as for numbers or "all" which, according to the "counting" category defined in one of our manuals (see Sect. 2.1 above), also qualify as iconic gestures. Some gestures seem to operate on propositional contents in the role of "modifiers". Furthermore, gestures can have pragmatic functions as shown by gestural "denials", others may even express full illocutionary acts, acting like constants. A new finding we have is that there are gestures related to turn allocation and dialogue structure, for example, self-selection of next speaker or next speaker selection by current speaker can both be done by pointing (see also the remarks on "interaction regulation" by means of pointing gestures by Kendon [35]). Finally, we have complex gestures produced by several agents in successive turns collectively. In sum, we see that there is evidence that gesture is not a mere construction-related or even propositional phenomenon but permeates all levels of conversation. All of these gesture shapes are based on a handful of typologically relevant motor behaviours.

The gain of a systematic gesture typology is manifold: The types factored out substantiate the very notion of a gesture morphology: we see that gestures are built out of regular, stable parts, the aforesaid points, lines and so on. In the end, we use the types isolated for computational gesture generation and gesture understanding. Concerning computation, we can establish a finite set of gesture building blocks to be used as a generating device triggering simulated motor behaviour. A study taking up this research line is Bergmann et al. [11]. On the understanding side we can associate conventionalised descriptions of partial ontology providing a gesture's meaning with these gesture types, whose interface with verbal meaning is conventionalised.

Our typological work was at first restricted to pointing and iconic gestures co-occurring with noun phrases [26,57]. In the sequel we developed classifications of gestures indicating dialogue structure [28] and presently we work on gestures related to full verb phrases [60].

3.3 A computational model of gesture production

To generate gesture forms from a given representation of content we have proposed *GNetIc*, a gesture net specialized for iconic gestures [5]. These networks implement the representation technique-based form-meaning relationship as described in Sect. 3.2, and even go beyond it in that they account for the empirical findings which indicate that a gesture's form is also influenced by specific contextual constraints like linguistic or discourse contextual factors (e.g., information structure, communicative goals, or previous gesture use of the same speaker) as well as obvious inter-individual differences (see Sect. 3.1). We employ a formalism called Bayesian decision networks (BDNs)—also termed *Influence Diagrams* that supplement standard Bayesian networks by decision nodes [5,7]. This formalism provides a representation of a finite sequential decision problem, combining probabilistic and rule-based decision-making. We are, therefore, able to specify rules for the mapping of meaning onto gesture forms and at the same time we can reconstruct data-based patterns in terms of probability distributions.

GNetIc provides a feature-based account of gesture generation, i.e., gestures are represented in terms of characterizing features as their representation technique and form features which correspond to those covered by the gesture typology (see Sect. 3.2). These make up the *outcome* variables in the model which divide into chance variables quantified by conditional probability distributions in dependence on other variables, ('gesture occurrence', 'representation technique', 'handedness', 'handshape'), and decision variables that are determined in a rule-based way from the states of other variables ('palm orientation', 'BoH orientation', 'movement type', 'movement direction'). Factors which potentially contribute to these choices are considered as input variables. So far, three different factors have been incorporated into this model: linguistic/discourse context (communicative goals, information structure, thematization, noun phrase type), features characterizing the previously performed gesture, and features of the referent (shape properties, symmetry, number of subparts, main axis, position).

The probabilistic part of the network is learned from the SaGA corpus data by applying machine learning techniques. The definition of appropriate rules in the decision nodes is based on our theoretical considerations of the

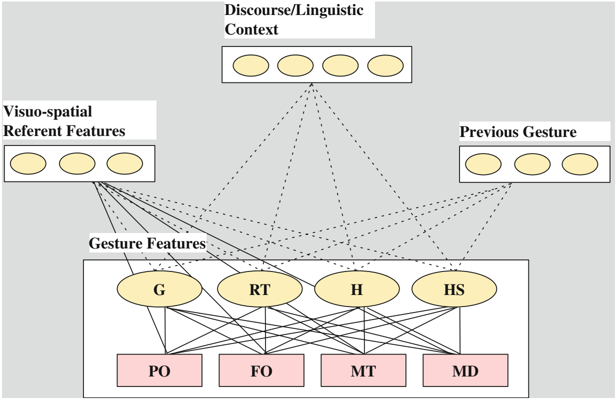

Fig. 3 Schema of a GNetIc network

meaning-form relation via gestural representation techniques and our corpus-based analysis of these techniques. That is, depending on the very representation technique, gesture form features are defined to be subject to referent characteristics as well as other gesture form features. See Fig. 3 for the generation network schema. With a set of rules the values for palm and BoH orientation, movement type, and movement trajectory of a gesture are determined. With respect to representation technique-specificity, the rules account, e.g., for the fact that drawing gestures are typically performed with a downwards palm orientation and fingers oriented away from the speaker's body. In addition, regarding movement type, the referent-characteristic shape properties are depicted.

Currently, the system has the ability to simulate five different speakers from the SaGA corpus by switching between the respective decision networks built as described above. These five speakers have been chosen because they gestured at a relatively high rate while giving the object descriptions so that the amount of data was sufficient to build the probabilistic parts of the networks. A total of 288 gestures (473 noun phrases) was used for this purpose.

The GNetIc model was finally evaluated in two ways. First, in comparison with empirically observed gestural behavior, the model was shown to be able to successfully approximate human use of iconic gestures, especially when capturing the characteristics of individual speakers' gesture style [7]. Second, when brought to application in a virtual agent, the generated gestural behavior was found to be positively rated by human recipients [10]. In particular, individualized GNetIc-generated gestures could increase the perceived quality of object descriptions. Moreover, the virtual agent itself was rated more positively in terms of verbal capability, likeability, competence, and human-likeness.

4 Gesture and unification-based grammar

Unification-based grammars make up leading formal frameworks for formulating grammars for natural languages. *Head-Driven Phrase Structure Grammmar* (HPSG [56,63]), *Lexical Functional Grammar* (LFG [19]), *Sign-Based Construction Grammar* (SBCG [62]), and *Fluid Construction Grammar* (FCG [65]) all rely on Attribute-Value Matrices (AVMs) as representation format and the unification of typed feature structures as central mechanism. Due to their prominent status and the flexibility of AVM representations, unification-based grammars are a first choice for formulating a speech-gesture interface.

4.1 AVM representations for gestures

Using AVMs as a representation format for gestures goes back at least to the work on a multimodal unification-based grammar as part of a handheld pen-input interface by Johnston [32]. However, Johnston's gesture representations are rather impoverished, they are strictly limited by the restricted pen-input application domain. In the later work of Kopp et al. [39], attribute-value pairs for representing a gesture are oriented at the "morphology" of gestures. The architecture of such gesture AVMs is straightforward: the orientations and movements of palm, back of hand and wrist make up features that are specified by appropriate orientation and movement values (including the empty value for, say, static gestures). The feature HANDSHAPE is assigned an (perhaps modified) ASL hand shape name. Such a morphology-driven AVM representation is used in the HPSG accounts of Lascarides and Stone [42] and Alahverdzhieva and Lascarides [1]. The more detailed SaGA gesture annotation can be readily linked to such a feature structure representation (see [43,57]).

4.2 Multimodal parsing

Johnston [32] defined and implemented a multimodal chart parser. This multimodal parser processes input on two different input streams and conjoins them into a combined structure. The constraint that licenses the junction is a temporal one: a gesture that overlaps with a word or follows it not more than four seconds can be combined into a multimodal structure. This constraint is backed by the phonological synchnrony rule formulated, for instance, by McNeill [48]. In this vein, Alahverdzhieva and Lascarides [1] and Lücking [43] proposed an intonation constraint: a gesture relates to a stressed verbal element in an overlapping time interval.

4.3 Multimodal linking

In addition to a bimodal, phonetically driven parser and AVM representations for gestures, we assume that the gesture morphology relates to a typology of entities of mixed complexity via the iconic mapping *rep*—see Sect. 3.2 above. For example, *rep* maps the hand shape *bent-B* of a static left hand onto

$$\begin{bmatrix} \textit{left-hand-gesture} & \\ \text{MORPH} & \begin{bmatrix}\text{HANDSHAPE} & \text{bent-B}\end{bmatrix} \\ \text{V-CON} & \begin{bmatrix} \textit{word} & \\ \text{PHON} & \text{stressed} \\ \text{CONT} & \begin{bmatrix}\text{INDEX} & i\end{bmatrix}\end{bmatrix} \\ \text{REP} & \left\langle \begin{bmatrix} \text{TYP} & \text{left-curved-side} \\ \text{INST} & s \\ \text{ARG} & i \\ \text{PERSP} & g \end{bmatrix} \right\rangle \end{bmatrix}$$

Fig. 4 Feature structure representation of a *bent-B* gesture

left-curved-side(s, i, g), where s is the side, i is the thing s is a side of, and g is the gesturer, a parameter which is needed to account for the perspectivity of "left". A representation of the characteristics of the *bent-B* gesture by means of an AVM is given in Fig. 4, where the *rep* function is captured in terms of the feature REP.

The *left-hand-gesture* is to be read as follows: the morphology (MORPH) of the simple static gesture introduces the hand shape value. The gesture's verbal connector, V-CON, has to be a stressed word. The entity represented by the gesture is part of the gesture's REP list. There, the typology entry is given as the value of TYP. The semantic relation between the gesture and its verbal connector is established via linking, that is, by the sharing of indices. The INDEX of the word is the ARGument of the gesture representation—see index i. Furthermore, the representation is relative to a certain perspective g, which has to be fixed contextually.

Gesture structures like the one from Fig. 4 can be "plugged into" a multimodal HPSG framework as developed by Alahverdzhieva and Lascarides [1] or Lücking [43], giving rise to a grammar modelling of speech and gesture integration.

5 Work on gestures supporting dialogue structure and interaction in SaGA

The experimental context for SaGA shows a kind of "Russian Doll structure" with respect to embedding. This forms a sort of precondition for gestures used in dialogical exchange: We have the route context making use of the conversational participants' (CPs') gesture spaces. It contains the topical or baseline information dealing with the task. In addition, there is the larger embedding context of the experimental situation (see Fig. 5) and a still larger one in which we have the physical VR environment, the institute, the University or Bielefeld town, etc. The discourse-related gestures observed can be grouped roughly into gestures used in turn allocation, feed-back gestures in second turn, those indicating assessment of evidence, gestures serving to highlight information, sequences of quick feed-back or monitoring gestures tied to sub-propositional contributions and, finally, truly interactive gestures exclusively social in character. All of these are accompanying speech. For these we have developed a manual (cf. [27]). Data descriptions can be found in Hahn and Rieser [28] on which this report is based and in Rieser [58].

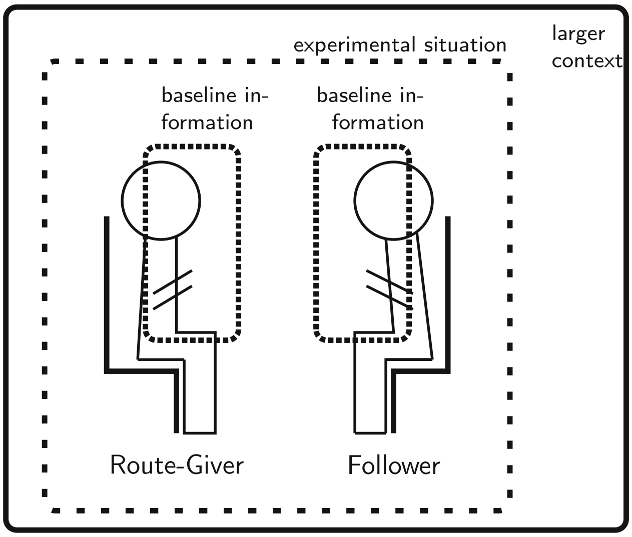

Fig. 5 Embeddedness

Here is a synopsis of the main findings: First we treat gestures related to turn allocation. With Sacks et al. [61] we assume a regularity for turn allocation in dialogue depending on the larger speech-exchange system: Current speaker selects next. If next is not provided this way, one of the other speakers self-selects. This option omitted, the first speaker may continue. Similar to observations already made in Clark [17], the SaGA data show that there is considerable freedom in this schema. It does leave room for quick interrupts of other. These become acceptable for CPs if motivated locally and related to the common purpose. Turn-related gestures exploit the "Russian Doll" property of the situation: current speaker points to other to select her as next. Somewhat surprisingly, indexing other to select self is also a possibility observed in roughly the same types of contexts. Further, matching a class of specific verbal contexts, current speaker may invite a contribution from other using a "gimme mo" gesture. In addition, current speaker may ask to be given a larger amount of time to carry on or complete his contribution. He may indicate a lapse to be tolerated by other and use a finger-to-lip or finger-below-lip gesture to express that. In tightly coordinated discourse there is an interesting "attack-ward-off pair": other may indicate that he wants to contribute at a non-turn-transition relevance place. Discouraging that, current speaker may try to fence him off with a characteristic posture. In the end, current speaker may yield and offer a "go ahead" for the intruding CP. Speaker of a second turn may use an iconic gesture of previous speaker in order to indicate acknowledgement or accept (see Bergmann et al. [11] for a simulation account of this phenomenon). As with the indexing, next speaker's gesture's imitation uses a topical gesture in a discourse function. Gestures can also indicate assessment of evidence—cf. the *hedging* category of gesture classification (introduced in Sect. 2.1 above). We observed two groups of gestures to indicate reliability of information. One is conveying doubt concerning the fit of a description, the other one is indicating an agent's epistemic state concerning a situation. Similarly, gestures can be used to highlight or to downgrade information: stressing information can also be suggested lifting a G-shaped hand, directing it against the addressee and moving it in a beat-like fashion. By contrast, we have the near-universal "brush-away" gesture indicating that information is considered to be not so relevant, a finding that supports the analyses by Teßendorf [71,72]. Perhaps our main finding was sequences of quick feed-back or monitoring gestures tied to sub-propositional contributions: CPs in near-to-natural task-oriented dialogue often converse quickly and in short thrusts. So we can have a Router's "don't interrupt" followed by the Follower's "let me interrupt" and, finally, the Router's acknowledgement and a "go ahead" gesture. This shows that full-blown dialogue acts do not always matter. From all of these we want to delineate gestures which are truly interactive such as hand and body postures to mollify someone or touching or caressing him. Here we have "calming down" and "don't bother" gestures.

Another phenomenon of gesture use in dialogue is *gestural alignment*. Here, two directions of alignment that involve gestures have to be distinguished, namely alignment between gesture and speech on the one hand and alignment between gesture and gesture on the other hand. Concerning the former case, viz. speech-and-gesture alignment, a first attempt to attest evidence for a mutual adaptation between both modalities has been undertaken by Lücking et al. [44]. Starting from the notion of *multimodal ensembles*, that are couples of gestures and their verbal affiliates Kendon [36], it has been found that the use of gestures has an influence on the distribution of words in multimodal discourse in such a way that the distributional patterns that governs monomodal spoken dialogue gets slightly distorted. Note, that this approach therefore provides a measurement procedure for speech-gesture alignment. Building on that work, the classification of multimodal ensembles has been formalized as a machine learning task by making use of the notion of *cross-modal alignment* in Mehler and Lücking [49]. In ongoing work, we assess cross-modal alignment in terms of a network model for measuring alignment in dialogue developed by Mehler et al. [50]. This model is the first one that goes beyond surface repetition in expressing and capturing alignment phenomena.

Concerning gesture–gesture alignment on the other hand, we recently found—in a first systematic study of gesture form convergence based on a large sample naturalistic dialogue data—that gesture use is also subject to inter-speaker influences. In other words, we found evidence for *gestural* alignment [9]. Remarkably, not all gesture features seem to be subject to this effect. While the form features 'wrist movement' and 'finger orientation' seem resistant to these contingencies, we found that the use of particular gestural representation techniques as well as the gesture form features 'handshape', 'handedness' and 'palm orientation' are significantly subject to inter-speaker convergence effects. In a detailed analysis of those sensitive features we addressed the question whether intra-speaker or inter-speaker influences on gesture form are stronger: for all features under consideration, alignment effects were found to be significantly stronger within speakers than across speakers. That is, same speaker's gestures influence each other more than the gestures an interlocutor performs, notwithstanding the effectiveness of other-alignment. Further, we investigated how gestural alignment depends on the temporal distance between gestures. Here a multi-faceted picture emerged: alignment in 'handshape' and gestural representation techniques gets weaker with greater distance, while alignment in 'handedness' and 'palm orientation' remains constant. It will be discussed whether this heterogeneous picture of gestural alignment at the level of different features may be due to the fact that particular features are communicatively bound, i.e., more crucial for conveying intended meaning and less amenable for interpersonal coordination.

Results from corpus investigation concerning dialogue moves have been modelled in multi-modal dialogue theory [55,58] and in simulation of multi-modal dialogue effects [11]. In ongoing work we extend our modelling accounts to cover further dialogue phenomena of multi-modal communication.

6 Conclusion

In this paper we presented the Bielefeld SaGA corpus, a collection of naturalistic, yet content-controlled speech-gesture data. The data is systematically and completely annotated, the annotation being based on a grid developed to cover the semantic and pragmatic fulcrum of iconic gestures, especially hand-shape and movements of hand in the gesture space. It has been rated for practices like drawing and modelling and for the fine-grained gesture morphology, both yielded in the end a stable foundation for specifying the semantics of gestures. In order to support gesture semantics and to initiate work on speech-gesture interfaces, gesture occurrences were grouped into types adopting specific functions, for example lines which can indicate trajectories of movements or borders of surfaces or regions. We found out that use of gesture types varies with speakers and contexts as does gestural (as well as all linguistic) behaviour in general. One of our main research questions was "Can we get at multimodal meaning related to gesture meaning and verbal meaning as input?". It was investigated using unification based grammars and given different answers depending on the role attributed to gesture meaning. Working through SaGA also led to the discovery of non-familiar types of gestures supporting dialogue structure and interaction regularities. Most importantly, machine-learning treatment of annotated SaGA structures led to the development of a talking and synchronically gesturing avatar. The gestures it produced have in turn been evaluated using model theory mapping them unto gesture occurrences in SaGA and finally submitted to judgements of human observers estimating naturalness and other social parameters [59]. Especially work along these lines is expected to be helpful in developing human-robot interfaces to facilitate communication using gestures on one or both sides. Investigation of SaGA structures and its various implementations have so far been mainly concentrated on dealing gesturally with objects, especially landmarks. As a consequence, dynamics in gesture execution did not play a dominant role. We now started to investigate gestures tied to the verbal constituents of utterances, for example, those indicating the direction of routes taken, the junction of roads or the flow of water from a fountain. From these investigations we hope to gain insight into the topology of the speakers' gesture spaces and to develop experimental tools for measuring them using body trackers. A further hope is that these measurements can be used to supplant the naïve observational annotation categories applied now and that this will in the end lead to developing avatars equipped with richer and more spontaneous gesture behaviour.

Acknowledgments This work has been supported by the German Research Foundation (DFG) and has been carried out in the CRC 673 "Alignment in Communication".

References

1. Alahverdzhieva K, Lascarides A (2010) Analysing language and co-verbal gesture in constraint-based grammars. In: Müller S (ed) Proceedings of the 17th international conference on head-driven phase structure grammar (HPSG), Paris, pp 5–25
2. Alibali M (2005) Gesture in spatial cognition: expressing, communicating, and thinking about spatial information. Spatial Cogn Comput 5:307–331
3. Allwood J, Cerrato L, Jokinen K, Navarretta C, Paggio P (2007) The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. Language Resources and Evaluation 41:273–287
4. Bergmann K (2012) The production of co-speech iconic gestures: empirical study and computational simulation with virtual agents. PhD thesis, Faculty of Technology, Bielefeld University
5. Bergmann K, Kopp S (2009) GNetIc—using Bayesian decision networks for iconic gesture generation. In: Ruttkay Z, Kipp M, Nijholt A, Vilhjalmsson H (eds) Proceedings of the 9th international conference on intelligent virtual agents, Springer, Berlin, pp 76–89
6. Bergmann K, Kopp S (2009) Increasing expressiveness for virtual agents—autonomous generation of speech and gesture in spatail description tasks. In: Decker K, Sichman J, Sierra C, Castelfranchi C (eds) Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems, Budapest, Hungary, pp 361–368
7. Bergmann K, Kopp S (2010) Modeling the production of co-verbal iconic gestures by learning Bayesian decision networks. Appl Artif Intell 24(6):530–551
8. Bergmann K, Kopp S (2010) Systematicity and idiosyncrasy in iconic gesture use: empirical analysis and computational modeling. In: Kopp S, Wachsmuth I (eds) Gesture in embodied communication and human-computer interaction. Springer, Berlin, pp 182–194
9. Bergmann K, Kopp S (2012) Gestural alignment in natural dialogue. In: Proceedings of the 34th annual conference of the cognitive science society, CogSci 2012 (forthcoming)
10. Bergmann K, Kopp S, Eyssel F (2010) Individualized gesturing outperforms average gesturing—evaluating gesture production in virtual humans. In: Allbeck J, Badler N, Bickmore T, Pelachaud C, Safonova A (eds) Proceedings of the 10th conference on intelligent virtual agents. Springer, Berlin, pp 104–117
11. Bergmann K, Rieser H, Kopp S (2011) Regulating dialogue with gestures—towards an empirically grounded simulation with conversational agents. In: Proceedings of the SIGDIAL (2011) Conference. Association for Computational Linguistics, Portland, Oregon, pp 88–97
12. Blache P, Bertrand R, Ferré G (2009) Creating and exploiting multimodal annotated corpora: the toma project. In: Kipp M, Martin JC, Paggio P, Heylen D (eds) Multimodal corpora. Lecture Notes in Computer Science, vol 5509. Springer, Berlin, pp 38–53
13. Blache P, Bertrand R, Bigi B, Bruno E, Cela E, Espesser R, Ferré G, Guardiola M, Hirst D, Magro EP, Martin JC, Meunier C, Morel MA, Murisasco E, Nesterenko I, Nocera P, Pallaud B, Prévot L, Priego-Valverde B, Seinturier J, Tan N, Tellier M, Rauzy S (2010) Multimodal annotation of conversational data. In: Proceedings of the fourth linguistic annotation workshop, LAW IV '10, pp 186–191
14. Buschmeier H, Bergmann K, Kopp S (2010) Adaptive expressiveness—virtual conversational agents that can align to their interaction partner. In: Proceedings of the 9th international conference on autonomous agents and multiagent systems, Toronto, Canada, pp 91–98
15. Calbris G (1990) The semiotics of French gesture. Indiana University Press, Bloomington (Advances in Semiotics)
16. Carletta J (1996) Assessing agreement on classification tasks: the kappa statistic. Comput Linguist 2(22):249–254
17. Clark HH (1996) Using language. Cambridge University Press, Cambridge
18. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46
19. Dalrymple M (2001) Lexical functional grammar, syntax and semantic. Academic Press, New York
20. Denis M (1997) The description of routes: a cognitive approach to the production of spatial discourse. Curr Psychol Cogn 16:409–458

21. Eco U (1976) A theory of semiotics. Indiana University Press, Bloomington
22. Gibbon D, Gut U, Hell B, Looks K, Thies A, Trippel T (2003) A computational model of arm gestures in conversation. In: Proceedings of the 8th European conference on speech communication and technology. EUROSPEECH 2003, pp 813–816
23. Gibbon D, Gut U, Hell B, Looks K, Milde JT, Thies A, Trippel T (2004) CoGesT: a formal transcription system for conversational gesture. In: Proceedings of LREC 2004
24. Grice HP (1975) Logic and conversation. In: Cole P, Morgan JL (eds) Syntax and semantics. Speech Acts, vol 3. Academic Press, New York, pp 41–58
25. Gwet K (2001) Handbook of inter-rater reliability. STATAXIS Publishing Company, Gaithersburg
26. Hahn F, Menke P (2011) Corpus-driven development of a gesture typology based on the Bielefeld speech and gesture alignment corpus. In: Corpus Linguistics 2011, Birmingham, UK
27. Hahn F, Rieser H (2009–2011) Dialogue structure gestures and interactive gestures, Annotation Manual. CRC 673, Alignment in Communication. Working Paper
28. Hahn F, Rieser H, (2011) Gestures supporting dialogue structure and interaction in the Bielefeld speech and gesture alignment corpus (SaGA). In: Proceedings of SEMdial 2011, Los Angelogue, 15th workshop on the semantics and pragmatics of dialogue. Los Angeles, California, pp 182–183
29. Halliday M (1967) Notes on transitivity and theme in English (part 2). J Linguist 3:199–247
30. Harthshorne C, Weiss P (eds) (1965) Collected papers of Charles Sanders Peirce. Harvard University Press, Cambridge, MA, repr. from 1932
31. Herskovits A (1986) Language and spatial cognition: an interdisciplinary study of the prepositions in English. Cambridge University Press, Cambridge
32. Johnston M (1998) Unification-based multimodal parsing. In: Proceedings of the 36th Annual Meeting on Association for Computational Linguistics—Volume I, Annual Meeting of the ACL, Association for Computational Linguistics, Montreal, Quebec, Canada, pp 624–630
33. Kendon A (1980) Gesticulation and speech: two aspects of the process of utterance. In: Key MR (ed) The relationship of verbal and nonverbal communication. Contributions to the sociology of language, vol 25. Mouton Publishers, The Hague, pp 207–227
34. Kendon A (1988) How gestures can become like words. In: Poyatos F (ed) Cross-cultural perspectives in non-verbal communication. Hogrefe, Toronto, pp 131–141
35. Kendon A (1996) An agenda for gesture studies. Semiot Rev Books 7(3):8–12
36. Kendon A (2004) Gesture—visible action as utterance. Cambridge University Press, Cambridge
37. Kipp M, Neff M, Albrecht I (2007) An annotation scheme for conversational gestures: how to economically capture timing and form. J Lang Resour Eval Special Issue Multimodal Corpora 41(3–4):325–339
38. Köhler R, Altmann G (2000) Probability distributions of syntactic units and properties. J Quant Linguist 7(3):189–200
39. Kopp S, Tepper P, Cassell J (2004) Towards integrated microplanning of language and iconic gesture for multimodal output. In: Proceedings of the international conference on multimodal interfaces (ICMI'04), ACM Press, pp 97–104
40. Krifka M (2008) Basic notions of information structure. Acta Linguist Hung 55(3–4):243–276
41. Krippendorff K (1980) Content analysis. The SAGE KommTexT Series, vol 5. SAGE Publications, Beverly Hills
42. Lascarides A, Stone M (2006) Formal semantics of iconic gesture. In: Schlangen D, Fernández R (eds) Proceedings of the 10th workshop on the semantics and pragmatics of dialogue, Universitätsverlag Potsdam, Potsdam, Brandial'06, pp 64–71
43. Lücking A (2011) Prolegoma zu einer Theorie ikonischer Gesten. PhD thesis, Faculty of Linguistics and Literary Studies, Bielefeld University
44. Lücking A, Mehler A, Menke P (2008) Taking fingerprints of speech-and-gesture ensembles: approching empirical evidence of intrapersonal alignmnent in multimodal communication. In: LonDial 2008: The 12th workshop on the semantics and pragmatics of dialogue (SEMDIAL), King's College London, pp 157–164
45. Lücking A, Bergmann K, Hahn F, Kopp S, Rieser H (2010) The Bielefeld speech and gesture alignment corpus (SaGA). In: Kipp M, Martin JC, Paggio P, Heylen D (eds) LREC 2010 workshop: multimodal Corpora—advances in capturing, coding and analyzing multimodality, pp 92–98
46. Lücking A, Ptock S, Bergmann K (2012) Assessing agreement on segmentations by means of *Staccato*, the *segmentation agreement calculator according to Thomann*. Gesture in embodied communication and human–computer interaction revised selected papers of the 9th international gesture workshop, GW 2011, Athens, Greece, May 25–27, 2011
47. Martell C, Osborn C, Friedman J, Howard P (2002) FORM: a kinematic annotation scheme and tool for gesture annotation. In: Proceedings of multimodal resources and multimodal systems evaluation, Las Palmas, Spain, pp 15–22
48. McNeill D (1992) Hand and mind—what gestures reveal about thought. Chicago University Press, Chicago
49. Mehler A, Lücking A (2009) A structural model of semiotic alignment: the classification of multimodal ensembles as a novel machine learning task. In: Proceedings of IEEE Africon 2009, IEEE, Nairobi, Kenya, 23–25 Sept
50. Mehler A, Lücking A (2010) A network model of interpersonal alignment in dialogue. Entropy 12(6):1440–1483
51. Menke P, Mehler A (2010) The ariadne system. A flexible and extensible framework for the modeling and storage of experimental data in the humanities. In: Proceedings of the seventh conference on international language resources and evaluation, LREC' 10, pp 944–948
52. Müller C (1998) Redebegleitende Gesten. Kulturgeschichte—Theorie—Sprachvergleich, Körper—Kultur—Kommunikation, vol 1. Berlin Verlag, Berlin
53. Pfeiffer T (2011) Understanding multimodal deixis with gaze and gesture in conversational interfaces. Shaker Verlag. http://www.shaker.de/shop/978-3-8440-0592-9
54. Pickering MJ, Garrod S (2004) Toward a mechanistic psychology of dialogue. Behav Brain Sci 27(2):169–190
55. Poesio M, Rieser H (2009) Anaphora and direct reference: empirical evidence from pointing. In: Edlund J, Gustafson J, Hjalmarsson A, Skantze G (eds) Proceedings of the 13th workshop on the semantics and pragmatics of dialogue, Royal Institute of Technology (KTH). Stockholm, Sweden, pp 35–43
56. Pollard C, Sag IA (1994) Head-driven phrase structure grammar. Chicago University Press, Chicago
57. Rieser H (2010) On factoring out a gesture typology from the Bielefeld speech-and-gesture-alignment corpus (SAGA). In: Kopp S, Wachsmuth I (eds) Proceedings of GW 2009, pp 47–60
58. Rieser H, (2011) Gestures indicating dialogue structure. In: Proceedings of SEMdial 2011, Los Angelogue, 15th workshop on the semantics and pragmatics of dialogue. Los Angeles, California, pp 9–18
59. Rieser H, Bergmann K, Kopp S (2012) How do iconic gestures convey visuo-spatial information? Bringing together empirical, theoretical, and simulation studies. In: Post-proceedings of GW 2011
60. Röpke I, Hahn F, Rieser H (2012) Gestures co-occuring with verb phrases in route-description dialogue (submitted)

61. Sacks H, Schegloff EA, Jefferson G (1974) A simplest systematics for the organization of turn-taking for conversation. Language 50:696–735
62. Sag IA (2011) Sign-based construction grammar: an informal synopsis. In: Boas H, Sag IA (eds) Sign-based construction grammar. Center for the Study of Language and Information—Lecture notes, chap. 2. University of Chicago Press, Chicago, pp 39–170
63. Sag IA, Wasow T, Bender EM (2003) Syntactic theory: a formal introduction. CSLI Publications, Stanford
64. Smith M, Wheeldon L (1999) High level processing scope in spoken sentence production. Cognition 73(3):205–246
65. Steels L (ed) (2011) Design patterns in fluid construction grammar. John Benjamins, Amsterdam
66. Stegmann J, Lücking A (2005) Assessing reliability on annotations (1): theoretical considerations. Tech. Rep. 2, SFB 360 Situierte Künstliche Kommunikatoren, Universität Bielefeld
67. Sternberg MLA (1981) American sign language: a comprehensive dictionary. Harper and Row, New York
68. Stevens SS (1958) Problems and methods of psychophysics. Psychol Bull 55(4):177–196
69. Stone M, Doran C, Webber B, Bleam T, Palmer M (2003) Microplanning with communicative intentions: the SPUD system. Comput Intell 19(4):311–381
70. Streeck J (2008) Depicting by gesture. Gesture 8(3):285–301
71. Teßendorf S (2007) Metonymy, metaphor, and pragmatic functions: a case study of a recurrent gesture. In: 3rd international conference of the international society for gesture studies. Evanston, IL, USA
72. Teßendorf S (2007) Pragmatic functions of gestures: the case of the 'brushing aside gesture' in spanish conversation. 10th International Pragmatics Conference, GÃuteborg, Sweden
73. Thomann B (2001) Oberservation and judgment in psychology: assessing agreement among markings of behavioral events. Behavior Res Methods Instrum Comput 33(3):339–248