The Bielefeld Speech and Gesture Alignment Corpus (SaGA)

Andy Lücking*, Kirsten Bergman*, Florian Hahn*, Stefan Kopp*, Hannes Rieser*†

*CRC 673 "Alignment in Communication", B1 "Speech-Gesture Alignment"
†CRC 673 "Alignment in Communication", X1 "Multimodal Alignment Corpora"
Bielefeld University
{Andy.Luecking|Kirsten.Bergmann|Fhahn2|Stefan.Kopp|Hannes.Rieser}@uni-bielefeld.de

Abstract

People communicate multimodally. Most prominently, they co-produce speech and gesture. How do they do that? Studying the interplay of both modalities has to be informed by empirically observed communication behavior. We present a corpus built of speech and gesture data gained in a controlled study. We describe 1) the setting underlying the data; 2) annotation of the data; 3) reliability evalution methods and results; and 4) applications of the corpus in the research domain of speech and gesture alignment.

1. Introduction

In face to face conversation, interlocutors co-produce language and gestures. The term 'gesture' refers to gesticulations according to Kendon's continuum (Kendon, 1980), that are spontaneous co-verbal hand and arm movements which are meaningful and contribute to the conversational participants' contributions. Both, words and gesture, are temporarily and semantically coupled so that they cohere into bimodal information units (McNeill, 1992). To put it in psycholinguistic terms: speech and gestures of a speaker are *aligned* (Pickering and Garrod, 2004). For the timespan of a dialogue they enter into crossmodal signs called *bimodal* or *multimodal ensembles* (Lücking et al., 2008). However, to date there is no systematic account for the division of labour between verbal and non-verbal means for their cooperative consitution of a common meaning.

We address this challenging topic 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 a precise modelling of the interface between speech and gesture. The implementation of theoretical models with computational means, on the other hand, enables us to simulate multimodal communicative behavior in virtual agents or robots. Both research lines necessitate a rich empirical basis in the form of a detailed and systematically annotated multimodal corpus. In Section 2 we present the Bielefeld Speech and Gesture Alignment (SaGA) corpus. We describe the primary experimental data as well as the secondary annotation data. Corpus evaluation in terms of interrater reliability is presented in Section 3. In order to compare the concordance of gesture performance transcriptions we distinguish two kinds of data types and apply chance-corrected agreement measures as well as a method developed in Bergmann and Kopp (2009a) that is based on the translations of annotation predicates into angle measures. Applications from linguistics and computer science that exemplify how the SaGA corpus is utilized in investigating and simulating the alignment of speech and gesture are given in Section 4.

(a) VR presentation phase

(b) Part of stimulus: church square with two churches

Figure 1: Experimental setting

2. The SaGA Corpus

The primary data of the SaGA corpus are made up of 25 dialogs of interlocutors which engage in a spatial communication task combining direction-giving and sight description. There is extensive evidence that "speakers gesture more when they talk about spatial topics than when they talk about abstract or verbal ones" (Alibali, 2005, p. 313). This scenario is, therefore, well-suited for systematically studying aspects of natural speech and gesture utterances used to communicate information about the shape of objects and the spatial relations between them. The stimulus is a model of a town presented in a Virtual Reality (VR) environment (see Figure 1(a)). The VR scenario affords better determination and experimental control of the content of multimodal messages. Additionally, it secures that all participants receive the same stimulus. Upon finishing a "bus ride" through the VR town along five landmarks (see for instance the church square with two churches in Figure 1(b)), a router explained the route as well as the wayside landmarks to an unknown and naïve follower.

2.1. Primary Data

Our primary data consists of 25 direction-giving dialogs. Audio- and videotapes were taken of each dialog. For the videotape, three synchronized camera views were recorded (see Fig. 2), as well as body movement data and eye-tracking data from the router. In total, the SaGA corpus consists of 280 minutes video material containing 4961 iconic/deictic gestures, approximately 1000 discourse ges-

Figure 2: Experimental dialogue situation from three camera views, capturing the router (left), the follower (right) and the dialog scenario as a whole (middle).

tures and 39,435 words. To our knowledge, this is the largest and most comprehensive collection of naturalistic, yet controlled, systematically annotated (see below) speech-gesture data currently available.[1] Our multimodal dialogue data are stored, retrieved, and transformed within the Ariadne system (Gleim et al., 2007) which is used as an *Alignment Corpus Management System*.

2.2. Secondary Data

The data has been completely and systematically annotated[2] based on an annotation grid that has been developed according to theoretical considerations and refined in pilot annotation sessions.

All gestures have been segmented in order to specify the *stroke* phase (Kendon, 1980). The gestures (i.e., strokes) are then typed for belonging to a certain kind, namely *deictic*, *iconic* or *discourse*. 'Iconic', a term coined by McNeill (1992), who alludes to a Peircean trichotomy (Peirce, 1867), is an "umbrella term" (cf. Eco (1976)) that covers a variety of different signifying methods. (Müller, 1998), drawing on the work of (Wundt, 1911), sets up a more fine-grained classification of gestures according to the distinction of four *techniques of representation* on the ground of what the hands *do*. According to our domain of application we adopt or modify the sets of representation techniques as posited by (Müller, 1998; Kendon, 2004; Streeck, 2008). The classification of gestures within SaGA now distinguishes the following eight representation techniques:

(1) *Indexing*: pointing to a position within gesture space;
(2) *Placing*: an object is placed or set down within gesture space;
(3) *Shaping*: an object's shape is contoured or sculptured in the air;
(4) *Drawing*: the hands trace the outline of an object's shape;
(5) *Posturing*: the hands form a static configuration to stand as a model for the object itself.
(6) *Sizing*: indicating distances or sizes;
(7) *Counting*: iconic representation of a tally sheet;
(8) *Hedging*: a depiction of uncertainty (typically by a wiggling or shrugging movement).

In addition, each gesture has been coded for its morphology consisting of handshape, wrist position, palm, back of hand orientation. Movement within any of these dimensions is coded in terms of movement features.

- To code *handshape* we use a modified ASL (American Sign Language) lexicon.
- *Palm orientation* is devoted in terms of the direction of an axis orthogonal to the palm, whereby the following six speaker-centric half-axes were used (Herskovits, 1986): forward, backward, left, right, up and down. Up to three of these basic values are combined to encode diagonal or mixed directions, e.g. 'up/right' or 'up/right/forward'. In order to capture palm movements it is possible to build a temporal sequence of these values by means of the ">"-concatenator. 'up>down', for instance, denotes an upwards-downwards movement sequence.
- 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". In addition, the extension of a gesture is specified *via* its distance to the gesturer's body.
- For dynamic gestures the *movement direction* is annotated in terms of the six cardinal directions in space. As already described for palm and back of hand orientation, combinations and sequences of the categories are used to describe directions in between the six basic values as well as temporal sequences.
- To further classify the type of movement trajectory, we distinguish between linear and curved movements. Assume, for instance, the sequence of orientations 'up>right>down>left'. If it is performed linearly, the resulting trajectory is a square whereas it would be a circle if the same sequence would be performed in curved fashion.

We also transcribed interlocutors' speech on the level of words. The dialogs of the corpus are 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. Each clause then is annotated by its associated communicative goal. Denis (1997) developed several categories of communicative goals that can be distinguished in route directions. We revised and refined these for our purposes into four categories: (1) Naming a landmark; (2) Landmark property description; (3) Landmark construction description; or (4) Landmark position description.

Following Halliday (1967) we distinguish the thematization structuring of clauses in terms of *theme* and *rheme*. Additionally, the information foci *given*, and *new* are annotated and, borrowing the terminology of (Stone et al., 2003), classified according to the information states 'private' and 'shared'

The gestures of a subset of seven dialogs have also been annotated semantically. Gestures used in object descriptions have been coded for the descriptions referent and some of the referent's spatio-geometrical properties. 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 unequivocally available for field data.

[1] There is a more multifarious collection of routes, though, hosted at the McNeill Lab (`http://mcneilllab.uchicago.edu/`) which comprises about 13 direction dialogs of different languages (English, Chinese, Huichol) – see also (McCullough, 2005) (McCullough, p.c.).

[2] We used Praat (www.praat.org) for speech transcription and Elan (www.lat-mpi.eu/tools/elan/) for gesture annotation.

3. Reliability Assessments

The annotation data has been evaluated in terms of interrater reliability. Here, a qualitative distinction has to be made, namely the distinction between Type I vs. Type II ratings (Gwet, 2001). Type I measurements are those where the human interpretation effort leading to a rating is well-understood and the outcome easily interpretable. To the contrary, this is not the case for measurements of Type II. Note that Type I ratings usually make up data on an interval or ratio scale, wheras Type II ratings are strongly associated with nominal scales. Accordingly, this difference has to be accounted for in evaluations of respective annotations: Type II ratings have to be adjusted for chance-based agreements (Cohen, 1960), whereas "chance" has no interpretation in Type I ratings. However, in the context of the latter but not the former one can speak of annotation errors. The gesture annotation comprises both types of annotation data, Type I and Type II. The classification of gestures in terms of representation techniques, reference ojects and dialogue context information is interpretive and therefore of Type II. The respective annotation labels are categories on a nominal scale. Descriptions of gesture morphology make up data of Type I. With one exception (hand shape, see below), the labels for annotating a gesture performance are ordered on an ordinal scale. Accordingly, we employ different methods in order to evaluate annotations of representation techniques and context information on the one hand, and annotations of gesture morphology on the other hand. 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 (2001). In order to assess the extent of association between annotations of the Type I gesture morphology, we employ an approach based on angle measures previously used by Bergmann and Kopp (2009a).

3.1. Type II Data.

In the run-up of the reliability study we set a reasonable agreement level of 70% with an α-error of 0.05 and a β-error of 0.85 for Type II annotations. The appropriate sample size of 477 gestures has been drawn from gesture annotations. The Type I morphology sample has been classified by four, the Type II technique sample by three annotators. The resulting first-order agreement coefficient AC1 for gestures' representation technique rating is 0.784. Its confidence interval is (0.758, 0.81). The proportion of agreement on gestures' representation techniques, given that the agreement is not due to chance, is significantly greater than 75%. In particular, this result complies with our reliability level initially demanded. The degree of reliability of the annotations of reference objects and context information was calculated for one dyad taken from the subset annotated for this information. The agreement coefficient AC1 for the classification of reference objects was 0.91, for information structure 0.95, for information state 0.86, and for communicative goal 0.88. All values are collected in Table 1. In sum, the highly interpretive Type II data show a reasonable degree of interrater reliability.

Technique	Referent	InfoStruc	InfoState	Goal
0.784	0.91	0.95	0.86	0.88

Table 1: Overview of Type II data reliability evaluation. Values denote AC1 coefficients.

3.2. Type I Data.

The annotations that make up the secondary Type I data of the SaGA corpus transcribe the movement of a gesture within gesture space – cf. the afore-mentioned annotation description. The gesture space is a three-dimensional region which is spanned over the saggital, transversal, and frontal planes of a speaker. The respective directions 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[3]. We address this problem by translating the annotation labels into angular measures which can be analyzed in terms of numeric differences. The smallest angular deviation is 2.36° for the movement direction of hand shapes, the biggest one is 46.16° for back of hand orientation. 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.

3.3. 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 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 agree on 83% (AC1 = 0.9, to give the Type II statistics for comparison) of the hand shapes within the reliability sample of gestures.

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, denote rather harmless dissent between annotators. Hence, the SaGA corpus provides a reproducible data base which can be exploited for empirically driven research.

[3] 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.

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)

Table 2: Overview of Type I data reliability evalution. 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 agreed Hand Shapes – for details, please consult the text.

4. Applications

So far, the SaGA corpus is put to use in two application domains. First, the gesture annotation is used to build an interpretive domain ontology, that is, an underspecified semantic representation of gesture morphology arranged in a typological grid. Second, the annotation data are used to trigger Baysian networks of gesture production as depending from semantic and discourse context factors. Both applications are shortly illustrated subsequently.

4.1. The Typological Grid Methodology

Considering SaGA, the question is: Are the gestures observed, lines, rectangles, the three-dimensional entities arising from them, idiosyncratic tokens or are they systematically used in one datum by two agents and throughout the whole SaGA corpus by many or even all agents? In order to investigate both these typological questions we set up a typological grid (Rieser, 2010) for one datum (SaGA video film 5) in the following way: gestures build a space consisting of hierarchies of simple and more complex morphological entities. The most basic properties we have are the individual annotation predicates like hand shape or palm orientation. For example, for a horizontal line we need the predicates hand shape, wrist movement and palm orientation. The annotation predicates' values are atoms of the gesture space, called features and represented in attribute-value matrices (AVMs). Only unified do these single bits of information describe a horizontal line, as represented by the following AVM (*RH* abbreviates "Right Hand", *FC* abbreviates "Feature Cluster"):

$$\begin{bmatrix} \textit{R-Line-RH} & \\ \text{R-FC-RH-1a} & \begin{bmatrix} \textit{R-FC-RH-1a-cat} & \\ \text{HandShape} & \textit{G} \\ \text{PalmOrient} & \textit{PDN} \\ \text{BoHOrient} & \textit{BAB} \end{bmatrix} \\ \text{R-FC-RH-2a} & \begin{bmatrix} \textit{WristMovement-RH-1a-cat} & \\ \text{PathofWrist} & \textit{Line} \\ \text{WristLocMovDirection} & \textit{MR or ML} \end{bmatrix} \end{bmatrix}$$

The single features form the most basic stratum of the gestural space and the kernel of our observational language. We also set up 0-dimensional entities originating from indexing which are considered to have no spatial extension. Lines come in different shapes and directions, straight, bent, horizontal, vertical and so on. They form the one-dimensional layer below the features and the theoretically motivated cluster layers. Similar to the line distinction, we have two-dimensional entities, rectangles, squares and so on, followed by three-dimensional entities such as cuboids, spheres or prisms. An interesting typological fact is that we get composites of n-dimensional entities, the most functionally conspicuous ones being lines touching circles orthogonally from the outside, horizontal and orthogonal lines meeting or two objects held and related to a previously introduced one – see Figure 3 for illustrations. Thus, the typological grid provides a compositional, semantic interpretation for gestures. The methodology applied here secures that even the most complex semantic features of gestures are strictly tied to annotation predicates. It is the systematic and fine-grained annotation of SaGA that makes this empirical backing of gesture meaning feasible. This in turn exposes the typological grid, and hence the reconstruction of gesture meaning, to Popperian falsifiability, a feature that sets this methodology apart from qualitative or interpretive inspections and exemplar-based analyses.

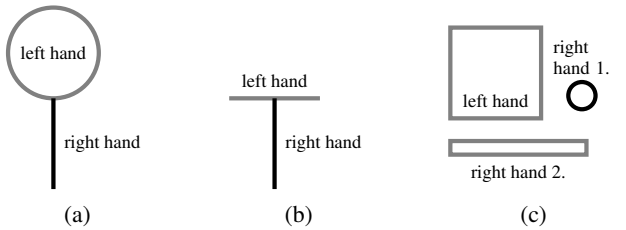

Figure 3: Illustrations of $n+m$-dimensional composite entities ($0 \leq n, m \leq 3$). Reproduced after (Hahn and Rieser, submitted).

The questions we have to investigate with respect to the grid are: How many features do we use, how many gestures of which dimensions do exist, how many composites of which dimensional parts are there and so on. Statistically based answers to these questions tell us which simple and complex gestural forms Router and Follower exploit. The following results emerged for the grid data (see (Hahn and Rieser, submitted)): Generally speaking, the Router concentrates on depicting routes, regions and locations as well as objects as (part of) landmarks. Composites consisting of $n \geq 2$ gestures provide the possibility to "hold" the landmarks and specify the route to them: at the same time both, landmark and route are relationally placed in Router's gesture space. Interestingly, the Follower sets up his interactive map using one-dimensional gestures most of his time. In other words, he concentrates on representing routes. For both, Router and Follower, the right hand is dominating when gesturing. The Router uses far more two-handed composites than the follower. He populates gesture space with more objects than the Follower does. Since gesture space functions as depictional model, his gesture space is more informative than the Follower's. A series of interest-

ing results emerged with respect to the "atoms" of the gesture hierarchy, the features and how they enter into clusters: The five features, HandShape, BoHOrientation, PalmOrientation, WristPosition, and WristMovementDirection are most frequently used by both Router and Follower in their left and and right hands, respectively. The annotationally motivated grouping of the features HandShape, BoHOrientation and PalmOrientation into feature cluster at the outset of the typological work thus gets statistical support. At the same time the large number of WristPosition features and WristLocationMovementDirection features motivates the set up of clusters for WristPosition and WristMovement. Both Router and Follower predominantly use their right hands. This can be seen from the greater number of feature clusters there.

4.2. Autonomous generation of speech and gesture

The SaGA corpus is also used as an empirical basis to model speech and gesture production. We have proposed an architecture that simulates the interplay between the two modes of expressiveness on two levels (Bergmann and Kopp, 2009b). First, two kinds of knowledge representations – propositional and imagistic – are utilized to capture the modality-specific contents and processes of content planning (i.e., what to convey). Second, specific planners are integrated to carry out the formulation of concrete verbal and gestural behavior (i.e., how to convey it). Of particular importance in this framework is the question how to generate gestural forms from an abstract representation. According to empirical results based on the SaGA corpus, iconic gesture generation on the one hand generalizes across individuals to a certain degree and these commonalities may pertain primarily to gesture's iconicity. On the other hand, inter-subjective differences must also be taken into consideration by an account of why people gesture the way they actually do (Bergmann and Kopp, 2010). Our research methodology to investigate this puzzle of iconic gesture production is based on computational modelling: we have proposed GNetIc (*Gesture Net for Iconic Gestures*), a probabilistic network to model decision-making in the generation of iconic gestures (Bergmann and Kopp, 2009a). Individual as well as general networks are learned from annotated corpora by means of automated machine learning techniques and supplemented with rule- based decision making. Three different types of factors are included in the network to influence the resulting gestures: (1) visuo-spatial referent features, (2) linguistic and discourse context, and (3) the previously performed gesture. A prototype of the generation model is employed in an architecture for integrated speech and gesture generation. 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. See Figure 4 for two simulation examples.

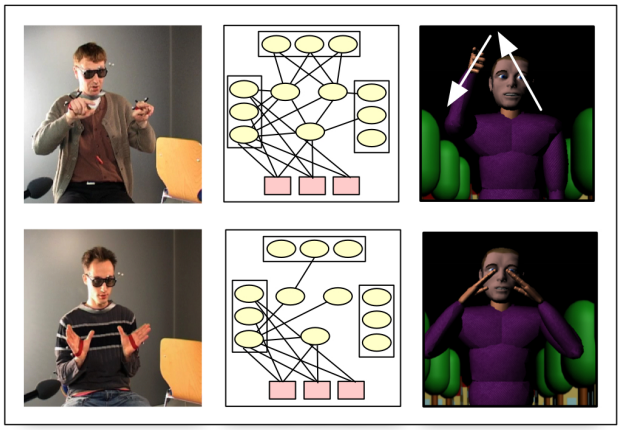

Figure 4: Two example networks (middle column) learned from individual speakers' data (left column) resulting in speaker-specific gesture production (right column). These gestures are simulated for the same referent (a tapered roof) in the same initial situation.

Analyzing the modelling results enables us to gain novel insights into the production process of iconic gestures: the resulting networks learned for individual speakers differ in their structure and in their conditional probability distributions, revealing that individual differences are not only present in the overt gestures, but also in the production process they originate from (as an example see the two differing networks in Figure 4 each of which learned from a particular speaker's data). Whereas gesture production in some individuals is, e.g., predominantly influenced by visuo- spatial referent features, other speakers mostly comply with the discourse context. So there seems to be a set of different gesture generation strategies from which individuals typically apply a particular subset.

In a comparison of learning algorithms for the network structure it turned out that at best 71.3% of the probabilistically modelled generation choices in individual networks could be predicted correctly. The accuracy achieved with general networks is 69.1%. Notably, all accuracy values clearly outperform the chance level baseline of 30%. The results show, by trend, that individual networks perform better than networks learned from non-speaker specific data (Bergmann and Kopp, to appear).

For the rule-based choices of in the model we calculated the angle between the predicted and the empirically observed orientation vector (as in the reliability study). Considering this, the mean deviation for palm orientation of 54.6° (SD = 16.1°) and the mean deviation for back of hand orientation of 37.4° (SD = 8.4°). As concerns the gesture's movement features, the movement type (linear or curved) could be predicted with 76.4% accuracy (SD=13.6). For the movement direction we distinguish between motions through the sagittal, transversal and frontal planes. Each segment in the generated movement description is tested for co-occurrence with the annotated value, resulting in a accuracy measure between 0 (no agreement) and 1 (total agreement). The mean similarity for movement direction

.75 (SD = .09). These are quite satisfying results with deviations which lie well within the natural fuzziness of communicative gestures.

To evaluate GNetIc-generated gestures in terms of their impact on the interaction between humans and machines, we are currently setting up a study to analyze if (1) semantic information uptake from gestures, and (2) the perceived interaction quality (expressiveness, naturalness etc.), is influenced by the agent's gesturing behavior. Generated gestures whose features do not fully coincide with our original data may still serve their purpose to communicate adequate spatial features of their referents – even in a speaker-specific way.

The conclusion to be taken is that the GNetIc simulation approach beside allowing an adequate simulation of speaker-specific gestures, is an valuable means to shed light onto the open research questions of (1) how iconic gestures are shaped and (2) which sources individual differences in gesturing may originate from.

5. Conclusion

The SaGA corpus is a large collection of naturalistic, yet content-controlled multimodal data. In order to make sure that its secondary data fulfill the scientific requirement of reproducibility, the data have been systematically annotated and evaluated in terms of interrater agreement methods. That ensured, the SaGA corpus is used in order to explore empirically the interplay of speech and gesture in giving directions and describing objects.

Achnowledgement

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

6. References

M.W. Alibali. 2005. Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information. *Spatial Cognition and Computation*, 5:307–331.

Kirsten Bergmann and Stefan Kopp. 2009a. GNetIc – Using Bayesian Decision Networks for iconic gesture generation. In Z. Ruttkay, M. Kipp, A. Nijholt, and H. Vilhjalmsson, editors, *Proceedings of the 9th International Conference on Intelligent Virtual Agents*, pages 76–89, Berlin/Heidelberg. Springer.

Kirsten Bergmann and Stefan Kopp. 2009b. Increasing expressiveness for virtual agents – Autonomous generation of speech and gesture in spatail 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, Budapest, Hungary.

Kirsten Bergmann and Stefan Kopp. 2010. Systematicity and idiosyncrasy in iconic gesture use: Empirical analysis and computational modeling. In S. Kopp and I. Wachsmuth, editors, *Gesture in Embodied Communication and Human-Computer Interaction*, pages 182–194. Springer, Berlin/Heidelberg.

Kirsten Bergmann and Stefan Kopp. to appear. Modeling the production of co-verbal iconic gestures by learning bayesian decision networks. *Applied Artificial Intelligence*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Michel Denis. 1997. The description of routes: A cognitive approach to the production of spatial discourse. *Current Psychology of Cognition*, 16:409–458.

Umberto Eco. 1976. *A Theory of Semiotics*. Indiana University Press, Bloomington.

Rüdiger Gleim, Alexander Mehler, and Hans-Jürgen Eikmeyer. 2007. Representing and maintaining large corpora. In *Proceedings of the Corpus Linguistics 2007 Conference, Birmingham (UK)*.

Kilem Gwet. 2001. *Handbook of Inter-Rater Reliability*. STATAXIS Publishing Company, Gaithersburg, MD.

Florian Hahn and Hannes Rieser. submitted. Corpus-based gesture typology for explaining speech-gesture alignment in mm dialogue. Submitted to SEMDial.

M.A.K. Halliday. 1967. Notes on transitivity and theme in english (part 2). *Journal of Linguistics*, 3:199–247.

Annette Herskovits. 1986. *Language and Spatial Cognition: An Interdisciplinary Study of the Prepositions in English*. Cambridge University Press.

Adam Kendon. 1980. Gesticulation and speech: Two aspects of the process of utterance. In Mary Ritchie Key, editor, *The Relationship of Verbal and Nonverbal Communication*, volume 25 of *Contributions to the Sociology of Language*, pages 207–227. Mouton Publishers, The Hague.

Adam Kendon. 2004. *Gesture – Visible Action as Utterance*. Cambridge University Press.

Andy Lücking, Alexander Mehler, and Peter Menke. 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)*, pages 157–164, King's College London, June 2–4.

Karl-Erik McCullough. 2005. *Using Gestures during Speech: Self-Generating Indexical Fields*. Ph.D. thesis, The University of Chicago, Chicago, Illinois.

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

Cornelia Müller. 1998. *Redebegleitende Gesten. Kulturgeschichte – Theorie – Sprachvergleich*, volume 1 of *Körper – Kultur – Kommunikation*. Berlin Verlag, Berlin.

Charles Sanders Peirce. 1867. On a new list of categories. In *Proceedings of the American Academy of Arts and Sciences Series*, volume 7, pages 287–298.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

Hannes Rieser. 2010. On factoring out a gesture typology from the bielefeld speech-and-gesture-alignment corpus

(saga). In Stefan Kopp and Ipke Wachsmuth, editors, *Proceedings of GW 2009*.

Matthew Stone, Christine Doran, Bonnie Webber, Tonia Bleam, and Martha Palmer. 2003. Microplanning with Communicative Intentions: The SPUD System. *Comput. Intelligence*, 19(4):311–381.

Jürgen Streeck. 2008. Depicting by gesture. *Gesture*, 8(3):285–301.

Wilhelm Wundt. 1911. *Völkerpsychologie. Eine Untersuchung der Entwicklungsgesetze von Sprache, Mythus und Sitte*, volume Erster Band: Die Sprache. Erster Teil. Wilhelm Engelmann, Leipzig.