The Production of Co-Speech Iconic Gestures: Empirical Study and Computational Simulation with Virtual Agents

Dissertation zur Erlangung des Grades eines
Doktors der Ingenieurwissenschaften (Dr.-Ing.)

vorgelegt von

Kirsten Bergmann

bei der Technischen Fakultät
der Universität Bielefeld

The Production of Co-Speech Iconic Gestures:
Empirical Study and Computational Simulation with Virtual Agents

Kirsten Bergmann
Sociable Agents Group
Faculty of Technology
Bielefeld University
Email: kbergman@techfak.uni-bielefeld.de

Abdruck der genehmigten Dissertation zur Erlangung
das akademischen Grades Doktor der Ingenieurwissenschaften (Dr.-Ing.)
der Technischen Fakultät der Universität Bielefeld,
vorgelegt am 09.03.2011,
verteidigt am 19.07.2011.

GUTACHTER:
apl. Prof. Dr.-Ing. Stefan Kopp
Prof. Dr. Catherine Pelachaud
Prof. Dr. Hannes Högni Vilhjálmsson

PRÜFUNGSAUSSCHUSS:
Prof. Dr. Philipp Cimiano
apl. Prof. Dr.-Ing. Stefan Kopp
Prof. Dr. Catherine Pelachaud
Dr. Manja Lohse

Gedruckt auf alterungbeständigem Papier nach DIN-ISO 9706.

Acknowledgements

The research reported in this thesis was carried out during my time as a member CRC 673 project B1 at Bielefeld University. First of all, I am most deeply grateful to my advisor Stefan Kopp for his enthusiasm, guidance, encouragement, helpful suggestions and comments throughout my research for this work!!! I'm also grateful to Florian Hahn, Andy Lücking, Hannes Rieser and Ipke Wachsmuth for the joint work in B1.

I would like to thank all members of the Sociable Agents Group, including Christoph Broschinski, Hendrik Buschmeier, Ulf Grossekathöfer, Christian Pietsch, Amir Sadeghipour, Maha Salem, and Ramin Yaghoubzadeh for the positive atmosphere, cooperation and help. Special thanks go to Volkan Aksu, Elmar Bienek, Hendrik Buschmeier, Oliver Damm, Lars Mödeker, Sebastian Ptock, and Nick Thomas for their contributions to programming the prototype system and helping with data analyses.

Further special thanks go to Catherine Pelachaud who agreed to be the second reviewer of this thesis. And I'm also grateful to Hendrik Buschmeier, Bridgette DeCot, Oliver Damm, and Ramin Yaghoubzadeh for proofreading the manuscript and giving helpful hints.

Finally, I'd like to thank my family and friends, and, in particular Michael for all their support and encouragement throughout the time of this work.

Contents

> "[...] why different gestures take the particular physical form they do is one of the most important yet largely unaddressed questions in gesture research."
>
> Bavelas et al., 2008

CHAPTER 1

Introduction

1.1 Motivation

When we are face to face with others, we use not only speech, but also a multitude of nonverbal behaviors to communicate with each other. A head nod expresses accordance with what someone else said before. A facial expression like a frown indicates doubts or misgivings about what one is hearing or seeing. A pointing gesture is used to refer to something. More complex movements or configurations of the hands depict the shape or size of an object. Of all these nonverbal behaviors, *gestures*, the spontaneous and meaningful hand motions that accompany speech, stand out as they are very closely linked to the semantic content of the speech they accompany, in both form and timing. Speech and gesture together comprise an utterance and externalize thought; they are believed to emerge from the same underlying cognitive representation and to be governed, at least in part, by the same cognitive processes (Kendon, 2004; McNeill, 2005).

Gestures are an integral part of human communication, as Goldin-Meadow (2003, p. 4) so aptly put it: "whenever there is talk, there is gesture". There is, actually, a growing body of evidence substantiating the significant role of gestures. The importance becomes apparent in the fact that gestures already develop early in our linguistic development: children at the one-word stage already systematically combine a word and a gesture (Goldin-Meadow and Butcher, 2003). Even congenitally blind people, who have never seen anybody gesturing, spontaneously produce gestures while talking (Iverson and Goldin-Meadow, 1998). The important role of gestures in communication is further supported by the fact that "to date there is no report of a culture that lacks co-speech gestures" (Kita, 2009, p. 146). The role of gestures of such significance that speech-accompanying gestures do not disappear when visual contact between speaker and listener is absent, e.g., on the telephone (Cohen, 1977; Bavelas et al., 2008).

What tempts us to use gestures? What are the circumstances under which speakers make use of a gesture? And under which circumstances do they not employ gestures?

What aspects of meaning do speakers select to be expressed gesturally? What perspective do they adopt when depicting an object? Having a particular object in mind that aims to be depicted gesturally—with which hand do speakers perform a gesture? The left? The right? Or both hands? How do speakers shape and orient their palm(s) and fingers to depict a particular shape? What kind of movement trajectory do the hands or fingers perform?

This thesis does not take up all of these questions, but focusses on those which concern the physical appearance of gestures. As shown by the epigraph by Bavelas et al. (2008) these issues are mostly unanswered. That is, it remains an ambitious objective to gain some understanding of the mechanisms that underlie gesture production in human speakers. Along the same lines, de Ruiter (2007, p. 30) recently put the problem thus: "generating an overt gesture from an abstract [...] representation is one of the great puzzles of human gesture, and has received little attention in the literature".

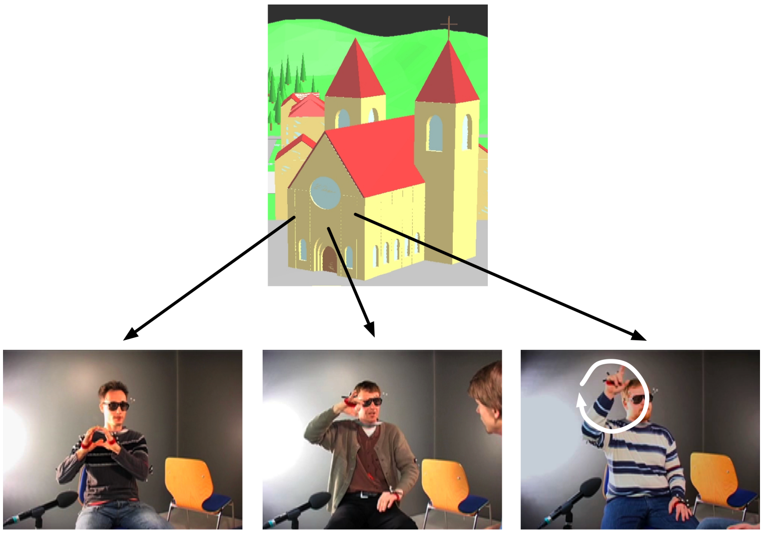


Figure 1.1: Gestures from three different speakers all of which depicting the same round church window.

In Figure 1.1, examples are given from three speakers who are describing the same stimulus, a round church window. The speaker on the left hand side uses a two-handed gesture in which the shape of the hands statically depicts the shape of the window. The gesture of the speaker in the middle is similar such that the hand is shaped in a way that bears a resemblance with the shape of the window. The difference is that the one in the middle is performed with only one hand. Finally, the speaker on the right hand side depicts the window by drawing its shape in the air. So the same stimulus is depicted in quite different ways, showing that there is no one-to-one mapping between a gesture and the object it depicts. McNeill and

Duncan (2000) termed this phenomenon ‘idiosyncrasy’ implying that gestures are not held to standards of good form, but are rather created locally by speakers while speaking. McNeill and Duncan (2000, p. 143) conclude that, “by virtue of idiosyncrasy co-expressive, speech-synchronized gestures open a ‘window’ onto thinking that is otherwise curtained”.

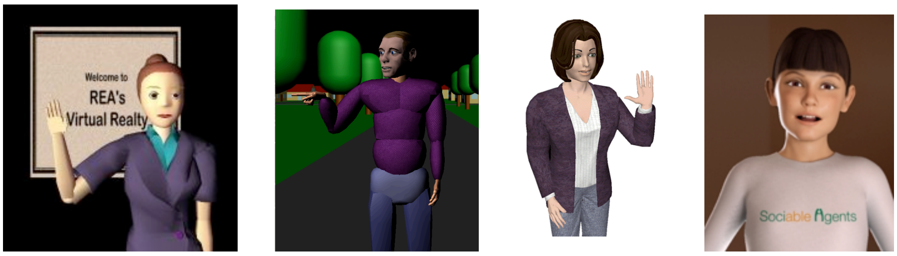
The fact that people are able to recover and interpret the meaning of iconic gestures, as shown by Cassell et al. (1999), suggests that there is at least some systematicity in the way speakers encode meaning in gestures. The examples above clearly indicate that there is a certain degree of iconicity in the gestures since the circular shape of the window becomes apparent in all of them. There are, however, differences in how the physical form of the gestures corresponds to the object they depict: in the gestures of the first two speakers, the hand(s) adopt the shape of the window to be depicted, that is, there is a resemblance between the hand configuration and the object shape. Likewise, in the third speaker’s gesture, the round shape of the reference object is depicted by the movement of the speaker’s drawing hand. That is, there obviously exist several ‘techniques’ to represent the same kind of meaning in gestures: whereas in the lefthand and the middle examples, the hands adopt a static posture as a model for the circular shape, the same shape is depicted by a circular movement of the drawing hand in the righthand example. For an adequate account of how meaning is transformed into gesture form, these representation techniques certainly have to be taken into account. Concrete mapping rules are more likely to be found within a set of gestures belonging to the same representation technique than across all instances of iconic gesture use.

Empirical studies, however, reveal that similarity to the referent cannot fully account for all occurrences of iconic gesture use (Kopp et al., 2007). Rather, recent findings actually indicate that a gesture’s form is also influenced by specific contextual constraints such as the discourse context (Holler and Stevens, 2007; Gerwing and Bavelas, 2004), the linguistic context (Kita and Özyürek, 2003; Gullberg, 2010; Bavelas et al., 2002), and gesture history (McNeill, 2005). In addition, human beings are all unique and inter-individual differences in gesturing are quite obvious (Hostetter and Alibali, 2007).

Modeling the Production of Iconic Gestures Psycholinguistic approaches aiming to model the production process of iconic gestures, however, do not account for these complex influences of referent characteristics, contextual factors, and inter-individual differences. Rather, they focus on particular aspects of the gesture production process, emphasizing either the mapping of referent characteristics onto gesture form, or the relationship between speech and gesture, or inter-individual differences in gesture production.

The same holds for computational accounts of gesture production, in which different modeling approaches have been tested in an effort to translate systematic characteristics of co-verbal gestures, shared among speakers, into generative models. Some of these models focus on the influence of contextual factors on gesture use (Cassell et al., 2000a, 2001) or on how meaning is mapped onto gestural form features (Kopp et al., 2007). Others have emphasized individual differences in communicative behavior trying to model individual gesture style (Ruttkay, 2007; Hartmann et al., 2006; Neff et al., 2008).

The Role of Gesture Use in Human-Computer Interaction Conversational skills have developed in humans in such a way as “to exploit all of the unique affordances of the human body” (Cassell, 2000, p. 1). This is obviously true for the case of gestures for which a speaker employs his hands and arms. Due to the significant role of the human body for communication, the metaphor of face-to-face conversation has been applied to human-computer interaction in the form of *Embodied Conversational Agents*. This term was introduced by Cassell et al. (2000b) for agents that are represented with a human or animal body to appear lifelike and believable. These agents may either be *physically* embodied, e.g., as mobile robots, or, *virtually* embodied, being represented by a graphical body. The latter are called virtual agents, some examples are given in Figure 1.2.



(a) REA (b) Max (c) Greta (d) Billie

Figure 1.2: Examples of virtual agents from left to right: REA (Cassell et al., 2001), Max (Kopp and Wachsmuth, 2004), Greta (Poggi et al., 2005), and Billie (Kopp, 2010).

According to Cassell (2000), virtual agents are not just computer interfaces represented by way of human or animal bodies, but rather exhibit the same properties as humans in face-to-face conversation, including the ability to generate verbal and non-verbal output. This definition of virtual agents entails the challenge of endowing them with human-like and multimodal expressiveness. As gestures play such an important role in human communication, the automatic generation of flexible gestural behavior is a major goal towards natural and intuitive human-computer interaction.

1.2 Research Aim

The major goal of this thesis is to develop a computational simulation model for the production of speech-accompanying iconic gestures to be realized in virtual agents. In particular, the circumstances under which gestures are used and how the physical appearance of gestures is shaped will be investigated. Substantial mechanisms and contextual factors are to be identified and computationally replicated with regard to the question of how to transform meaning into gesture form.

The rationale behind this objective is twofold. First, devising and probing a predictive model of gesture production strives to increase our understanding of the cognitive mechanisms underlying gesturing as an intuitive form of human communication. Second, a computational model that allows agent-based interfaces to compose their gestural behavior adequately should improve human-agent interaction such that it progresses towards intuitive and human-like communication. Accordingly, the success of the computational simulation model is to be evaluated in two ways. First, by its accuracy in predicting gestural behavior as observed in human speakers: To which extent can the model automatically generate gestures comparable to those of the human archetype? Second, by exploring if and how automatically generated gestures can be beneficial for human-agent interaction. In other words, what are the effects of the gesture generation account on a user's perception of virtual agents?

The scope of this work is restricted to gestures used in object descriptions. These gestures typically accompany noun phrases in speech. Other forms of gesture use, like verb-phrase aligned gestures (e.g., pantomime gestures or typical direction-giving gestures as in "turn right"), or dialogue regulating gestures (e.g. indicating someone to take the turn, or to 'brush away' what someone else said), remain unconsidered.

1.3 Methodology

To investigate the research objective, this thesis follows the design methodology of modeling communicative behavior (Cassell and Tartaro, 2007) as illustrated in Figure 1.3. This design process allows the investigation of communicative behavior, the construction of formal models of interaction, and the evaluation of those models.

The design process starts with an *empirical study* of natural human communication, collecting data that reflects the surface level of the behavior to be modeled. This data is enriched with annotations for the purpose of analyzing how speech, gestures, and other non-verbal behaviors are used and combined. The additional coverage of the context in which the communicative behavior occurs is the basis for the next step of the design process, building a *predictive model*. This is done either by extracting rules from the analysis of the human behavior data (Cassell et al., 2001; Kopp et al., 2004) or by using the data directly as a resource for data-driven techniques, as demonstrated in, for example, Neff et al. (2008).

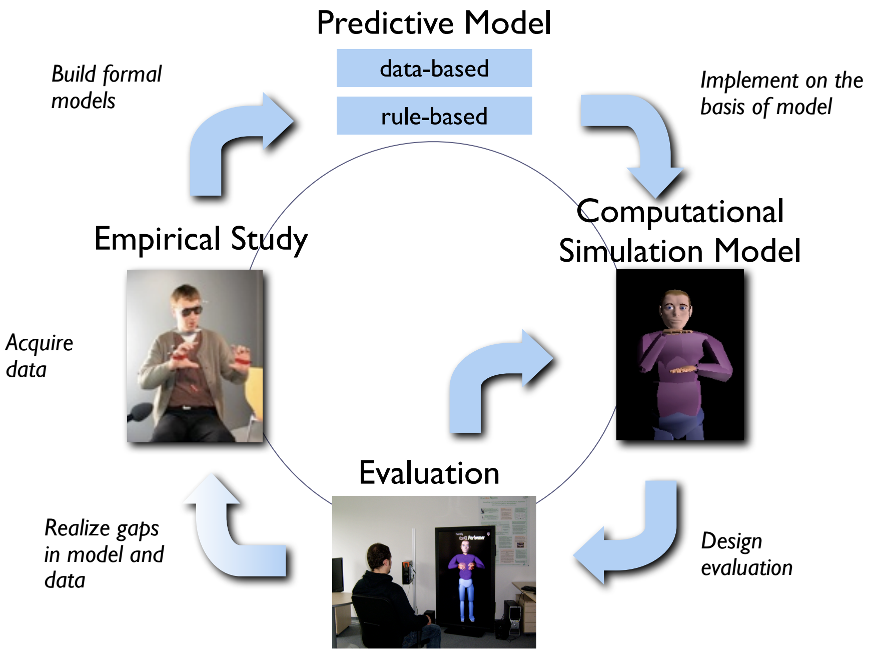


Figure 1.3: Iterative design methodology of modeling communicative behavior, following Cassell and Tartaro (2007).

Such a model is then translated into an *implementation* for a virtual agent or a robot, taking into account the constraints of the particular computational platform. This realization is *evaluated* by having people interact with it. An analysis of evaluation results allows identification of shortcomings and gaps in the model. Implementations in virtual agents or robots also allow modification the communicative behaviors model, e.g., one may easily switch particular components of the model on or off to analyze the effects on human users and their perception of the interaction. At this stage, the development cycle may become iterative: returning to the data may start a new pass through the cycle in order to extract improved models of communicative behavior.

1.4 Thesis Structure

Chapter 2, **Theoretical and Empirical Background**, reviews and discusses relevant research literature. It covers the phenomenological view on communicative gestures, introduces the relevant terminology, and frames the type of gestures to be investigated in this thesis. As a first step towards a computational simulation model, evidence is collected to identify substantial mechanisms and contextual factors of gesture use. The chapter is closed with a review and discussion of theoretical gesture production models with regard to the question of how far they account for those mechanisms and factors.

Chapter 3, **Generating Gestures—The Computational Perspective**, puts the focus on computational approaches and analyzes the strengths and weaknesses of existing systems that generate gestural behavior and simulate it with virtual agents.The computational perspective on gesture production will be completed with a survey of work concerning the representation of visuo-spatial and shape-related knowledge, an essential prerequisite for the generation of iconic gestures.

Following the design methodology of modeling communicative behavior, Chapter 4, **Empirical Study**, describes the empirical basis of the generation model. It introduces the Bielefeld Speech and Gesture Alignment (SaGA) corpus and presents results with regard to the question of how meaning gets transformed into gesture form under consideration of modulating factors.

Having so far covered the empirical basis, Chapter 5, **A Model of Iconic Gesture Generation**, deals with the conception of a simulation model. Requirements for a generation model are defined based on insights from the empirical results and an adequate formalism will be identified meeting those requirements. In a similar way, in the second part of the chapter, requirements for a comprehensive speech and gesture generation model into which the gesture production model is to be integrated are formulated. This is followed by the conception of an overall generation architecture for the integrated generation of speech and gestures.

Chapter 6, **Realization**, is concerned with the implementation of the concepts developed in the previous chapter. It describes how the gesture production model is realized and integrated into the overall production architecture. A generation example of a whole multimodal utterance is given and modeling results are presented and discussed.

In Chapter 7, **Evaluation**, the gesture production model developed so far is evaluated in two respects, recalling the two-fold rationale followed in this thesis. First, in terms of a corpus-based evaluation, the model's prediction accuracy is measured in comparison with the empirically observed gesturing behavior in the SaGA corpus. Second, in terms of a perception-based evaluation, the degree to which the automatically generated gesturing behavior is able to improve the interaction of human users with virtual agents is investigated.

Finally, Chapter 8, **Conclusion**, summarizes and discusses the results and provides an outlook on future research perspectives.

CHAPTER **2**

Theoretical and Empirical Background

The previous chapter introduced the motivation, research aims, and methodology of this thesis with the primary research objective of developing a computational simulation model of iconic gesture production. In preparation for this task, this chapter reviews and discusses relevant research literature. Section 2.1 covers the phenomenological view on communicative gestures, introduces the relevant terminology, and frames the type of gestures to be investigated in this thesis. Section 2.2 collects evidence for a number of factors modulating gesture use, including their underlying representation, the mapping of meaning onto gesture form, contextual factors, and inter-individual differences. Section 2.3 deals with theoretical models of gesture production which aim to outline the process of gesture production from a psycholinguistic point of view and discusses in how far the different models consider for those factors.

2.1 Gestures

What is a gesture? The understanding of the term 'gesture' in this thesis follows Kendon (2004, p. 7), who defines the term as "*visible action when it is used as an utterance or part of an utterance*". To make sense of this definition, it is necessary to specify what is meant by 'utterance' and 'action'. To start with the former, Kendon recalls Goffman (1981) who understands an utterance as any ensemble of action that counts for others as an attempt to provide some kind information to others. This comprises any unit of activity treated by those co-present as a communicative move, turn, or contribution. These units of activity "may be constructed from speech or from visible bodily action or from combinations of these two modalities" (Kendon, 2004, p. 7). Whereas it is relatively easy to recognize whether or not speech is involved in an utterance, the notion of 'visible bodily action' deserves some elaboration.

A dictionary[1] entry defines the term ‘gesture’ as “*A motion of the hands, head, or body to emphasize an idea or emotion, esp. while speaking.*” This description, however, is very broad since it includes motions of any body part as well as movements such as clothing adjustments or hair-pattings. To narrow the scope of the above dictionary entry, a definition given by McNeill and Levy (1982, p. 5) is suitable: “any visible movement of the hand(s) excluding selfadaptors (scratching the head, fizing the hair)".

On the Semiotic Character of Gestures Psycholinguists focus particularly on the semiotic character of gestures: “The gesturer is transforming stored information (internal representation) into patterned movement” (Tuite, 1993, p. 92). That is, gestures are treated particularly as referential acts to convey meaning, depict events, and represent ideas. They specify and often clarify verbal references, and they can denote meanings that may not be in the accompanying words (Bavelas et al., 1992). In this view, gestures are seen as “manual symbols [...] analyzable as paired signifiers and signifieds” (McNeill, 1985, pp. 351f.), which is building on the sign model proposed by de Saussure (1916), who offered a dyadic model of the sign as a composition of the form which the sign takes (the ‘signifier’) and the concept it represents (the ‘signified’). De Saussure, however, focused on the linguistic sign and the arbitrary nature of these signs. A different model of the sign was proposed by Peirce (1965) who offered a triadic model consisting of the form the sign takes (its ‘representamen’), (2) an ‘object’ to which the sign refers, and (3) the sense made of the sign by an interpreter (the ‘interpretant’).

Peirce further devides signs into three classes characterized by the relation between a sign’s components: ‘icon’, ‘index’, and ‘symbol[2]. Icons are characterized by the fact that the representamen is perceived as resembling or imitating the object. The resemblance is typically visual, as in a portrait, or auditory, as in onomatopoeia. The three gestures in Figure 1.1 have an iconic character because their physical appearance bears some relationship to the circular shape of the object they represent. An index is a sign which is is directly connected to the object in some way, e.g., physically or causally. Examples of indexical signs are signposts, indexical words like ‘this’ or ‘that’, and pointing gestures. Indexical signs are context-dependent; all the aforementioned examples become meaningless or change their meaning if they are deprived of context. A symbol is a sign for which the relation between representamen and object is fundamentally arbitrary or conventional - such that the relationship must be learned. Words are typical examples of symbolic signs, e.g., there is no reason why a window is called a ‘window’. There are also some gestures which have a symbolic character, such as the ‘thumbs up’ sign.

[1] Collins English Dictionary, 7th Edition, 2005

[2] This trichotomy inspired McNeill’s classification scheme of iconic gestures which will be explained in Section 2.1.1

2.1.1 Classifying Gestures

To systematize these semiotic gestures, Kendon (1988) proposed a continuum of five types of hand and arm movements. This ordering was named *Kendon's continuum* by McNeill (1992), who identified two cross-cutting dimensions of the continuum: conventionalization and the degree of speech presence. In a later version, McNeill (2000, 2005) added two further dimensions: the extent of linguistic properties hand movements possess and their semiotic content. Four types of hand and arm movements are distinguished and placed along each continuum (see Figure 2.1).

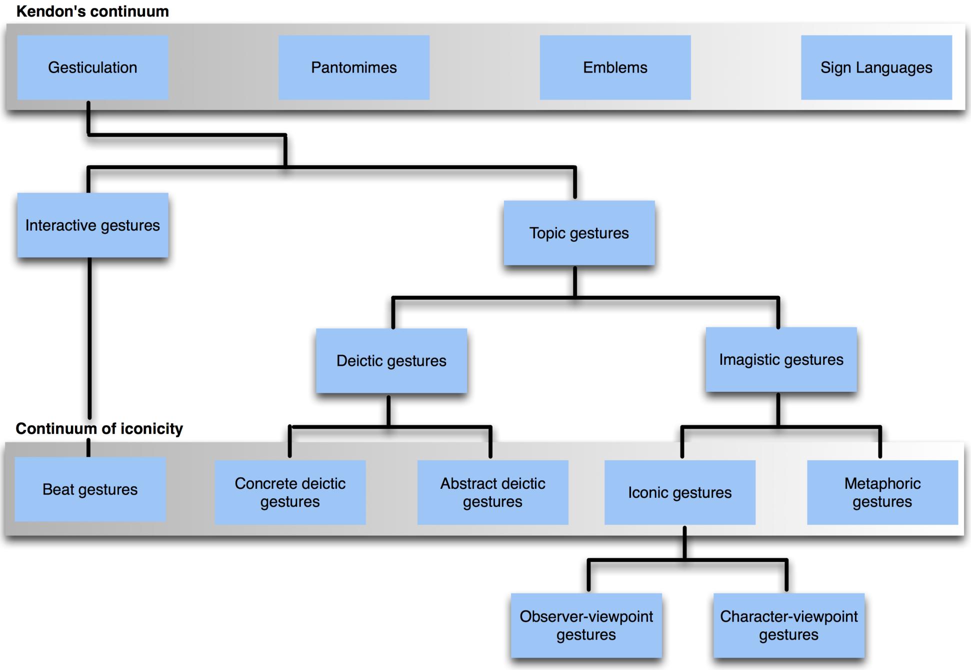


Figure 2.1: Summary of gesture typologies proposed by McNeill (1992, 2005), Bavelas et al. (1992), and Gullberg (1998). See text for details.

The first continuum controls the co-occurrence of gestures and speech. Speech becomes less obligatory from left to right. While gesticulations are meaningful only in conjunction with verbal utterances, the simultaneous production of sign language and speech is disruptive for both speech and sign. The second continuum, the degree of linguistic property, refers to how much the hand and arm movements obey any system constraints with regards to well-formed vs. not-well-formed ways of producing the movements. The third continuum, the degree of conventionality, is reflected in the measure of agreement about the manual form among users. The fourth continuum regards the semiotic differences of hand and arm movements.

At one extreme of the continuum are *sign languages* which are independent languages which display phonology, morphology, semantics, syntax and pragmatics just like spoken languages (Stokoe, 1972). Sign languages such as American Sign Language (ASL) develop spontaneously and independently within communities of deaf people throughout the world and are subject to regional variation just like spoken languages. Next on the continuum are *emblems* as highly conventionalised gestures whose forms and meanings do not differ from one speaker to another within the same culture or group and often replace speech. While some emblems have different meanings across cultures, others are transcultural, or even largely global. Examples of the latter are the 'thumbs up' sign or the 'OK' sign made by forming a circle with the forefinger and the thumb with the other fingers extended outward. The next type of gesture on the continua away from sign languages are *pantomimes*. They are re-enactments of actions which are not as conventionalised as emblems. In contrast to emblemes, pantomines are characterized by absence of standardized well-formedness leaving some space for individual variations in expression. Finally, at the other end of the continuum, opposite sign languages, are *gesticulations*—unconventionalised hand and arm movements which are nearly always accompanied by speech. There are no rules according to which these hand and arm movements are produced. Notably, this does not exclude the possibility that their production exhibits certain patterns among speakers. As an example of this gesture type, consider the variations of co-speech gestures depicting a circular window in Figure 1.1. The emphasis in this thesis is on these gesticulations and all usages of the term 'gesture' to follow are should be understood as such.

Types of Speech-Accomanying Gestures

Many attempts have been made to categorize and classify speech-accompanying gestures. Most classification schemes are based on early work by Wundt (1973) and Efron (1970). The most common classification scheme used in gesture research is a four-way distinction among 'beats', 'deictics', 'iconics', and 'metaphorics' (McNeill and Levy, 1982; McNeill, 1992, 2005), the main characteristics of which will be described briefly while keeping in mind McNeill's claim that none of these categories is exclusive to members of the others. Asserting that, "most gestures are multifaceted—iconicity is combined with deixis, deixis is combined with metaphoricity, and so forth", McNeill (2005, p. 38) suggests differentiating gesture in terms of dimensions rather than disjunctive categories.

Beat Gestures Beats are simple, rhythmic, and repetitive flicks of the hands in a vertical or horizontal direction. There is no obvious relationship between gesture form and the semantic content of the accompanying speech (Feyereisen et al., 1988). Beats are coordinated with speech prosody and tend to fall on stressed syllables (McClave, 1994). McNeill and Levy (1982) called them ‘beats’ as they take the form of the hand movement marking, or beating, time. Other names adopted in classifications are ‘batons’ (Efron, 1970; Ekman and Friesen, 1972), ‘speech-focused movements’ (Butterworth and Beattie, 1978), ‘speech-marking’ (Rimé and Schiaratura, 1991), and ‘motor gestures’ (Krauss et al., 2000).

Deictic Gestures Deictic gestures are “pointing movements, which are prototypically performed with the pointing finger” (McNeill, 1992, p. 80). Other hand shapes, however, may also be used, such as an open hand with outstretched fingers. Deictic gestures have no fixed meaning, but rather, their meaning is the act of indicating the objects (Krauss et al., 2000). They refer not only to objects or persons, but also to object properties. Deictics are used to indicate concrete or abstract entities, depending on whether the target is physically present or not. Whereas concrete deictic gestures are among the first to develop in children, abstract pointing develops late, around the age of twelve (McNeill, 1992).

Iconic Gestures According to McNeill (1992), iconic gestures bear a close formal relationship to the semantic content of speech by depicting some property of the speech referent. For instance, drawing a circular shape in the air when talking about a window (as in the example in Figure 1.1) iconically depicts the circular shape of the window. In addition to objects and object properties, iconic gestures may also depict actions, as in, “he was walking,” accompanied by a gesture in which the index and middle fingers move back and forth in opposition. Iconic gestures are typically performed on the fly—as there are no standards of form, individuals create their own gestural depiction at the moment of speaking. Another term used for this class of iconic gestures is ‘physiographic gestures’ (Efron, 1970; Rimé and Schiaratura, 1991).

The fact that listeners actually understand the meaning of iconic gesture (Cassell et al., 1999) is explained by its similarity or resemblance to that which it depicts. That is, similarity (or resemblance) is the core notion by which *iconicity* is defined. Ekman and Friesen (1969, p. 60), accordingly, suggested that an iconic gesture, “looks in some way like what it means.” This relation between signifier and signified has, however, been subject to a philosophical discussion on the nature of resemblance. Typically, if two entities share a number of characterizing properties, this is what is understood by *resemblance*. This definition, however, does not meet the notion of depiction in a satisfactory way since resemblance is defined to be reflexive and symmetric (Goodman, 1976). That is, a gestural depiction would represent its referent and symmetrically the referent would, vice versa, represent the gesture. Streeck (2008) proposed a way to deal with this insufficiency based on Goodman (1976). He assumes that the gesture that depicts an object or process of any kind offers an analysis of the signified. That is, the gesture does not mirror, but analyze the object: “the gesture *is* not like its referent, but rather shows *what the referent is like*" (Streeck, 2008, p. 286, emphasis in the original). This notion of iconicity is the one used in this thesis.

Metaphoric Gestures Metaphoric gestures "are similar to iconics in that they present imagery, but present an image of an abstract concept" (McNeill, 1992, p. 80) that is not physically present. As an example, consider an upward or downward movement of the hands while talking about the economy rising or falling. According to McNeill (1992), metaphorics are more complex than iconics as they depict two things: (1) the base, which is the concrete entity or action presented in the gesture, and (2) the referent, which is the concept that the base stands for. Drawing a distinction between iconic and metaphoric gestures is not always easy, as Krauss et al. (2000, p. 276) put it: "it makes more sense to to think of gestures as being more or less iconic rather than either iconic or metaphoric," labelling these 'lexical gestures'. Metaphorics are also called 'ideographics' by Efron (1970) and Ekman and Friesen (1972).

Other researchers proposed modifications to this four-way distinction, either via additional categories or by merging labels for some of the four types. Based on research on face-to-face dialogue, Bavelas et al. (1992) proposed a further distinction between 'topic gestures' and 'interactive gestures'. Topic gestures depict semantic information directly related to the topic of discourse, whereas interactive gestures refer to some aspect of the process of conversing with another person. The latter are, thus, independent of the topic and addressed to the interlocutor. They subsume, but are not limited to McNeill's class of beat gestures. Gullberg (1998) proposed a continuum of iconicity based on McNeill's typology of gestures. While beats, at one end of the continuum, have no representational character, fully iconic gestures are placed at the other end of the continuum. In-between are abstract deictic gestures, metaphoric gestures, and concrete deictics.

Figure 2.1 summarizes the gesture typologies reviewed so far. The focus of this thesis will predominantly be on iconic gestures. In addition, abstract deictic gestures will also be taken into consideration as there is "a relatively transparent form-meaning relationship" (Kita, 2000, p. 162) for both of these gesture types: for iconic gestures there is a kind of isomorphism between a gesture's form and its referent, and abstract deictic gestures establish a virtual object in gesture space. Kita, therefore, labels iconic and abstract deictic gestures collectively as 'representational gestures', which are exactly the scope of this work.

2.1.2 Spatio-temporal Structure of Gestures

Temporal Structure Regarding the question of how gesturing behavior unfolds in time, Kendon (1972, 1980, 2004) analysed the structure of gesticulations. He identified *gesture units* as the largest interval starting when the articulators "begin to depart from a position of relaxation until the moment when they finally return to one" (Kendon, 2004, p. 111). In other words, a gesture unit is delimited by successive rest positions of the limbs, i.e., poses where the hands either hang down at the side of the body, lie in the lap or on an armrest.

Within the course of a gesture unit one or more *gesture phrases* may be distinguished. These prototypically consist of three consecutive gesture phases: *preparation*, *stroke*, and *retraction*, to borrow terms from McNeill (1992). In the preparation phase, the hands are brought into a position in gesture space where the stroke can begin. According to Kita et al. (1998), the preparation phase starts with an optional *liberating movement* such as unclasping interlocked fingers. Then, the *location preparation* brings the hand to a starting position in gesture space, while hand shape and hand orientation are set to starting values in the *hand-internal preparation* phase. Location preparation and hand-internal preparation usually overlap in time, whereby the hand-internal preparation does not precede the location preparation. The *stroke* is the meaningful phase of the gesture characterized by a "distinct peaking of effort" (Kendon, 1980, p. 212). Similarly, McNeill (1992, p. 83) defines the stroke as "the peak of effort" in which "the meaning of the gesture is expressed". In the *retraction* phase the hands are either brought back into a rest position, or made ready for another stroke. The latter variant is called a *partial retraction* (Kita et al., 1998). In addition to these three basic phases, temporary cessations of motion might occur either before or after the stroke (Kita, 1993): *pre-stroke holds* and *post-stroke holds*. These holds are used to synchronize the stroke with speech. Post-stroke holds may additionally extend the meaning conveyed by the stroke for the duration of the hold. (Kendon, 2004) refers to the phase of action that includes the stroke and the post-stroke hold as the *nucleus* of the gesture phrase. Notably, all gesture phases are organized around the stroke being the only obligatory phase of a gesture. It is optionally prepared for, held and finally retracted from.

An alternative to Kendon's and McNeill's view that the stroke is the most effortful gesture phase proposed by Kita et al. (1998), which argues that a gesture phrase does not necessarily contain a dynamic stroke. Instead, Kita and colleagues refer to the obligatory and semiotically active phase of each gesture as the *expressive phase*. It contains either a stroke or an *independent hold* which is a static way to convey meaning gesturally (*stroke hold* in terms of McNeill).

Gesture Form The analysis of physical gesture form is an indispensable prerequisite of any account of gesture generation. The notion of *gesture features* (McNeill, 2000) is based on the observation that a gesture can be decomposed into parts, such as handshape, movement, size, hand orientation, and location in gesture space. Attempts at describing the physical features of gestures and suggestions for coding schemes have been proposed by several researchers from different disciplines (Calbris, 1990; McNeill, 1992; Müller, 1998; Kendon, 2004; Gibbon et al., 2004; Kopp et al., 2007; Lausberg and Slöetjes, 2009). For a detailed overview and comparison of form description attempts see Bressem (2008).

Despite differences between the attempts due to the fact that they were developed for different purposes and against different theoretical backgrounds, it is nevertheless widely accepted to describe gesture forms by the following four features: (1) handshape, (2) hand orientation, (3) position, and (4) movement characteristics. Support for the adequacy of this four-part description also comes from research on virtual agents: in representation formats which aim to specify multimodal behavior for virtual agents the same attributes are successfully employed to control the hand and arm movements of virtual characters (e.g., Kranstedt et al., 2002).

2.2 What Shapes the Use of Gestures?

The previous section dealt with the classification of gestures and their physical appearance. How is gesture use—in terms of both particular gesture types and their physical appearance—shaped at the moment of speaking to convey a particular meaning? In the literature, there is a growing body of evidence suggesting that gesture use is subject to the influence of multiple factors. A review of research regarding this question follows, arranged by type of modulating factors.

2.2.1 Mental Representation

Gesture production means to transform an abstract representation into an overt gesture (de Ruiter, 2007). The nature of the underlying representation is, therefore, of major relevance for the production process of gestures. The kind of representation that has received most attention in the literature on gesture production is *mental imagery*, “a quasi-perceptual experience [...] [that] resembles perceptual experience, but occurs in the absence of the appropriate external stimuli” (Thomas, 2010, p. 1). Evidence for this claim comes from frequent co-occurrences of images and gestures. Gestures were found to occur more often with speech about spatial information than with speech about nonspatial information (cf. Alibali, 2005). Similarly, gestures often occur when speakers express information that evokes images. A study by Beattie and Shovelton (2002) revealed that imageability had a significant effect on the probability of the core propositional unit being accompanied by a gesture. Participants rated written clauses originally accompanied by gestures as more evocative of images than those unaccompanied by gestures.

Hostetter and Alibali (2008) argued that frequent co-occurrence of gestures and images is due to an isomorphism between images and gestures in terms of being both *global* and *synthetic*. Global means that the meanings of the parts (gesture form features) depend in a top-down fashion on the meaning of the whole gesture. Synthetic means that several meanings (e.g., a referent’s shape, position, and orientation) are synthesized into one gesture (McNeill, 1992; McNeill and Duncan, 2000). The global and synthetic properties of gestures are similar to the global and synthetic properties of images. Images, like gestures, convey meaning globally, such that an image's meaning as a whole influences the interpretation of each part. Similarly, images are synthetic because they can integrate several meanings.

Thus, gestures and (mental) images seem to be closely related; gestures seem to be a natural and intuitive way to express spatial information. However, the question remains whether gestures actually originate from spatial and imagistic representations? Empirical evidence for this claim includes the work of Hostetter and Hopkins (2002), who found that speakers used more iconic/deictic gestures while retelling a cartoon story if they had watched the cartoon than if they had read a verbal description of it. Further evidence comes from studies of inter-individual differences. Individuals with weak spatial visualization skills were found to gesture less than individuals with stronger skills (Hostetter and Alibali, 2007). Similarly, stroke patients who have visuospatial deficits were found to gesture less than do age-matched controls (Hadar et al., 1998). This evidence converges to suggest that gestures derive, at least in part, from spatial or imagistic representations.

2.2.2 Form-meaning Relation

Closely related to the issue of representations underlying gesture production is the question of how meaning is transformed into an overt, observable gesture: how is a mental representation mapped onto gesture form? Literature addressing the relationship of form and meaning in iconic gesture is, however, sparse. In the following, a survey of results from research in different domains is given.

Gesture type, handedness and object complexity Marsh and Watt (1998) investigated the relationship of gesture form and meaning in a study in which 12 participants were provided with the name of an object or shape and their task was to describe the items non-verbally. The set of 15 items consisted of *primitive* shapes such as circles, triangles, and squares as well as *complex* shapes like chairs, cars, and houses. Marsh and Watt analyzed the relationship of object type gesture use with respect to handedness and gesture type.

It was observed from the study that participants generally preferred drawing gestures (hands outline or trace a picture of an object or shape) over posturing gestures (the hands are shaped to match the form of an object as if it was being held or grasped). With regard to handedness, Marsh and Watt further found that 3D shapes were always, and 2D shapes predominantly, expressed using two-handed gestures. Some participants only used a single hand to perform gestures when tasked with depicting the primitive shape of a circle. Similarly, iconic two-handed gestures were the preferred means to depict complex shapes.

Gesture form and visuo-spatial object features Sowa (2006) analyzed the relationship between gesture form and meaning in a corpus of 383 iconic gestures from 37 speakers. The participants' task was to describe parts and aggregates from a toy construction kit (screws, cubes, bars, etc.). Although use of hands was mentioned in the instructions, gestural explanations were not explicitly enforced. Subjects were told to explain the objects' appearance in such a way that others who would watch the video afterwards would be able to imagine the object.

Sowa identified a small set of pictorial strategies in terms of frequent pairings of gesture form and visuo-spatial referent features by evaluating the frequency of semantic features in referents as expressed by certain gestural form features. Due to the chosen referents in the study, the analysis focused on linear and circular shape features. To express *linear* extent in one dimension, Sowa identified three form features of gestures used with high frequency. The form feature most frequently used to express linear extent in one dimension was linear movement. The maximum distance between hands was used as second most frequent feature, while the two-point focus area was ranked third. *Circular* profiles were mainly expressed by circular motion, curved handshape, and distance in two-handed gestures.

Gesture features and object features McNeill and Levy (1982) analyzed cartoon retellings of six speakers with regard to the relationship of physical gesture features and mental representations suggested by word meaning. They considered the following properties of physical gesture features:

- Handedness (left-handed, right-handed, both-handed)
- Handshape (fist, curled fingers, extended fingers, extended index finger)
- Palm orientation (downwards, towards oneself)
- Movement trajectory (up, down, left-right, straight line)
- Movement repetition
- Two-handed configuration (both hands moving the same way)

These gesture features were correlated with meaning features such as 'downward', 'horizontal', 'rotation' expressed in speech. The analysis focused on 74 iconic gestures accompanying verbs and revealed several positive as well as negative correlations. Verbs with a 'downward' meaning feature tended to co-occur with gestures with downward motion and with curled fingers. In contrast, 'downward' does not go together with gestures whose meaning is opposite of 'downward', i.e., with upward movements and extended fingers, with both hands going in the same direction, and with reduplicated gestures. Verbs with a horizontal meaning feature were found to be correlate positively with left-to-right movements and with two-handed configurations. On the other hand, 'horizontal' does not go together with upward gestures (due to meaning incompatibility) as well as handshapes depicting closure and contact (fists and pointing). Further, the semantic features 'end state' and 'entrance/exit' tended to co-occur with two-handed gestures.

Going beyond a direct mapping between gesture features and object features, Kopp et al. (2004) hypothesize that the relationship between the form of iconic gestures and visuo-spatial aspects of their referents does not pertain to gestures at a whole-gesture level, but to their subparts. That is, referent features of shape, spatial properties, or spatial relationships are associated with more primitive form features of gesture morphology, such as hand shapes, orientations, locations, movements in space, or combinations thereof. In particular, Kopp and colleagues proposed a framework to deconstruct such gestural images into semantic units (so-called *image description features*) and to link these units to morphological gesture features such as hand shape or movement trajectory. Figure 2.2 illustrates the hypothesized relationship between a building with a vertical plane feature, a vertical plane (surface) definable in terms of image description features, and a flat, vertically oriented gesture form.

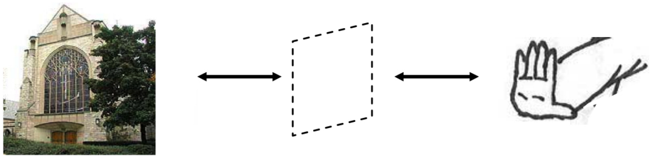


Figure 2.2: Mapping of concrete referent features (left) onto gesture morphology (right) via imagistic description features (middle) as proposed by Kopp et al. (2007).

Based on data from an empirical study on language and gesture use in direction-giving (∼1000 gestures), three gesture features were investigated to determine whether they correlated positively with the corresponding features of their referents.

Actually, it was found that flat hand shapes oriented vertically (with fingers pointing up), horizontally (with palms facing down), and sideways were positively correlated with referents that possess salient visual characteristics corresponding to vertical, horizontal, and sideways planes, respectively. Although the correlation was significant for all three morphological classes, the vertical flat morphological form was the best predictor of referent form. However, there was also a sizeable number of false negatives for each of the three relationships. For instance, despite the frequent linking of a flat vertical image description feature to a flat vertical handshape, there were also cases in which a referent exhibiting that feature was not described using a flat handshape. The conclusion to be drawn from this is that feature-based iconicity is not the sole driving force behind a representational gesture.

Although none of the abovementioned studies was able to provide a comprehensive account of the form-meaning relationship in iconic gestures, at least a few strategies became apparent, summarized in Table 2.1. Relations with referent features were found for several aspects of gesture use concerning both the type of gestures and gesture form features, but there seems to be no way to straightforwardly map referent characteristics onto gesture form. This suggests that the use of iconic gestures is determined by factors beyond the features of its referent.

Table 2.1: Overview of empirical findings concerning the form-meaning relationship of iconic gestures.

	Gesture Features	**Meaning Features**
Gesture Type	virtual depiction	primitive shape
	virtual depiction and pantomimic gestures	complex shape
Handedness	2-handed	horizontal motion
	2-handed	end state
	2-handed	entrance/exit
Handshape	curled fingers	downward motion
Palm-/Finger Orientation	extended fingers + vertical orientation	vertical plane
	extended fingers + horizontal orientation	horizontal plane
	extended fingers + sideways orientation	sideways plane
Movement	downward	downward motion
	left-right	horizontal motion

2.2.3 Sub-categories of Iconic Gestures

Another issue which is crucial for the generation of gestures is that the class of iconic gestures contains various gestures for which it is appropriate to make a further sub-division. This is due to the fact that there are different means of representation employed in iconic gesture use. Returning to the introductory example given in Figure 1.1, the same referent—a round window—is depicted in entirely different ways, either by tracing the outline in the air as if using a pen, or by bringing the hands into a form and position such that they act as a model for the referent itself. Apparently, the transformation of a referent representation onto overt gesture form is sensitive to such differences. In the following, different attempts that aim to further systematize and sub-classify iconic gestures will be described.

Wundt (1973) offered a classification of gestures based on the examination of sign languages under the assumption that these provide insight into psychological processes of language use. Wundt's classification was mainly focused on the semiotic status of gestures, but it also distinguishes techniques of representation. In particular, he assumes a single **dichotomy of modes** which distinguishes between *drawing* gestures whereby "the outline of the object is drawn in the air by the index finger", and *plastic* gestures in which "the image of the object is imitated three-dimensionally with the hands" (Wundt, 1973, p. 78). Wundt described the former as the more primitive of the two, as it predominates the natural gesture of deaf-mutes, whereas sign languages with a longer tradition make use of hand-formed shapes.

Müller (1998, p. 121) compared the production of gestures with the work of visual artists: "Just like pictures and sculptures in fine art, [gestures] are shaped by the properties of the pictorial devices". In her view, thus, methods of gestural depiction are to be considered analogous to artistic drawing or sculpting. Müller distinguished gestural **modes of representation** in considering two major aspects: what the gesticulating hands stand for, and what the hands are doing. On this basis, she differentiates the following four modes of gestural representation: (1) the acting hand, (2) the modeling hand, (3) the drawing hand, and (4) the representing hand. In *acting*, the hands simulate an activity pantomimically, often with imaginary objects involved in the activity. With the *modeling* hands, an object or the course of an action is represented by moving the hands as though they were actually touching the referent. In other words, the hands are used like those of a sculptor when forming figures from clay. *Drawing* means that the hands are used like a pen to trace the outline of an object. Typically the stretched index finger represents the pen. *Representing* hands slip inside the represented object and depict the ensuing events from the object's perspective.

According to Müller (1998), gestures are always a compromise between shape features (based on a rudimentary shape analysis by the speaker) and representation means of the hands. Gesture use, therefore, depends on the ability to extract characteristic features from the perceived environment and transform these with different means of representation into gestures.

Another sub-categorization of iconic gestures was proposed by Kendon (2004). Based on previous attempts to classify **techniques of representation**, Kendon argues for a three-fold division consisting of modeling, enactment, and depiction. In *modeling*, a body part is used as if it is a model for some object. The hands may be shaped so that the form of the hands bears some relationship to the shape of the object. *Enactment* means that the gesturing body parts engage in a pattern of movement that shares common features with the patterns of action to which it refers. In *depiction*, the gesturing body parts engage in a pattern that is recognized as 'creating' an object in the air, for example, in outlining the shape of a cake.

In his analysis of iconic gestures using a toy description corpus, Sowa (2006) analyzed the relationship between gesture form and meaning based on a set of 84 different *gesture prototypes*, each of these being a pairing of (1) gesture form and (2) visuo-spatial referent features. For these prototypes, he identified four general **gesture types** or strategies. First, *dimensional gestures* relate to the overall shape of the referent by indicating spatial extent. These may be further sub-grouped according to the number of depicted object dimensions into 1D-, 2D-, and 3D-gestures. These gestures do not always depict a referent three-dimensionally, but in an abstract version in only one or two dimensions, i.e., dimensionally underspecified. Second, *surface property gestures* depict certain features of an object's surface without reference to its shape as a whole, for instance, the use of a flat hand to depict a planar side. Third, *placeholder gestures* are characterized by the body part representing the object itself, characterized by a one-handed gesture with a distinctive handshape. Placeholder gestures are, thus, comparable to what Müller (1998) called 'representing hands' and what Kendon (2004) and Streeck (2008) labelled 'modeling gestures'. Finally, *spatial relation gestures* indicate the relative position and/or orientation of two object parts using one hand for each with the handshape depicting the shape of the respective part. This class of gestures is closely related to placeholder gestures, whereby spatial relation gestures are a particular case since they require both hands to indicate the relative position of the objects or object parts. Note that, Sowa studied a limited domain, restricted to the kinds of objects involved in the task given to the subjects. His four-part classification should, therefore, be seen in that context.

Streeck (2008) proposed a set of **gestural depiction methods** that emerged haphazardly during micro-analytic inquiry into meaning-making by hands in various cultural settings and communicative contexts. He found that speakers deploy twelve different practices which further sub-categorize the broader classes proposed by Wundt, Müller, and Kendon. For instance, Streeck (2008) differentiates between *making* gestures, which simulate the making and shaping of things, and *scaping* gestures, which differ from making gestures by giving shape to undivided domains and terrains instead of discrete objects. Another fine-grained distinction in Streeck's categorization is made between *handling* gestures, in which objects are indirectly represented by a schematic act that 'goes with' them, and *acting*, which is not very different from handling, but additionally includes actions in which no object is involved.

The fact that several researchers came up with sub-classification schemes shows that the class of iconic gestures is not homogeneous. Table 2.2 displays the relations between these classifications. Despite differences between the accounts, e.g., due to the domain of investigation and the level of granularity, there are also some basic points of agreement. Although researchers differ in their terminology, there is one category identified by most of them: *posturing* gestures characterized by the hands forming a static configuration which functions as a model for the object itself. This representation technique has been referred to as ‘representing’ (Müller, 1998), ‘modeling’ (Kendon, 2004; Streeck, 2008), and ‘placeholder’ (Sowa, 2006). There is, further, at least some consensus about a class of *drawing* gestures (in the terminology of Wundt, Müller and Streeck) in which the hands trace the outline of an object. In addition, there is a set of gestures remaining—henceforth termed *shaping* gestures—which fall into Müller’s category of ‘modeling’ gestures. For these, a varying number of categories are employed; whereas Wundt and Kendon did not identify any further sub-types, Sowa and Streeck proposed several different categories of shaping techniques. Finally, there is agreement among Müller (1998), Kendon (2004), and Streeck (2008) with respect to a category of *pantomime* gestures in which the hands simulate an activity pantomimically. While Müller and Kendon termed this technique generally ‘acting’ or ‘enactment’, Streeck identified three sub-types of pantomime gestures (‘handling’, ‘acting’, and ‘pantomime’).

Table 2.2: Overview of classification schemes for gestural representation techniques from literature.

Wundt (1973)	**Müller (1998)**	**Kendon (2004)**	**Sowa (2006)**	**Streeck (2008)**
Drawing	Drawing	Depiction	Dimensional, Surface Property	Drawing
Plastic	Modeling			Bounding
				Scaping
				Making
				Marking
	Representing	Modeling	Spatial Relation, Placeholder	Modeling
	Acting	Enactment		Handling
				Acting
				Pantomime
				Abstract motion
				Self-marking
				Model-world making

2.2.4 Linguistic Factors

Coverbal gesture has long been considered a by-product of language production (Kendon, 2004). Indeed, upon closer inspection, gestures were found to be influenced by the conceptual, syntactic, and lexical structure of concomitant speech.

Kita and Özyürek (2003) proposed the *Interface Hypothesis* according to which gestures are generated at the interface between spatio-motoric thinking and language production processes. That is, in this view, gestural depiction is not only shaped by spatio-motoric properties of the referent, but also by how information is organized linguistically.

Initial evidence for the Interface Hypothesis came from a cross-linguistic study by Kita and Özyürek (2003), in which they could show that the packaging of content for gestures parallels linguistic information packaging. Speakers of Japanese, Turkish and English had to re-tell cartoon events for which their languages provide differing means of encoding. English speakers, for example, used the verb 'swing' to describe a cartoon character's action, encoding an arc-shaped trajectory, whereas Turkish and Japanese speakers employed a trajectory-neutral, change-of-location predicate such as 'move'. Participants' gestures followed their linguistic information packaging in so far as Japanese and Turkish speakers were more likely to produce straight gestures, while most English speakers produced arced gestures. In another cartoon-event, the character rolled down a hill. Again, speakers of English typically described this by combining manner and path of the movement in a single clause (e.g. 'he rolled down'), accompanied by a single gesture encoding both semantic features. In contrast, Turkish and Japanese speakers encoded manner and path separately in two clauses (e.g. 'he descended as he rolled') and also used two separate gestures for these two features.

Further evidence along the same lines comes from a study comparing the gestures of native Turkish speakers, whose proficiency levels in English as their second language vary (Özyürek, 2002). Advanced L2 speakers typically encoded manner and path information in one clause, just as native English speakers do, and the gestures they used followed this packaging of information. In contrast, L2 speakers at lower proficiency levels typically used two-clause constructions in speech, accompanied by separate gestures for manner and path, as they are used to doing in their native language. Özyürek concluded that syntactic packaging in speech shapes information packaging in gestures in both L1 and L2.

Another subsequent study showed that this effect also occurred when stimulus events were manipulated in order to make first language speakers of English produce one- and two-clause descriptions of manner and path (Kita et al., 2007). In line with previous evidence, one-clause utterances conveying both manner and path were found more likely to be accompanied by gestures expressing manner and path simultaneously. Two-clause utterances, in contrast, were more likely to be accompanied by gestures conveying manner and path separately. That is, the use of manner and path representing gestures depended on the syntactic construction used in speech.

Recently, Gullberg (2010) analyzed how the semantics of placement verbs affect co-speech gestures. French, which typically uses general placement verbs like 'mettre' (English: 'put'), was contrasted with Dutch, which uses a set of fine-grained posture verbs such as 'zetten' (English: 'set/stand'), or 'leggen', (English: 'lay'). Analysis of concomitant gesture production in the two languages revealed a patterning toward two distinct, language-specific event representations. The object of the placing action is an essential part of the Dutch representation, and the gestures typically represent the types of the object to be placed. French speakers instead focus only on the (path of the) placement movement.

Impact on gesture has also been observed when problems in verbalization occur. Bavelas et al. (2002) investigated to what degree the verbal encodability of a stimulus affects gesture use. They compared the descriptions of two different stimuli, both being highly visual and easily to be described with gestures, but in one condition (stimulus: familiar geometric patterns), speakers were likely to have a strong verbal vocabulary available, whereas in another condition (stimulus: an unusual dress), they would be unlikely to have the necessary vocabulary. Although the poorer verbal encodability of the dress did not produce a higher gesture frequency, this stimulus did elicit a higher proportion of nonredundant gestures and more deictic expressions. Bavelas et al. conclude that speakers seemed to rely on gestures when words were not readily available.

2.2.5 Discourse Context

Gesture occurrence is not limited to the context of lexical affiliates in concomitant speech, or as Kendon (1987, p. 90) put it, "a speaker will select a model of formulation, not only in the light of a comparison between its adequacy of representation and the image that it is intended to convey, but also in the light of what the current communication conditions are". Rather, each single multimodal utterance is integrated into a greater context. It is carried out within a larger utterance fulfilling a particular function in order to achieve an overall communicative intention, which itself occurs within a particular situation including specific people talking to each other in a specific environmental situation. Gestures are, thus, embedded in cascading levels of discourse context (Gerwing and Bavelas, 2004). Furthermore, there is indeed evidence that gesture use is sensitive to the different levels of the overall discourse context.

Initial Common Ground A considerable amount of studies have investigated the role of *common ground*, which is the sum of mutual, common or joint knowledge, beliefs and suppositions that interlocutors share, providing the background for their interaction (Stalnaker, 1978; Clark, 1996).

With regard to gesture frequency, results are mixed. Holler and Wilkin (2009) used a between-subjects design with a 'common ground' condition, in which some shared knowledge about the stimulus material was experimentally induced, and the 'no common ground' condition, in which participants did not share any experimentally induced common ground. The findings showed that speakers gesture at higher rates (gestures per one hundred words) when they shared a common ground. Holler and

Wilkin conclude that gestures play an important communicative function, even when speakers convey information that is already known to their addressee. On the other hand, Jacobs and Garnham (2007) used a cartoon story re-narration setting and found that gestures were produced at lower rates when speakers believed the listener was also able to view the content of the stimulus, i.e., when speaker and listener had common ground.

Other studies went beyond the pure consideration of gesture frequency, and investigated the qualitative nature of gestures. Gerwing and Bavelas (2004) asked participants to describe play actions they performed with a certain set of toys, either to someone who played with the same or a different set of toys. Their findings revealed that speakers used less precise, less complex, and less informative gestures when talking to participants with whom they shared mutual knowledge as compared with those who did not have access to the same knowledge. Similarly, Holler and Stevens (2007) found that gestures' spatial extent was sensitive to information structure in the representation of size information. They used a referential communication task which involved participants locating a target entity in an array of other entities of different sizes. Their analysis focused on the representation of the size of particularly large entities in the array. It was found that entities were represented gesturally as significantly smaller when their actual size was already known to the addressee, as compared to when addressees had no pre-existing shared knowledge regarding the respective entities' size. Parrill (2010) found an effect of common ground on the encoding of semantic information in gestures. Participants described a target event in which a cat 'melted down some stairs'. Her analysis focused on the number of times speakers mentioned the stairs, i.e., the ground component, in their descriptions. When the content of the stimulus cartoon was absent from speaker-addressee common ground, production of ground information in gestures was increased.

In sum, research results regarding the correlation of gesture use and common ground yielded mixed results. While some suggest that gestures are produced at a lower rate, less precisely, and with less information in situations with common ground, others provided evidence that gesture rate is increased, and that gestures remain similar with regard to their form and the amount of information they convey.

Information Structure Apart from the knowledge that interlocutors share from the outset of a conversation, common ground also accumulates over the course of it. *New* information is defined as what the listener is not expected to know already. *Given* information, in contrast, is defined as what the listener is expected to already know (Haviland and Clark, 1974)[3].

[3] This dichotomy has been referred to in the literature with a variety of terms and meanings, e.g. differentiating between 'focus' and 'presupposition', between 'hearer-old' and 'hearer-new', and between 'discourse-old' and 'discourse-new' (cf. Prince, 1981)

Within these lines, Levy and McNeill (1992) as well as McNeill (1992) employed the notion of 'communicative dynamism' as a variable that correlates with gesture use[4], this being defined as the extent to which the message at a given point is "pushing the communication forward." The authors argued that sentences which are low in communicative dynamism, are typically not accompanied by gesture. Gestures, rather, pick out the sentence elements which are of high communicative dynamism. And, the more complex a gesture is, the higher are the peaks in communicative dynamism. As a measure of communicative dynamism, the amount of linguistic material used to make the reference was chosen. Pronouns have less communicative dynamism than full nominal phrases, which have less communicative dynamism than modified full noun phrases. This implies that the communicative dynamism can be estimated by looking at the syntactic structure of a sentence.

Gerwing and Bavelas (2004) went beyond pure gesture frequency to consider how the physical form of gestures mark the status of information. Their findings revealed that speakers marked new information with gestures that were larger in size and more precisely articulated. Given information, in contrast, was accompanied by gestures that were smaller and less well articulated.

Previous Gestures Another correlate with gesture use is the set of gestures a speaker has used previously in a given discourse. Quek and McNeill (2000) introduced the notion of 'catchments' for gesture features which recur in at least two (not necessarily consecutive) gestures. The authors interpret catchments as clues to cohesive linkages in the discourse; a common discourse theme is expected to produce gestures with recurring features. Examples for such catchments are, for instance, a single moving hand (referring to a single moving entity), a round handshape referring to a ball, or a particular spatial configuration of the hands referring to the relative position of two entities (McNeill, 2005).

2.2.6 Inter-Individual Differences

People differ substantially in the way they use gestures while speaking, as Gullberg (1998, p. 51) put it: "One of the most salient aspects of gesture is that people differ in their use of it." This becomes particularly obvious in gesture frequency: while some speakers rarely move their hands at all, other use gestures all the time. Empirical evidence for this observation comes from a number of studies. For instance, Jacobs and Garnham (2007) reported from a narrative task study that large differences in gesture rate were observed among the participants. Krauss et al. (1996) reported gesture rates ranging across speakers from 1.0 to 28.1 gestures per minute in a referential communication experiment involving abstract line drawings as stimuli. Similarly,

[4] This notion was originally introduced by Firbas (1971).

gesture rates were found to range from 0.5 to 30 gestures per minute when participants were retelling cartoon stimuli (Krauss et al., 1996; Rauscher et al., 1996).

In addition to gesture frequency, other aspects of gesture use are also subject to individual differences. McNeill claimed that different speakers display the same meaning in the same context in different ways: "Lacking standards of form, individuals create their own gesture symbols for the same event, each incorporating a core meaning but adding details that seem salient, and these are different from speaker to speaker" (McNeill, 1992, p. 41). Marsh and Watt (1998) found that individuals vary in their preferences for particular gestural techniques of representation. In their study on iconic hand gestures as a mode of human-computer interaction for the input and manipulation of objects and shapes within 3D environments, some participants preferred virtual depiction over substitutions, while others employed both strategies. Webb (1996) compared metaphoric gestures across subjects to determine whether metaphoric gestures are idiosyncratic or shared among speakers. It was found that although speakers share a single, limited lexicon of metaphoric gestures, individuals used different subsets of that lexicon. Moreover, regarding handedness in the use of gestures, a right or left hand preference was reported (Kimura, 1973a,b).

Widely unexplained is the question of how these differences among individuals arise. Although a number of correlates are possible, such as cultural background, personality traits, or cognitive skills, only very few studies have addressed this question so far. Regarding gesture frequency, Hostetter and Alibali (2007) found that individual differences in gesture rate are associated with the speakers' verbal and spatial skills. Individuals with low verbal fluency and high spatial visualization skill were found to gesture the most. Likewise, Chu and Kita (2009) examined the relationship between co-speech gestures and gestures produced during a silent mental rotation task ('co-thought' gestures). They found that how often people produce co-speech gestures in a verbal description task can be predicted by how often they produce co-thought gestures. Chu and Kita argue that gestures originate from an action-based representation system, and how often people gesture is partly determined by their degree of preference for using this representation system. For individual differences in gestural hand preference, Kimura (1973a,b) found a positive correlation with speakers' handedness. Right-handed participants' co-speech gestures occurred primarily with the right hand, while self-touching movements occurred equally often with either hand. Lausberg and Kita (2003) narrowed the scope of this finding, however, as they presented data suggesting that hand choice in iconic gestures with observer viewpoint is predominantly influenced by the content of the message: participants preferred their left hand to refer to objects that had been on the left in the stimulus scene. Analogously, the right hand was preferred for objects that had been on the right in the stimulus scene.

2.3 Theoretical Models of Gesture Production

The previous section identified a number of factors to which gesture use is sensitive. This set of factors should, although not necessarily complete, at least be considered by a comprehensive simulation account of gesture production. Now, the degree to which these crucial aspects are considered in psychological and psycholinguistic theories about the production of gestures will be investigated. For this purpose, the following five frameworks will be reviewed: (1) a process model for lexical access via lexical gestures (Krauss et al., 2000), (2) the sketch model (de Ruiter, 1998, 2000, 2007), (3) the interface model (Kita and Özyürek, 2003), (4) growth point theory (McNeill, 1992, 2005; McNeill and Duncan, 2000), and (5) the gestures-as-simulated-action framework (Hostetter and Alibali, 2008).

Some of these models investigate the production of speech and gestures in terms of representations and processes (RP models, henceforth). RP models are information processing models in which information "is taken to mean 'representation' " (de Ruiter, 1998, p. 6). Building blocks of RP models are usually information processing modules interfacing with each other via representations. RP models are particularly adequate for checking the coherence of psycholinguistic theories using computational simulations. Besides RP-based modeling, there are also non-modular interactive production models proposing an interactive and dynamic development of gestures. There is no clear distinction between representations and processes in these models.

A process model for lexical access via lexical gestures Krauss et al. (2000) proposed a model that makes the assumption that an important function of gestures is to facilitate lexical retrieval. This is based on empirical evidence that the restriction of gesturing adversely affects speech (see Rauscher et al. (1996) for a review). The scope of the model includes iconic and metaphorical gestures, labelled as *lexical gestures* (cf. Section 2.1.1).

Gesture generation is based on a memorial representation of a so-called *source concept* that consists of a set of semantic features, such as size, color, shape etc. that are encoded in propositional and/or spatial format: some are represented in both formats, whereas others are represented only spatially or only propositionally (see Figure 2.3). The central hypothesis in this account is that "lexical gestures derive from non-propositional representations of the source concept" (Krauss et al., 2000, p. 268). The question of whether or not at all a gesture is produced is, therefore, answered implicitly: this depends on the existence of non-propositional semantic features in the source concept. If any features are available, these are then transformed into a set of *spatial/dynamic specifications*. These are abstract properties of movements which are in turn translated into a *motor program*, a set of instructions for executing

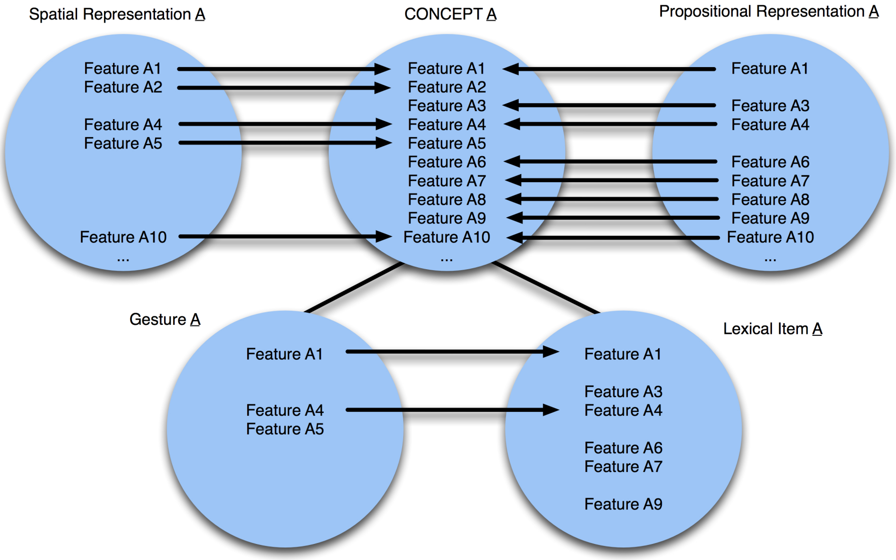


Figure 2.3: The feature-based mental representation of a source concept and its reflection in gesture and speech. Redrawn from Krauss et al. (2000)).

the gestural movement. Krauss et al. (2000, p. 276) were confident that their feature model provides a "satisfactory way of accounting for a gesture's form." How exactly the mapping of semantic features onto gesture form is realized, however, they did not exactly elaborate on. The authors also do not consider any sub-classification of lexical gestures in terms of gestural representation techniques.

Krauss and colleagues' model is based on Levelt's speech production model (Levelt, 1989). The key idea—gestural facilitation of lexical retrieval—is realized in the following way: the lexical gesture provides input to the phonological encoder. Because the features conveyed in the gesture may also be features of the sought-after lexical item, it may facilitate the retrieval of the word form by a process of cross-modal priming. Krauss and colleagues haven chosen the phonological encoder (and not the grammatical encoder) as an anchor for the lexical facilitation link as there is empirical evidence from tip-of-the-tongue studies that retrieval failures are phonological rather than semantic. Interactivity between speech and gesture production processes, however, is restricted to these two points of interaction. Neither an earlier influence of speech production processes on gesture production, nor any other contextual factors to which gesture production is sensitive, is designated.

The Sketch Model Similar to the model proposed by Krauss et al. (2000) is de Ruiter's *Sketch Model* (de Ruiter, 1998, 2000, 2007), displayed in Figure 2.5. Al-

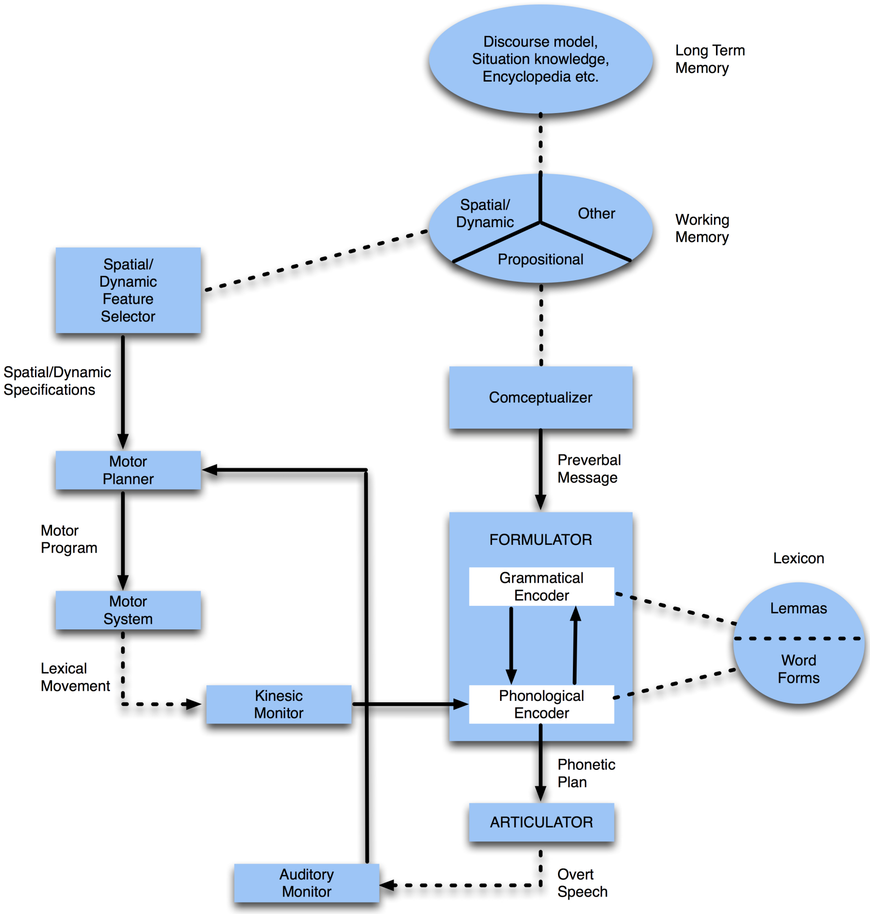


Figure 2.4: The cognitive architecture for speech and gesture production processes based on Levelt's speech production model. Redrawn from Krauss et al. (2000).

though de Ruiter assumes, in contrast to Krauss and colleagues, that gesture is a communicative device, both models are alike in that they assume that gestures are generated before the linguistic formulation process takes place, as well as that speech and gesture production are, to a large extent, independent processes.

According the Sketch Model, gesture generation beginning in Levelt's conceptualizer is responsible for the selection of information to be expressed and the assignment of a perspective for the expression. As both tasks are structurally similar for gesture and speech, they are performed in one and the same module for both modalities: the communicative intention is divided into two parts, a propositional representation which is transformed into a preverbal message, and an imagistic representation that is transformed into a so-called *sketch*. Depending on the type of gesture, the sketch contains different information: for pointing gestures, it contains a vector in the direction of the referent position; for pantomimic gestures, it references a motor program encoded in the sketch; for iconic gestures, it consists of spatio-temporal representations (so-called *trajectories*). As an example, de Ruiter (1998) cites a gesture indicating a rectangular sign post for which the sketch would contain a large rectangle. Apart from the conceptualizer, speech and gesture are processed independently. In the gesture planner, the sketch is transformed into a physical gesture. As in the above-mentioned example, this is achieved by tracing the rectangular trajectory. De Ruiter, however, is aware of the question of "whether for all gestures a pre-existing action schema can be identified, or whether there are gestures (esp. the ones that are not pantomimes) that rely on newly generated motion patterns created ad-hoc for gestural communication." (de Ruiter, 2007, p. 30). For the case that the sketch contains more than one gesture, e.g., a pointing vector and an iconic trajectory, de Ruiter (2000) touched the problem of fusing multiple gestures, proposing a mechanism to reduce the motoric degrees of freedom.

De Ruiter further pointed out three computational problems to be considered in the gesture planner. First, there is empirical evidence that gesture planning must be sensitive to effects of recipient design. Second, de Ruiter mentioned the problem of hand allocation, which is the question of whether the gesture is to be performed with the left or right hand, or with both. Third, environmental constraints have to be considered in the process of gesture generation. The speaker has to take restrictions imposed by objects or other persons in the environment into account.

The Interface Model Kita and Özyürek (2003) proposed a model based on the idea that language shapes iconic gestures. This hypothesis accounts for empirical findings from cross-linguistic studies (Kita and Özyürek, 2003), second language acquisition (Özyürek et al., 2005), and psycholinguistic experiments (Kita et al., 2007), all of which show that information packaging for gestures parallels linguistic packaging in concomitant speech. Therefore, Kita and Özyürek (2003, p. 17) suggest that "gestures originate from an interface representation between speaking and spatial thinking".

In the proposed model, gestures, "are generated from a general mechanism of action generation which can be used in both purely communicative and practical purposes" (Kita and Özyürek, 2003, p. 29). The content of a gesture is determined by three factors: (1) the speaker's communicative intention, (2) action schemata selected on the basis of features of imagined or real space, and (3) bidirectional interaction between speech and gesture production processes to account for the influence on gestures by linguistic constraints of the speaker's language.

The model, as displayed in Figure 2.6, is based on Levelt's conceptualizer, which is split into two halves: the *Communication Planner* to generate communicative intentions, and the Message Generator to formulate a verbal utterance from a propositional representation. As the first step in the production process, the Communication Planner

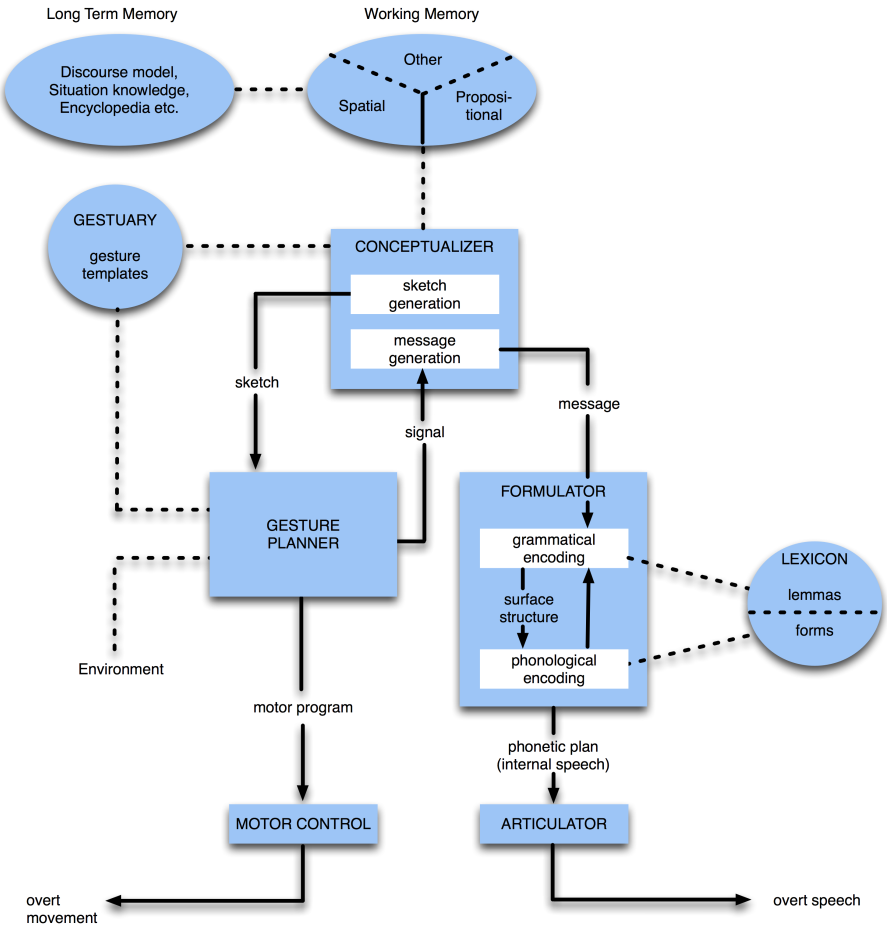


Figure 2.5: The sketch model of speech and gesture production based on Levelt's speech production model. Redrawn from de Ruiter (2000).

generates communicative intentions making a first rough decision on the information to be expressed and deciding which modalities should be involved. These specifications of intent are sent to the Action Generator and the Message Generator. The Action Generator generates a spatio-motoric plan for the gesture to be performed. It has access to the part of working memory where relevant spatial imagery—action schemata based on features of imagined or real space—is now active. The Message Generator, taking into account the communicative goal and the discourse context, formulates a propositional preverbal message to be sent to the Formulator. Both generators constantly exchange information, which also involves transformations between the two informational formats. Additionally, the Message Generator receives feedback from the Formulator as to whether a proposition can be readily verbalized or not.

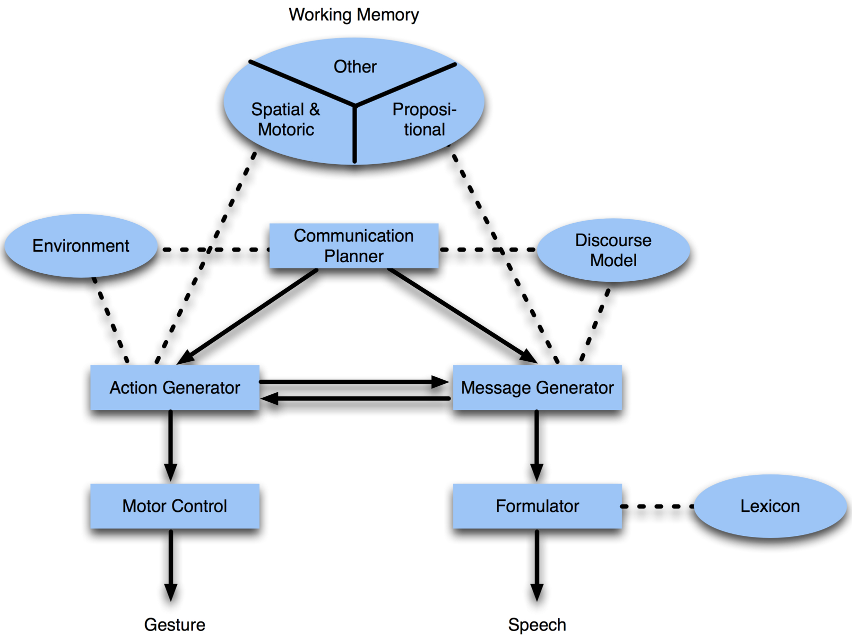


Figure 2.6: The interface model of speech and gesture production. Redrawn from Kita and Özyürek (2003).

These interactions among the three components are thought to go on until equilibrium is reached. Not until this point, verbal formulation starts and the spatio-motoric representation generated by the Action Generator is sent to Motor Control for execution. How exactly the mapping of an action representation onto the spatio-motoric gesture plan is realized is not further specified, however.

Growth Point Theory According to the growth point theory, "speech and gesture are two aspects of one process" (McNeill, 1992, p. 245): the process of speech and gesture production is regarded as a dialectic of gesture imagery and linguistic structure. Gestures and speech emerge from *growth points* (GPs) which are minimal units that combine linguistic and imagistic components as equal parts. That is, in line with the interface model, gestures are not based solely on visuospatial imagery, and they are influenced by linguistic factors. Growth points are dynamic units, i.e., they grow and unfold into speech and gesture over the course of time. This development is due to contextual influences in terms of background and contrast, as well as the dialectic between imagistic and linguistic parts. Further, McNeill (2005, p. 131) states that "in the GP, imagery is categorized linguistically" and "never purely visuospatial". That is, gestures are assumed to be influenced by linguistic factors.

Gestures as Simulated Action Framework Figure 2.7 presents the gesture-as-simulated-action (GSA) framework proposed by Hostetter and Alibali (2008). The framework asserts that "gestures emerge from the perceptual and motor simulations that underlie embodied language and mental imagery" (Hostetter and Alibali, 2008, p. 502). The central point of the framework, in line with embodied theories of cognition, is the idea that action and perception influence each other mutually. Language processing and mental imagery are accomplished via simulations of perception and action. These simulations involve activating premotor actions. In other words, this activation has the potential to spread to motor areas and to be realized as overt action such as gestures. Speech and representational gesture production are based on the same underlying system of mental imagery, namely on simulated action and simulated perception.

According to the proposed model, three factors contribute to whether activation involved in simulation will be realized as an overt movement, such as a gesture. The first factor is the *strength of activation* of the underlying representation, which is assumed to spread from the planning stage to the execution stage. Representational gestures are assumed to be more likely produced with more activation strength in terms of a higher degree of active processing. The GSA framework assumes differences in activation strength for different types of mental imagery. More activation underlies motor imagery than spatial imagery, and more activation underlies spatial imagery than visual imagery. The second factor is the individual speaker's current *gesture threshold*. The gesture threshold is "the level of activation beyond which a speaker cannot inhibit the expression of simulated actions as gestures" (Hostetter and Alibali, 2008, p. 503). This level may vary from speaker to speaker due to neural factors, cognitive factors, and aspects of the social communicative situation which gives rise to the influence of contextual factors on gesture use in terms of, for instance, recipient design. The third factor is the simultaneous engagement of the motor system for speaking. During speech production, the activation of motor plans spreads quickly from the planning to the execution stage. Because actions by the mouth and the hands have a common underlying motor system, the activation of motor plans for hand gestures is also likely to spread from the planning to the execution stage. Hence, gestures often accompany speech.

Discussion In Table 2.3, an overview of the models is given with respect those factors, gesture was shown to be sensitive to (Section 2.2). These factors roughly devide into three kinds. First, since the meaning of iconic gestures is explained by similarity or resemblance to their referent, the way how meaning is mapped onto gesture form is decisive. This implies that iconic gesture use is dependent on (1) the nature of the *representations* that underlie gestures, as well as the question (2) how this representation is *transformed* into gesture form constrained by (3) the use of different gestural *representation techniques*. Second, there is a body of evidence suggesting

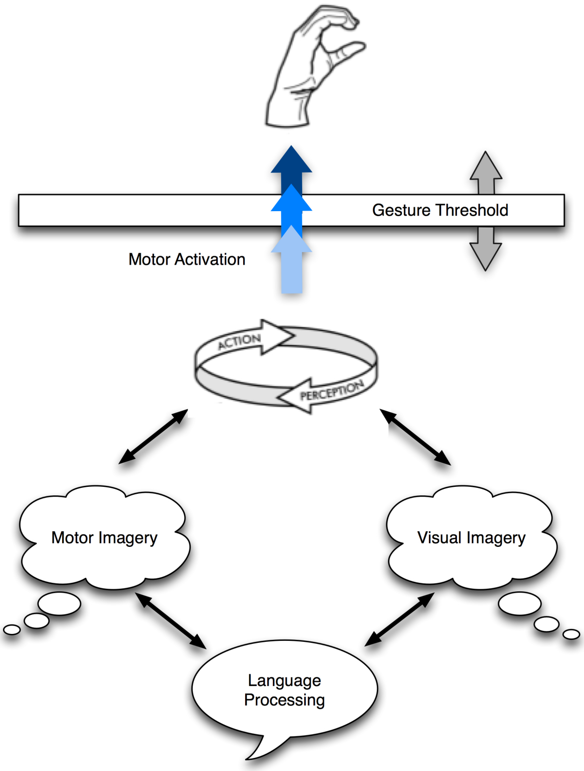


Figure 2.7: The gesture-as-simulated-action framework is based on the assumption that gestures arise from perceptual and motor simulations. Redrawn from Hostetter and Alibali (2008).

that a gesture's form is also influenced by (4) specific *discourse contextual constraints* as well as its (5) *linguistic context*. Finally, (6) *inter-individual differences* in gesturing are quite obvious. This set of factors should, although definitely not complete, at least be considered by a comprehensive simulation account of gesture production.

Regarding the nature of the representations underlying gesture production, the models essentially agree that gestures arise from visuo-spatial representations. The frameworks, however, make different claims about whether these spatial representations are stored as non-decomposable units that are retrieved holistically, or as sets of spatial features. Krauss et al. (2000) assume in their Lexical Access Model that memorial representations reflected in gestures are made up of sets of elementary features which can, in turn, be mapped onto gesture form features. The other frameworks, in contrast, are based on the holistic imagery view. That is, these models are concerned with the problem of how to transform an image into gestural movements. To this end, the sketch generator in de Ruiter's Sketch model selects features of an image to be expressed in gesture. In the other frameworks, however, it remains unclear as to how this transformation is realized.

Concerning the influence of linguistic factors on gesture production, the models differ considerably. Whereas the Lexical Access Model and the Sketch Model do not assume gesture production to be sensitive these factors, Kita and Özyürek (2003) propose a bi-directional interaction between speech and gesture production processes. Along the same lines, in Growth Point Theory, interactions between spoken form and imagery occur continuously and in both directions. That is, there are mutual effects between imagistic and linguistic components. Finaly, the GSA framework is closely aligned with the Interface Model with regard to the influence of linguistic factors. Although Hostetter and Alibali (2008, p. 508) do not explicitly propose bi-directional communication between gesture and speech, they assume "that the way in which speakers simulate visuospatial events is influenced by the constraints of the speakers' languages."

Regarding the influence of discourse context, the frameworks agree that gesture use is sensitive to these factors with exception of Krauss et al. (2000) who do not explicitly mention any influence of this kind. The other models, despite assuming a general influence of discourse context, do not make any explicit assumptions about how exactly gestures are shaped by these factors. Additionally, the models differ concerning the type of context to be considered.

In the Sketch Model, gesture planning is designed to be sensitive to effects of recipient design as well as environmental constraints in terms of objects or other persons in the environment. Kita and Özyürek (2003) designed their communication planner to have access to a discourse model so as to take into account what has been communicated so far, and to project how discourse should develop to achieve the overall communicative goal. This way, the communication planner may give more prominence to certain information and, therefore, the discourse model exerts an (indirect) influence on gesturing. In Growth Point Theory, the growth point's unfolding in time is shaped by contextual influences in terms of background and contrast. The contextual background is constrained by external (social and material) conditions. The speaker shapes the background in such a way as to render the intended significant contrast with it possible. That is, meaning has a dual character of being simultaneously a focal point and an implied background. In the GSA framework, the gesture threshold level is, among other factors, biased by the social communicative situation giving rise to recipient design.

Inter-individual differences are only considered in the GSA framework. They come into play through the concept of the gesture threshold, which is speaker-specific. The model, however, does not deal explicitly with other inter-individual differences in gesture use, for instance concerning individual preferences for particular gesture types or gesture form features.

Finally, concerning the varying character of iconic gestures with regard to gestural representation techniques, none of the models considers adequately the multi-faceted class of iconic gestures. Only in de Ruiter's Sketch Model are gestures of different kinds treated differently in the sketch generator. The sketch contains different information depending on the type of gesture. For pointing gestures, it contains a vector in the direction of the referent position. For pantomimic gestures, a reference to a motor program is encoded in the sketch For iconic gestures, it consists of spatio-temporal representations. There is, however, no further differentiation of representation techniques in the iconic gestures category.

In conclusion, none of the theories accounts for *all* six key aspects of gesture production This is due to the models' focus on different aspects of gesture production. The Lexical Access Model and the Sketch model mainly emphasize the problem of planning gestures from a given representation, the Interface Model and Growth Point Theory have their focus on the interaction of speech and gesture production, and the GSA framework aims to provide an embodied perspective on gesture use which takes inter-individual differences into account. That is, there is, to date, no comprehensive theory of gesture production available.

Table 2.3: Overview of theoretical models of speech and gesture production with regard to (1) the nature of the representations that underlie gestures, (2) the form-meaning mapping, (3) how gesture and speech are integrated, (4) the consideration of the discourse context, (5) in how far they account for inter-individual differences in gesture use, and (6) regarding iconic gestures, whether the model takes different representation techniques into account.

	Representation	**Form-meaning mapping**	**Linguistic factors**	**Discourse context**	**Inter-individual differences**	**Representation techniques**
Lexical Access Model (Krauss et al., 2000)	Feature-based source concept representation	Transformation of semantic features into into spatial-dynamic representations [3]	Gestural facilitation of speech; gesture production prior to linguistic formulation	discourse model [1]	— [2]	—
Sketch Model (de Ruiter, 2000)	Imagistic representation	Gesture production prior to linguistic formulation	Recipient design, environmental constraints	— [2]	— [2]	—
Interface Model (Kita and Özyürek, 2003)	Spatial and motoric representations	Mapping of action representation on spatio-motoric gesture plan [3]	Bi-directional interaction between speech and gesture production processes	discourse context [1]	— [2]	—
Growth Point Theory (McNeill, 1992)	Visuo-spatial imagery	— [2]	Imagery is categorized linguistically	GP is shaped by context	— [2]	—
GSA Framework (Hostetter and Alibali, 2008)	Simulated action and simulated perception in mental imagery	— [2]	Simulations (and therefore gestures) are constrained by the speakers' languages	Gesture threshold biased by social factors	Individual gesture threshold	—

[1] Influence/interaction not further specified

[2] Not explicitly considered

[3] Details of the mapping are not specified

2.4 Summary

This chapter reviewed the research literature relevant in the scope of this thesis. Section 2.1 introduced the phenomena of gesture and narrowed the scope of this thesis to deal with one particular kind of hand and arm movement: spontaneous, unconventionalized, and speech-accompanying gestures which are characterized by a relatively transparent form-meaning relationship known as iconic and abstract deictic gestures.

Another objective in gesture research of importance for a computational generation model are the physical aspects of gestures and the question of how to describe them. This issue was addressed by exploring, firstly, how the hands' movements are organized temporally, and, secondly, with regard to gesture form features which describe the physical appearance of gestures. Concerning the latter, a four-feature system turned out to be widely accepted to describe gesture forms, namely by (1) handshape, (2) orientation, (3) position, and (4) movement features.

Concerning the crucial question of how the physical features of a gesture are determined in the production process of gestures, light was shed on relevant literature in order to identify factors by which the use of gestures is, at least, modulated:

Mental representation With regard to the question of what kind the underlying representation of gestures is, evidence was presented for the claim that gestures derive, at least in part, from spatial or imagistic representations.

Form-meaning mapping Concerning the question of how meaning is transformed into an overt, observable gesture, a few pictorial strategies were identified from research in different domains, e.g., positive correlations between movement direction in referent motions and movement direction gestures, or, between salient planes in referent objects and hand orientation in gestures. None of the reviewed studies, however, was able to provide a comprehensive account of the form-meaning relationship in iconic gestures.

Representation techniques The class of iconic gestures contains a variety of different techniques of representation. Obviously, for each of these techniques, the transformation of imagistic meaning into a corresponding gestural form is different. Attempts were reviewed which aimed to further sub-divide the class of iconic gestures in particular. Despite the differences, there are some basic points of agreement among researchers with respect to the following techniques of gestural representation:

- *Posturing*: Gestures are characterized by the fact that the hands form a static configuration standing as a model for the object itself.
- *Drawing*: The hands trace the outline of an object.
- *Shaping*: Gestures are characterized by a sculpting or contouring movement of the hands

- *Pantomime*: The hands simulate an activity pantomimically.

Linguistic factors Gestures were found to be shaped by their linguistic context in terms of the conceptual, syntactic, and lexical structure of concomitant speech. In particular, information packaging for gestures was found to parallel linguistic packaging. Further evidence also suggests that linguistic choice and verbal encoding problems are positively correlated with gesture use.

Discourse context Gestures are, in addition, sensitive to the conditions of communication. There is evidence for an impact of contextual factors such as common ground, information structure, and the previously preformed gesture(s).

Inter-individual differences Inter-individual differences are another important issue as people differ substantially in the way they use gestures while speaking. This concerns the rate of gesturing as well as preferences for particular gesture types and handedness.

The chapter was completed from a psychological and psycholinguistic theoretical perspective about the production of gestures. Five models were reviewed with respect to fundamental design choices concerning the abovementioned factors. None of the models, however, provides a comprehensive account of how gestures are produced with consideration for all the factors shown to affect gesture use.

CHAPTER **3**

Generating Gestures—The Computational Perspective

The previous chapter provided empirical, theoretical, and psycholinguistic perspectives on how gestures are produced. This chapter will show how *computational* approaches investigate the challenge of building systems that generate gestural behavior to be realized with virtual agents.

An essential prerequisite for the generation of iconic gestures is the representation of imagistic knowledge. This chapter, thus, will begin with a review of work concerning the representation of visuo-spatial and shape-related knowledge in Section 3.1. Following this, related research in gesture generation will be presented and broken down into categories of model-driven attempts (Section 3.2.1), customization of generated behavior (Section 3.2.2), and data-based attempts (Section 3.2.3). The chapter will be concluded with a discussion in Section 3.3.

3.1 Representation of Imagistic Knowledge

Theoretical models of gesture production basically agree on the fact that gestures derive from some kind of spatial representation (Section 2.3). It is, therefore, reasonable to assume a visuo-spatial representation on which the production of gestures is based.

A prominent modeling approach for visuo-spatial imagery, which has proven to be computationally efficient, is to use two- or three-dimensional, matrix-like structures that are cell-wise occupied by object entities (Glasgow, 1993; Kosslyn, 1987). Such arrays represent the objects' visual appearance as well as their spatial relationship to one another, and can be hierarchically refined to locally allow for a higher level of detail.

Another prominent account is the *3-D model* by Marr and Nishihara (1978) which employs perceptual object axes as basic elements of shape. Axes were hierarchically

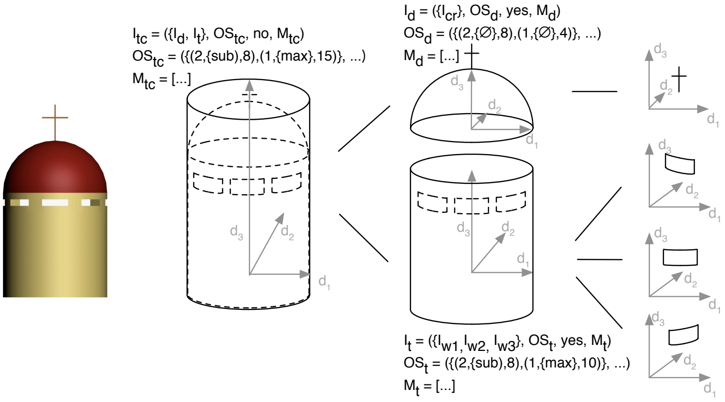

I_{tc} = ({I_d, I_t}, OS_{tc}, no, M_{tc})
OS_{tc} = ({(2,{sub),8),(1,{max},15)}, ...)
M_{tc} = [...]
I_d = ({I_{cr}}, OS_d, yes, M_d)
OS_d = ({(2,{∅},8),(1,{∅},4)}, ...)
M_d = [...]
I_t = ({I_{w1},I_{w2}, I_{w3}}, OS_t, yes, M_t)
OS_t = ({(2,{sub),8),(1,{max},10)}, ...)
M_t = [...]
d_1
d_2
d_3

Figure 3.1: A sample representation of a church tower as an Imagistic Description Tree (IDT). Each node contains an imagistic description including a set of childnodes *I*, an object schema *OS*, and a transformation matrix *M*.

arranged according to different levels of granularity. The disposition of a lower-level axis (part) with respect to the higher-level axis (whole) is explicitly encoded. The 3D model, however, is not able to represent objects without a dominant axis such as coins or spheres. Biederman (1987) proposed a competing model which contained a set of volumetric primitives (geons) as qualitative descriptions of object parts. The *geon model*, however, lacked part-whole relations in the sense of different levels of abstraction. Furthermore, all volumetric approaches inherently define 3D shapes and do not support underspecification, which is highly relevant for gesture use.

More recent models of computational imagery have utilized a graph structure that represents an object or multi-object scene as a tree, with geometrical primitives at the leaf nodes and geometrical transformations (among other properties) at the intermediate nodes (Croft and Thagard, 2002).

A similar model of visuo-spatial imagery, called *Imagistic Description Trees* (IDT) (Sowa, 2006; Sowa and Wachsmuth, 2009), was developed based on empirical data to represent shape-related information acquired via gesture and speech for use in a gesture understanding system. It is, thus, designed to cover all meaningful visuo-spatial features one finds represented in shape-depicting iconic gestures. Important aspects include (1) a tree structure for shape decomposition, (2) extents in different dimensions as approximations of shape, and (3) the possibility of underspecified dimensional information.

The IDT model unifies the abovementioned models from Marr and Nishihara (1978), Biederman (1987), as well as the semantic theory on dimensional adjectives Lang (1989). In this theory, so-called 'object schemas' were defined to describe basic gestalt properties of objects.

Figure 3.1 illustrates an IDT model for a certain building. Each node in an IDT contains an Imagistic Description (IMD), which includes an object schema representing the shape of an object or object part, respectively (Lang, 1989). An object schema contains up to three axes, each representing a spatial extent and assigned a dimensional attribute like 'max' or 'sub' to classify its extent relative to the other axes. Each axis may cover more than one dimension to account for rotation symmetries (becoming a so-called 'integrated axis'). An object boundary is defined by a profile vector that states symmetry, size, and edge properties for each object axis or pair of axes. The size property reflects change of an extent as one moves along another axis; the edge property indicates whether an object's boundary consists of straight segments that form 'sharp' corners, or of smooth, curved edges. The links in the tree structure represent the spatial relations that hold between the parts and wholes and are defined quantitatively by transformation matrices. It is thus possible to represent decomposition and spatial coherence.

3.2 Gesture Generation—State of the Art

The challenge of generating communicative behavior for virtual agents has typically been met in a series of three stages of processing: (1) content planning, (2) behavior planning, and (3) realization of the planned behaviors (cf. Kopp et al., 2006; Vilhjalmsson et al., 2007). This modularization corresponds to the stages usual in Natural Language Generation (Reiter and Dale, 2000). As an interface between these stages, representation formats allow delivery of input to the next stage and feedback of data to the previous stage where appropriate. See Figure 3.2 for an illustration of the modules and interfaces.

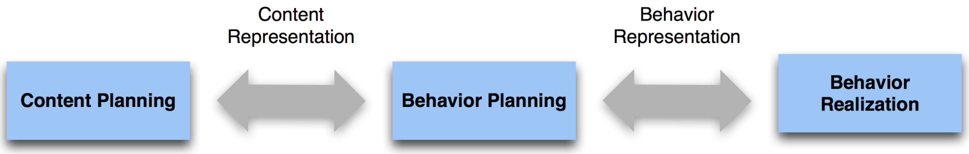


Figure 3.2: The process of generating communicative behavior in three stages of content planning, behavior planning, and realization.

Content planning, as the first stage, is concerned with the selection and structuring of domain-specific information to be conveyed. The resulting content representation is taken as input by the behavior planning module, which addresses the planning of coordinated linguistic terms and gestures. Finally, the realization module takes the behavior specification and turns it into synthesized speech and gestural movements for the virtual agent.

Gesture generation models, like the one to be developed in this thesis, fall into stage of behavior planning—sometimes realized in close interaction with content planning. With regard to behavior realization, the virtual agent Max (Kopp and Wachsmuth, 2004) will be employed in this thesis. In the Max system, the interface between behavior planning and surface realization is determined by the using the *Articulated Communicator Engine* (ACE) for behavior realization. ACE processes input that is specified according to the *Multimodal Utterance Representation Markup Language* (MURML, Kranstedt et al., 2002; Kopp, 2003). Therefore, output from the the behavior planning stage is to be specified accordingly[1].

3.2.1 Model-based Gesture Generation

Model-based behavior generation aims to formalize (aspects of) human communicative behavior, thus providing the basis for behavior simulation. In general, models are typically not built from scratch. Rather, there are two standard methods of building formal models of communicative behavior, either on the basis of empirical insights, or as a refinement of existing models from literature. In the former case, insights from the analysis of natural human-human communication data, in terms of rules and relationships, are implemented. In the latter case, a theoretical model is realized via computational means. Combinations of the two methods are also reasonable.

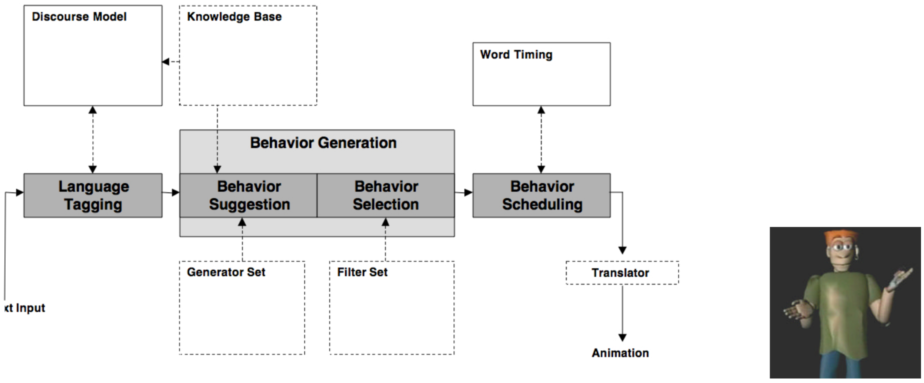
Lexicon-Based Gesture Generation

The first systems investigating the challenge of iconic gesture generation were *lexicon-based* approaches. In general, these systems were characterized by a straightforward mapping of meaning onto gesture form.

The **Behavior Expression Animation Toolkit (BEAT)** was among the first of a new generation of toolkits to allow the generation of synthetic speech alongside synchronized nonverbal behaviors, such as hand gestures and facial displays, to be realized with an animated human figure (Cassell et al., 2001). This approach of mapping text onto multimodal behavior was characterized by representing linguistic and social context and applying behavior generation rules based on empirical results.

The BEAT system was built to be modular, extensible, and to operate in real-time. The toolkit processes text in three major modules, interfaced with an XML-based representation of information. The pipeline approach provides support for user-defined filters and knowledge bases.

[1] Other representation languages have been designed as well, e.g. BEAT/Spark (Cassell et al., 2001) or APML (DeCarolis et al., 2004). Moreover, there are ongoing activities to define the *Behavior Markup Language* (BML) as an approach to unify existing formats (Kopp et al., 2006; Vilhjalmsson et al., 2007).



(a) The pipeline architecture consists of three modules to map typed input texts onto synthetic multimodal behavior. Reprinted from Cassell et al. (2001).

(b) Output utterance of the BEAT system for the text "Are you a good or a bad witch?".

Figure 3.3: Architecture of the Behavior Expression Animation Toolkit (BEAT) and an example utterance generated by the system.

In the first module, *Language Tagging*, input text gets broken down into clauses representing propositions and tagged with linguistic and contextual information: information structure (theme and rheme), word newness, contrast, as well as object and action positions in the text.

The output of language tagging is further processed in the *Behavior Generation* module, which consists of two parts. The first, *Behavior Suggestion*, augments the tagged text input with suggestions for appropriate nonverbal behaviors, based on a set of rule-based behavior generators including a 'beat gesture generator', a 'surprising features iconic gesture generator', an 'action iconic gesture generator', and a 'contrast gesture generator'. Each of these is in charge of realizing the mapping from annotated text onto particular nonverbal behaviors on the basis of empirical findings. For instance, the 'action iconic gesture generator' implements the strong tendency of rhematic actions to be accompanied by iconic gestures.

In the second step of behavior generation, this over-generated set of suggested behaviors is filtered down to yield the set actually used later in the animation, taking conflict resolution and priority into account. The system, thus, regulates how much nonverbal behavior is finally exhibited by the virtual character. These filters might, although not implemented in the toolkit, "reflect the personality, affective state and energy level of characters" (Cassell et al., 2001, p. 6). This kind of individualization, however, only considers the *composition* of several nonverbal behaviors and disregards the matter of individual style on other levels of the generation process. Insofar as gesture generation is concerned, the BEAT system is able to produce beat gestures, discourse gestures marking contrast, and iconic gestures for actions and object features. The production of iconic gestures in the BEAT system is lexicon-based. That is, depending on annotated features, predefined gesture specifications are selected from a knowledge base.

The last module in the pipeline is *Behavior Scheduling*. It converts its XML input into a set of instructions to be executed by an animation system. Nonverbal behavior of any kind is produced in synchrony with the tagged text span they with which they are affiliated.

The BEAT system was designed to be extensible in several ways. First, new hand gestures can be added to the knowledge base in order to correspond to domain object features and actions. Second, the definition of new behavior suggestion generators allows expansion of the range of nonverbal behaviors as well as the strategies for generating them. Similarly, behavior suggestion filters can be added or modified, e.g., tailored to the behavior of a particular character in a particular situation, or to a particular animator's style.

Incidentally, a similar approach was taken with the **Nonverbal Behavior Generator (NVBG)**, proposed by Lee and Marsella (2006). The system analyzes the syntactic and semantic structure of surface texts and takes the affective state of the virtual agent into account to generate appropriate nonverbal behaviors. Based on a study from the literature and a video analysis of emotional dialogues, the authors developed a list of nonverbal behavior generation rules. Each rule includes associated nonverbal behaviors together with a set of words usually spoken when the given nonverbal behavior is exhibited. The rule set applied to input texts accounts for the linguistic and discourse contextual use of nonverbal behavior. Linguistic context is, for instance, considered by the 'first noun phrase rule', whereas the 'response request rule' is an example for the consideration of discourse context. Although the focus of the NVBG is on head movements, the approach also provides the possibility to associate pre-animated gesture clips with particular verbal expressions. One example of this is the 'negation rule' associating words like 'no', 'not', and 'nothing' with an animation clip in which the virtual agent puts his hand up and shakes his head.

The **Real Estate Agent (REA)** is a more elaborate system as it aims to model the *bi*-directional process of communication (Cassell et al., 2000a; Cassell, 2000). That is, in addition to the generation of nonverbal behaviors, the system also seeks to understand aspects of these same modalities' use by a human interlocutor.

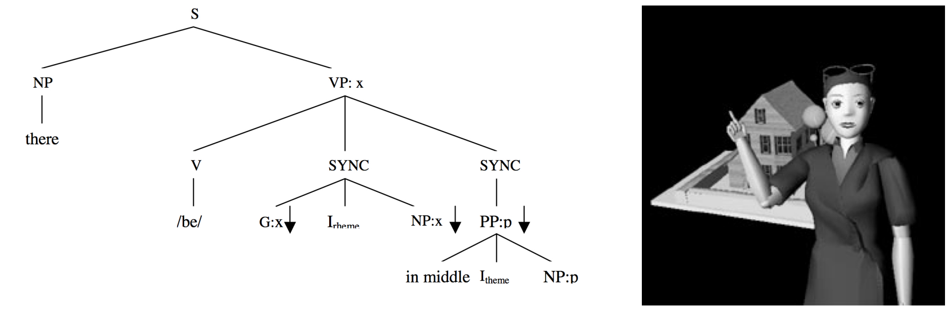
The REA architecture deals with input and output from multiple devices such as speech, gestures, and gaze. Of particular interest in the context of this thesis is the gesture generation process centered around the SPUD (Sentence Planning Using Descriptions) natural language generator (Stone et al., 2003). SPUD takes three types of input: (1) a set of communicative goals, (2) a grammar consisting of LTAG trees which are associated with a lexical anchor as well as semantic and pragmatic information, and (3) a knowledge base of logical formulae consisting of facts about the domain, explicitly labeled with information about their conversational status, i.e., whether the fact is private or shared. Provided with this input, SPUD treats sentence planning as a search problem in the search space spanned by the grammar and the knowledge base. Starting from an LTAG initial tree consisting of a root node and the set of communicative goals, the algorithm adds linguistic structures to the tree in each step, until a complete utterance that meets the given communicative goals is found.

For the REA system, SPUD's linguistic resources were extended to include a set of predefined gestures, upon which it draws to express its communicative goals. That is, multimodal utterances are generated with a single, uniform algorithm. To this end, Cassell et al. (2000a) introduced the *SYNC construction* to combine a whole gesture with the syntactic structure of a spoken constituent to be realized in temporal synchrony. For an example of a SPUD-generated utterance tree including SYNC-nodes for the integration of gestures, see Figure 3.2.1: gesture G co-occurs with the words 'a staircase', whereby the spiral shape of the staircase is mapped to the trajectory of gesture G.

The focus of gesture generation in the REA system is the context-dependent coordination of (lexicalized) gestures with speech, accounting for the fact that gestures do not always carry the same meaning as speech. Reflecting this, Cassell et al. (2000a) distinguished between *complementary* gestures that carry information not present in the simultaneous speech, and *redundant* gestures that convey the same information as speech. In a study of human-human conversation, rules of information distribution across modalities were extracted. These indicate the appropriateness of expressing semantic features in the gesture modalities, communicative goals, linguistic features in terms of theme/rheme, and other pragmatic information such as surprising semantic features and contrast. Integrating them into the generation framework resulted in the production of appropriate complementary or redundant gestures.

An extension of the REA system called REA*3D* was provided by Gao (2002). In this account, iconic gestures are directly derived from a 3D graphics scene augmented with information about 3D locations of objects and their basic shapes in terms of boxes, cylinders, spheres, user-defined polygons, or composites of these. These were directly mapped onto a set of hand shapes and spatial hand configurations in a rule-based way. This method allows for the derivation of a range of new gesture forms, but does not provide a unified way of representing and processing the knowledge underlying coordinated language and gesture use.

The **Virtual Guide (VG)** is a multimodal dialogue system represented by an embodied conversational agent that can help users to navigate a virtual environment while adapting its affective linguistic style to that of a human user in terms of polite-



(a) Syntactic structure of the multimodal example utterance in which a lexicalized gestures is integrated via a SYNC node.

(b) The virtual agent REA saying "a staircase in the middle of it." accompanied by a gesture which depicts the spiral shape of the staircase.

Figure 3.4: An example utterance tree generated with the REA system and its realization. Reprinted from Yan (2000).

ness (Hofs et al., 2010). The input for the route description consists of the shortest path from the starting point to the destination, specified as a list of 3D coordinates. This list gets translated into a sequence of markers associated with turn directions and landmarks. From this input, a route description is generated using a collection of sentence templates tagged for politeness. The templates are then organized in a specialization hierarchy, where specialized templates can augment or override the more general ones.

To generate appropriate gestures to accompany the verbal route description, the words in the route description are associated with tags representing different types of relevant gestures which may potentially be generated. An animation planner is at this point concerned with generating the required animations in synchronization with text-to-speech output. To this end, gesture selection follows a 'suggest-and-reduce' approach somewhat similar to that of the BEAT system (Cassell et al., 2001). First, the system creates a collection of all possible gestures that could be used to accompany the landmark references and direction words in each sentence of the route description. Subsequently, once the full route description has been generated, a selection from all possible gestures is made, based on weighted randomization. The weights used by the Virtual Guide were determined by hand, however the authors stated that more realistic weights might be determined empirically based on the results of video analysis. The type of gestures considered by the system are deictic and iconic. Deictics are generated from an objective viewpoint using the location of the target object as input parameter, while iconic gestures are generated using canned animations. Currently, the system always selects pointing gestures. This is the result of drawing from the findings of an evaluation study in which 68% of the participants preferred directions with objective viewpoint gestures.

Feature-Based Gesture Generation

The **NUMACK** system tried to overcome the limitations of lexicon-based gesture generation by formulating the meaning-form mapping at the level of single gesture features (Kopp et al., 2004). This approach was based upon the notion of *Imagistic Description Features* (IDFs) as described in Section 2.2.2. IDFs are used by NUMACK as an intermediate level of meaning, which explicates the imagistic content of iconic gestures, consisting of separable, qualitative features describing the meaningful geometric and spatial properties of entities.

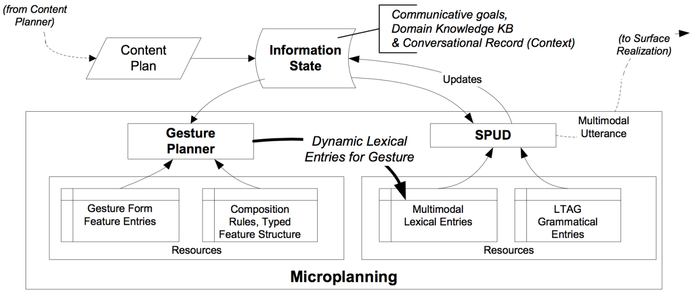

(from Content Planner)
Content Plan
Information State
Communicative goals, Domain Knowledge KB & Conversational Record (Context)
(to Surface Realization)
Updates
Gesture Planner
Dynamic Lexical Entries for Gesture
SPUD
Multimodal Utterance
Gesture Form Feature Entries
Composition Rules, Typed Feature Structure
Resources
Multimodal Lexical Entries
LTAG Grammatical Entries
Resources
Microplanning

Figure 3.5: The architecture of the NUMACK system in which the SPUD framework and a seperate Gesture Planner are employed. Reprinted from Kopp et al. (2004).

NUMACK extended the REA system described above such that a new subsystem for gesture planning within the microplanning stage was introduced, as illustrated in Figure 3.5. Gesture planning is based on an input specification of domain knowledge and a set of form feature entries connecting IDFs to morphological gesture features. The implementation of form feature entries is based on empirical evidence regarding the form-meaning mapping as described in Section 2.2.2. The gesture planner takes a set of IDFs as input and searches for all combinations of form feature entries which can be potentially realized. To this end, a feature structure unification is employed to combine morphological features, whereby any two form features may combine provided that the derived feature structure contains only one of any feature type at a time. This operation is applied iteratively until the given communicative intention gets encoded. Figure 3.6 shows the mapping of IDFs onto gesture features via form feature entries.

The gesture planner returns all possible combinations of gesture features, which are added to the resources of SPUD. Drawing from a set of dedicated SYNC constructions outlining all possible combinations of speech and gestures, SPUD chooses the

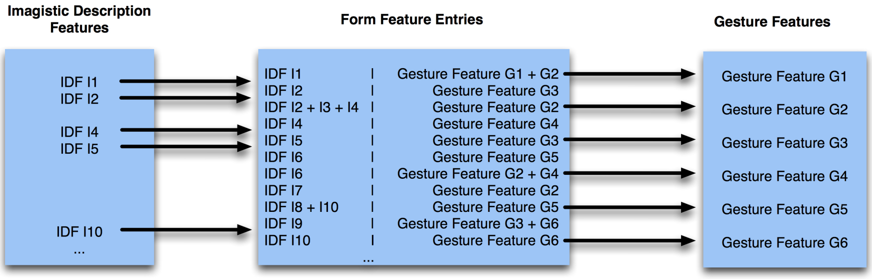
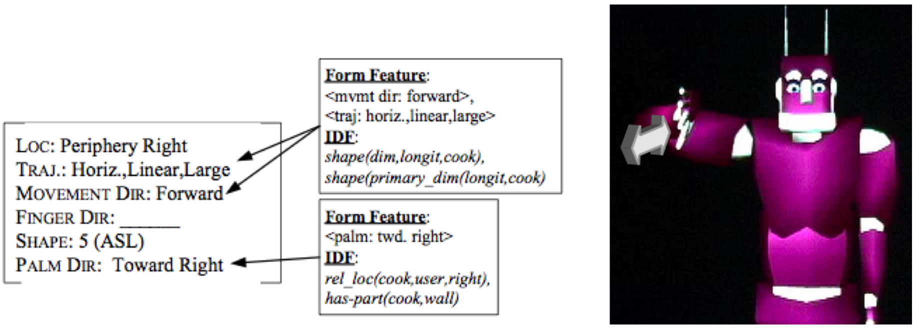

Imagistic Description Features
Form Feature Entries
Gesture Features
IDF I1
IDF I2
IDF I4
IDF I5
IDF I10
...
IDF I1 | Gesture Feature G1 + G2
IDF I2 | Gesture Feature G3
IDF I2 + I3 + I4 | Gesture Feature G2
IDF I4 | Gesture Feature G4
IDF I5 | Gesture Feature G3
IDF I6 | Gesture Feature G5
IDF I6 | Gesture Feature G2 + G4
IDF I7 | Gesture Feature G2
IDF I8 + I10 | Gesture Feature G5
IDF I9 | Gesture Feature G3 + G6
IDF I10 | Gesture Feature G6
...
Gesture Feature G1
Gesture Feature G2
Gesture Feature G3
Gesture Feature G4
Gesture Feature G5
Gesture Feature G6

Figure 3.6: Feature based form-meaning mapping in the NUMACK system: Imagistic description features are mapped on gesture features via form feature entries. Any two form features may be combined provided that the derived feature structure contains only one of any gesture feature type at a time.

gesture feature structure and the construction that, when combined with appropriate words, allows for the most complete intended interpretation in context. Finally, the resulting tree of the multimodal utterance is converted into an XML description, containing the textually defined words along with the feature structures for gestures to be performed by a virtual agent (Figure 3.7).



(a) A gesture feature structure is filled with form features entries.

(b) “You will see Cook Hall on your right.” accompanied by a gesture which depicts location and shape of the referent Cook Hall.

Figure 3.7: An example gesture generated with the NUMACK system and its realization. Reprinted from Kopp et al. (2004).

3.2.2 Customized Gesture Generation

The research reviewed so far was devoted to building general models of gesture use, i.e., systematic inter-personal patterns of gesture use are incorporated exclusively. What has not yet been considered in these systems is individual or group-specific variation, another line of research which has recently been making headway. In the following, approaches will be reviewed that provide means to customize gesture behavior.

Modulating Expressivity Parameters Based on the observation that gesture use is subject to considerable inter-individual differences, Hartmann et al. (2006) focused on the way individuals differ in the manner and execution of their gestures. Based on perceptual studies, six expressivity parameters of gesture quality were extracted as an intermediate level of behavior parametrization between holistic, qualitative communicative functions such as mood, personality, and emotion on the one hand, and low-level animation parameters like joint angles on the other hand:

- **Overall Activation** General amount of activity, e.g., passive vs. static or animated vs. engaged.
- **Spatial Extent** Amplitude of movements.
- **Temporal Extent** Duration of movements.
- **Fluidity** Smoothness and continuity of overall movement, e.g., smooth versus jerky.
- **Power** Dynamic properties of the movement, e.g., weak versus strong.
- **Repetition** Tendency to rhythmic repeats of specific movements.

These parameters were applied to the Gesture Engine of the virtual agent GRETA, whose library of known prototype gestures are tagged for communicative function (Hartmann et al., 2002). The parameterization approach allows for a large range of pre-defined gesture variants. See Figure 3.8 for an example of variation of the parameter ‘spatial extent’.

The expressivity parameter approach is driven by a perceptual standpoint, i.e., the question of how expressivity is perceived by humans. To investigate how the six parameters were recognized by human users, Hartmann et al. (2006) employed an evaluation study. It turned out that the recognition of single parameters was best for the dimensions ‘spatial extent’ and ‘temporal extent’, whereas the dimensions ‘repetition’ and ‘overall activation‘ were much less recognizeable.

Recently, Mancini and Pelachaud (2010) implemented the concept of expressivity parameters to create distinctive behavior patterns for ECAs. Their proposed algorithm generates nonverbal behavior for a given communicative intention and emotional

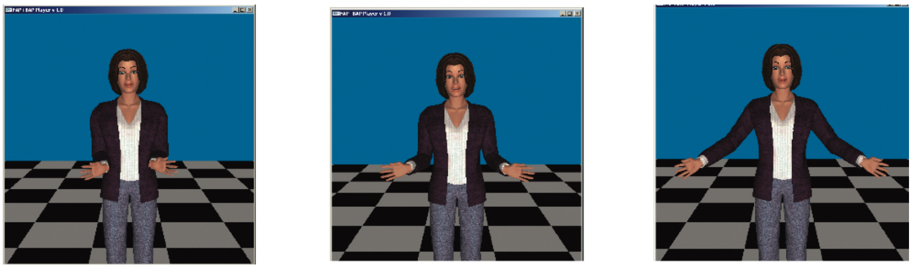


Figure 3.8: Variation of the parameter ‘spatial extent’ in the virtual agent GRETA: The neutral key pose in the middle is contracted in the left and extended in the right gesture execution. Reprinted from Hartmann et al. (2006).

state, driven by the agent’s general behavior tendency (‘Baseline’) and modulated by dynamic factors such as emotional states, relation with interlocutor, physical constraints, social roles. etc. For each modality (face, gesture, torso, head), the Baseline structure includes a modality preference and specifies values for the six expressivity parameters. The Baseline of a person can be automatically extracted from a video analysis, and then modulated by the agent’s ‘Dynamicline’, which influences the its behavior at two levels: the selection of multimodal signals to display and the specification of the behavior execution quality. In an evaluation study, human subjects were able to discern the achieved distinctiveness in the agent’s behavior.

Rehm et al. (2008) presented another variant which made use of the gestural expressivity parameters to generate culture-specific gestures (CSG for culture-specific generation). Their approach to CSG in embodied agents relies on a multimodal corpus analysis of human interactions in two cultures. The analysis of corpus data focuses on gestural expressivity in terms of the above-mentioned expressivity parameters depending on a speaker’s cultural background. Differences between the cultures can be identified and integrated in a probabilistic model for generating agent behaviors. In a Bayesian network, the culture to be simulated is connected with five dimensions of culture as a middle layer: hierarchy, identity, gender, uncertainty, and orientation. These culture-specific dimensions were then connected with gesture parameters.

This model was applied in an application called ‘cultural mirror’. As a user’s gestural expressivity (measured with a WiiMote) was analyzed and the classification result was set as evidence for the output nodes of the Bayesian network. By diagnostic inference, the user’s cultural background was estimated and this information was then set as evidence to the input nodes of a second network in which an agent’s behavior was parameterized via causal inferences, resulting in behavior congruent to user input.

Providing virtual agents with gesture style In another approach, Ruttkay (2007) aimed at endowing virtual humans with a unique style in order to appear typical of some social or ethnic group. The focus of this work was a markup language to define different aspects of style which were handcrafted to model the behaviour of stereotypic groups rather than individual group members. The markup language GESTYLE allows the generation of speech and accompanying gestures by tagging a text for meaning and declaring a style in which an utterance will be performed. To this end, GESTYLE contained the following set of constructs:

- **Character Markup** Defines the static characteristics of a character, such as culture, personality, age, sex, and other individual characteristics like handedness.
- **Situation Markup** Specifies the situation by setting dynamical aspects of the speaker (like mood, physical state) and the environment (social relation between interlocutors, characteristics of the addressee etc.)
- **Communicative Markup** The text to be uttered by the agent is tagged, e.g., for emphasis, get-attention.
- **Gesture Markup** Specifies the gesturing behavior to be expressed at certain points in time. Specific parameters are provided to modify the characteristics of the motions, e.g., amplitude, motion-manner.[2]

A set of gesture dictionaries was compiled for particular cultures, social, ethnic groups, and personalities or even for individual subjects. In these dictionaries, one or more gestures are stored to express a particular meaning to offer the alternatives of expressing a communicative function. Each gesture entry is augmented by the probability of using the specific gesture in this role and optional gesture modifying parameters specifying the motion characteristics of the gesture.

The effect of high-level character and situation parameters on the motion characteristics is given in terms of low-level gesture parameters. That is, a single gesture is selected from the possible alternatives, prescribed by different gesture dictionaries according to the given character and situation specification. Provided with an interface which translates the GESTYLE representation to the control parameters of an animated player, it is possible to give expression to this specification in the overt behavior of a virtual agent.

3.2.3 Data-Driven Gesture Generation

Another line of research uses data-driven methods to simulate (individual) speakers' gesturing behavior.

[2] The term 'gesture' does not only include hand and arm movements, but also head nods, eyebrow movements etc.

Creating Characters from Human Motion Capture Data Stone et al. (2004) proposed a method for using a database of recorded speech and captured motion to create an animated conversational character (RS+CM approach). The framework tied together offline activities of content authoring and data preparation with online processes that use the prepared content and data for generation and animation.

In *content authoring*, a scriptwriter designs what the character will say. Automatic tools then compute the utterance units implicit in the specification, formulate a concise script for a performer, compute a database specification that organizes the anticipated sound and motion recordings, and compile an application-specific generator that will index the resulting database.

For *data preparation*, automatic tools for speech and motion data analysis are combined with manual annotations. In particular, the data is coded for points of perceived prominence in speech and gesture. In addition, gestures are classified into two categories: descriptive or expressive. Descriptive gestures elaborate the referential content of the utterance, which is typical of iconic gestures that represent objects or events in space. Expressive gestures, in contrast, highlight the attitude of the speaker towards which she is saying and comment on the relationship of speaker and addressee, which is typically the case for metaphorical and beat gestures.

To plan a new utterance, the generator automatically determines the content and communicative function of each of the phrases the character needs to realize. To animate these phrases, suitable sound recordings must be combined with adequate gesture performances. As this is the unit selection problem, a cost function is defined to determine the best combination of a sound s and a motion m that minimizes the degree to which a unit of performance must be modified in the final realization. The function takes two measures into account. First, the difference in two successive motions, temporally averaged across the short overlay window where corresponding samples are interpolated, and second, the differences in pitch (a ratio) between the peak of two successive sounds.

The unifed approach makes it possible to capture the data needed for a character with a limited number of performances, to catalogue performance data with limited human effort, and to synthesize novel utterances.

Generating Gestures from a Speaker-Specific Model Another data-driven gesture generation approach was developed by Neff et al. (2008): a system for generating believable gesture animations for novel text which reflect the gesturing style of particular individuals by building speaker models (SM approach). The approach is mainly data-driven, using a video corpus of the human performer, but also incorporates general, character-independent mechanisms. The approach is divided into two phases: an *offline* processing phase to build gesture profiles from speakers, and an *online* processing phase to generate speaker-specific gestures for arbitrary input texts. For an overview see Figure 3.9.

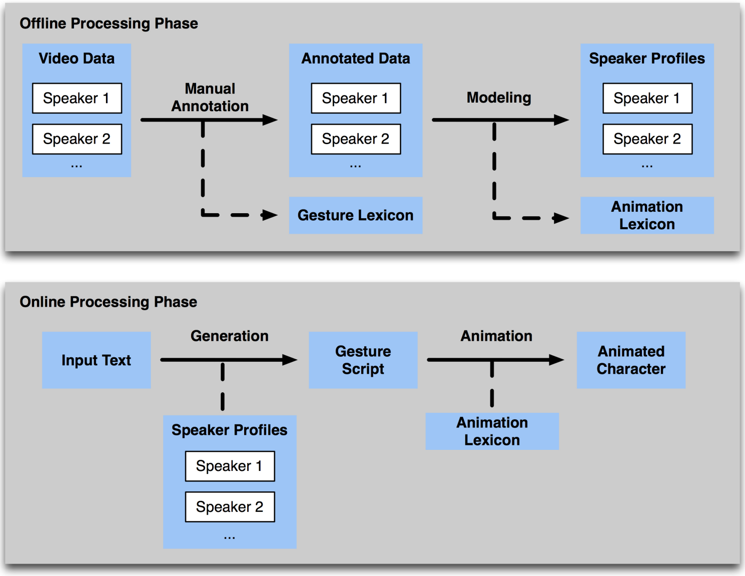


Figure 3.9: System to generate believable gesture animations for novel text that reflect the gesturing style of particular individuals (Neff et al., 2008). The approach is divided into two phases: an *offline processing phase* to build a gesture lexicon and gesture profiles from speakers, and an *online processing phase* to generate speaker-specific gestures for arbitrary input texts.

The offline processing phase, done individually for each speaker, and begins with the annotation of video data from the particular speaker with respect to speech and gestures. Spoken words are grouped into clauses and annotated for their information structure. The gestural part of the annotation follows the hierarchical organization of gestures in phases, phrases, and units (cf. Section 2.1.2). In addition, four attributes are coded for each gesture: (1) handedness, (2) lexical affiliate, (3) co-occurrence, and (4) lexeme. The lexeme denotes the lexicon entry to which the gesture corresponds in a *gesture lexicon* built beforehand to capture the semantics of gestures (Kipp, 2004). A total of 39 gesture types, i.e., recurring gesture patterns, are identified and described with respect to gesture form constraints in terms of handshape, hand location, hand orientation, hand/arm movement, handedness, shoulder movement, and facial expression.

From the annotated corpus, a profile of a speaker's gesturing behavior is built. This profile consists of a sample database, a statistical model, and average values. For the database, the annotations for each gesture in the corpus are stored as a reproducible sample of the specifc speaker. To build the statistical model, the speech transcription is processed in order to assign *semantic tags* such as 'agreement' ("yes") or 'quest_part' ("why"). The model is then automatically computed from the annotations and used in generation to trigger gestures, to predict where they are placed relative to speech, and to determine parameters such as handedness and frequency.

Once a speaker's gesture profile is created, the system can process any text that has been segmented into utterances and (manually) coded for its information structure. Words are stemmed and mapped to a semantic tag, just as in the offline modeling step. The generation then proceeds in two steps: gesture creation/selection, and gesture formation.

In the first step, a large number of underspecified gesture candidates are created and then reduced by a selection criterion. For each semantic tag in the input text, the generation system computes the conditional probability that a gesture g occurs with given semantic tag s. Additionally, a bi-gram model of gesture sequence is considered, i.e., the conditional probabilities that gesture g_i follows a previous gesture g_{i1}. Similarly, the handedness is determined on the basis of a bi-gram model that captures handedness sequences. Handshape is determined by consulting a lexicon in which all suitable handshapes for a particular lexeme are specified. Following the rule of economy, a handshape is chosen if it is equal to the handshape employed in the previous gesture. Otherwise the handshape is changed to a suitable one.

In the second online step, timing details for the realization of gestures are planned. To this end, gesture and speech are arranged by positioning the end of the stroke at the end of the corresponding word, using a random onset based on the speaker's mean value. Neighboring gestures are merged resulting in multiple strokes synchronized to enforce a minimum time span.

3.3 Summary and Discussion

This chapter aimed to cover the computational perspective on gesture generation. The first part dealt with the question of content representation underlying gesture production and addressed the state of the art in computational imagery. The second part presented related work in terms of gesture generation approaches, structured by model-based approaches, customization attempts, and data-based methods. Insights from both issues will be summarized and discussed in the following.

Computational Imagery

The presentation of content representation approaches focused particularly on the IDT (Imagistic Description Tree) model. This was developed based on empirical data, to represent shape-related information in a gesture understanding system. It is, thus, designed to cover all meaningful visuo-spatial features one finds represented in shape-depicting iconic gestures. IDTs are characterized by (1) providing a tree structure for object decomposition, (2) defining extents in different dimensions as approximations of shape, and (3) allowing the possibility of underspecified dimensional information.

The IDT model is particularly suited for gesture-related representation of imagistic knowledge for the following three reasons. First, the IDT model permits the represention of underspecification and vagueness, both of which are immanent in gesture. Dimensional underspecification (e.g., when representing a 2D circle or simply a 1D width) is given when the axes of an object schema cover less than all three dimensions of space. Vagueness can hold with respect to the extent along a certain dimension (e.g., when representing something 'longish') or the decomposition of a shape into subparts (e.g., when representing a church without being able to recall all its individual parts or geometrical details).

Second, complex and structured objects are typically not depicted all at once with a single, complex gesture. Instead, objects are usually decomposed and the components are described with simple sequential dimensional gestures. That is, successive gestures tend to organize cohesively in space and reflect the spatial arrangement of the referents they depict. A tree-like representation format like that of the IDTs, therefore, is particularly suited to reflect this.

Third, shape properties which are often depicted by gestures can be defined semantically in terms of the extent and profile of intrinsic object axes. Dimensional features as 'fat' or 'longish' are, for instance, well expressible as relations between object extents. The IDT model covers these object features adequately.

Gesture Generation

In Table 3.1, an overview of the computational models is given with respect to the six crucial aspects of gesture production (as identified in Section 2.2), which already have been considered in the evaluation of theoretical models (Section 2.3). Although this set of factors is not necessarily complete, a comprehensive computational model of gesture production should consider this set, at least.

Concerning the nature of the representations underlying gesture production and its mapping onto gesture form, lexicon-based systems rely on a dictionary of pre-defined gesture forms associated with a particular meaning (BEAT, NVBG, REA, VG). In combination with semantically tagged input texts, meaning is straightforwardly tied to the physical appearance of a gesture. Along the same line, technical approaches that aim to customize gesture use rely on pre-defined gesture forms to be modified to account for individual speakers (GRETA) or groups of speakers (GESTYLE, CSG). Similarly, data-based systems also make use of dictionaries of pre-defined gesture forms in combination with semantic information; the RS+CM system uses segments of motion capture data and the SM system employs a set of gesture types.

Table 3.1: Overview of computational gesture generation approaches and their major characteristics: Behavior Expression Animation Toolkit (BEAT, Cassell et al., 2001), Nonverbal Behavior Generator (NVBG, Lee and Marsella, 2006), Real Estate Agent (REA, Cassell et al., 2000a), REA*3d* (Gao, 2002), Virtual Guide (Hofs et al., 2010), NUMACK (Kopp et al., 2004), GRETA (Mancini and Pelachaud, 2010), GESTYLE (Ruttkay, 2007), Culture-Specific Generation (CSG, Rehm et al., 2008), Recorded Speech and Captured Motion(RS+CM, Stone et al., 2004), Speaker Models (SM, Neff et al., 2008).

	BEAT	NVBG	REA/REA*3d*	VG	NUMACK	GRETA	GESTYLE	CSG	RS+CM	SM
Representation	-	-	-/FB	-	FB	-	-	-	-	-
Form-meaning mapping	-	-	-/RB	-	RB	-	-	-	-	-
Discourse context	RB	RB	RB/RB	-	-	-	-	-	-	DB
Linguistic context	-	RB	RB/RB	-	-	-	-	-	DB	-
Customization	-[2]	-[2]	-[2]/-[2]	(DB)[1]	-[2]	P	P	P	(DB)[1]	DB
Representation techniques	-	-	-/-	-	-	-	-	-	-	-

[FB] Feature-based

[RB] Rule-based

[DB] Data-based

[P] Parameterization

[1] not implemented, yet

[2] not explicitly covered

The NUMACK system stands out in that it tried to overcome the limitations of lexicon-based gesture generation by implementing an intermediate level of meaning which explicates the imagistic content of iconic gestures, consisting of separable, qualitative features describing the meaningful geometric and spatial properties of entities. A similar strategy was applied in REA*3d*, however, the 3D graphics representation tagged with referent features was directly mapped onto gesture features.

None of the technical systems—explicitly—made use of the fact that iconic gestures realize different gestural representation techniques. Since most of the systems under consideration rely on pre-defined gesture sets (in which different representational techniques may implicitly be present), there are only two systems in which this distinction would be reasonable, REA*3d* and NUMACK. Actually, the limited prediction accuracy, as shown for the NUMACK system (Kopp et al., 2004), could be due to the fact that the correspondence between combinations of morphological features and the visual or geometrical similarity of referents was investigated without considering different techniques of representation.

Technical systems differ with regard to the consideration of linguistic and/or discourse context. The BEAT framework considered contextual information such as information structure (theme, rheme) or novelty in its behavior generators. The REA/REA*3d* system relied on discourse functional information, but additionally focused on the relation between gesture use and speech. In particular, it implemented rules of information distribution to account for the fact that speech and gestures are sometimes redundant and other times complementary. The NUMACK account followed the same strategy as the REA system by using SPUD to compose full, multimodal utterances. Extended with a flexible gesture planner (instead of using a static set of predefined gestures), gestures were then dynamically incorporated into SPUD's resources and utilized in the same way as described in REA.

The probabilistic SM system made probabilistic generation choices on the basis of conditional probabilities taking two aspects into account: the previously performed gesture and the input text tagged with theme, rheme, and focus. RS+CM relied on aspects of the linguistic context as it searched for the best combination of sound and motion appropriate for given the content and communicative function to be realized. A cost function determined which combination minimizes the degree to which a unit of performance must be modified in the final realization.

In VG, sentence templates were associated with tags representing different types of gestures to be possibly generated. That is, any kind of contextual influence was not explicitly modeled, but at most implicitly defined by these templates. In GRETA, CSG, and GESTYLE no kind of contextual information was considered.

Some of the systems under consideration tried to achieve generality by abstracting from individual speakers: BEAT, REA/REA*3d*, VG, NUMACK, and RS+CM[3]. Others, by contrast, focused on inter-individual or group-specific differences in gesture use. In GRETA, the concept of *expressivity parameters* was implemented to create distinctive behavior for virtual agents. The system was able to account for automatically extracted individual speaker characteristics in terms of expressivity parameters and modality preference. This kind of individualization of gesture style was evaluated as successful in that the parameters could be perceived by human subjects. The same concept of expressivity parameters also came to application in the CSG system to provide virtual agents with culture-specific gesture style. Similarly, in GESTYLE, different styles were defined in a dictionary of meaning-to-gesture mappings with optional modifying parameters to specify the characteristics of a gesture in terms of group and culture specificity. In both systems, however, the individualization of gesturing behavior was limited to parameters of gesture style. What was not considered in this individualization account was the formation of meaningful gesture forms and how these also reflect inter-individual differences.

In conclusion, previous research in computational gesture generation has either emphasized common patterns in the formation of iconic gestures or concentrated on the individualization of gesture use. Common patterns have throughout been investigated in a *rule-based* way, whereas individualization was realized either in a *data-based* fashion (probabilistically or using a cost function) or by parameterization of pre-defined gestures. That is, so far, there is no computational system which combines a substantiated mapping of meaning onto gesture form under contextual constraints which additionally accounts for inter-individual differences in gesture use.

This can be explained in the following way. On the one hand, a comprehensive data-based account necessitates a rich database consisting of fine-grained gesture coding in combination with contextual information about gesture use. To date, no gesture corpora are not available in a size sufficient to adequately apply data-based techniques. In the absence of such corpora, rule-based modeling is an alternative. Inter-individual differences can, however, hardly be investigated in a rule-based way. As a result, it has not been possible previously, either with data-based or rule-based methods, to adequately model gesture production comprehensively. This situation is the starting point for the modeling approach to be developed in this thesis.

[3] Although not explicitly designated by the authors, one could imagine to also apply the basic principles of these systems for customized gesture generation

CHAPTER 4

Empirical Study

The introductory Section 1.3 laid out the cyclic design methodology of modeling communicative behavior. The developmental cycle begins with the collection of natural human communication data that reflects the surface level of the behavior to be modeled. The physical appearance of a gesture does not depend on iconicity alone, but also on specific contextual constraints such as discourse context, the linguistic context, gesture history, and inter-individual differences in gesturing (Section 2.2). So far, no literature describes in detail how iconic gesture use is correlated with those variables. As demonstrated in Section 2.2, iconicity is not only decisive for the physical appearance of a gesture, but also specific contextual constraints such as discourse context or verbal context in which a gesture is used. In addition, intersubject differences in gesturing are quite obvious, but exactly how iconic gesture production is subjectively biased remains poorly understood so far. The sensitivity of gesture use to these factors is, however, decisive for the design of a generation model.

An adequate empirical basis for a detailed account of gesture use must, therefore, fulfill several requirements. First, the physical form of gestures has to be annotated in a fine-grained way. Second, gesture referents and their visuo-spatial properties have to be clearly defined to provide the basis for mapping meaning into gesture form. Third, the data must be supplemented with information about the gesture context. Finally, the corpus should consist of a not too small number of speakers to capture inter-individual differences in the data. Because there is, to date, no such collection of data available, an empirical study on spatial communication in a setting resembling a potential application scenario was conducted in the scope of the CRC 673 project B1 at Bielefeld University in February 2007. This chapter recounts the study, results, and conclusions with respect to a computational model of iconic gesture production. The *Bielefeld Speech and Gesture Alignment* (SaGA) corpus, including details of the study setting, annotation of the data, and reliability issues was described in Lücking et al. (2010). Initial findings were published in Bergmann and Kopp (2009b, 2010b).

The remainder of this chapter is structured as follows. In Section 4.1.1, the study setting will be described. In Section 4.1, the SaGA corpus will be introduced in terms of experimental data as well as the secondary annotation data. Corpus evaluation in terms of inter-rater reliability will be presented. Finally, results from the statistical analyses of the data will be described in Section 4.2.

4.1 Building a Corpus of Speech and Gesture Use

4.1.1 Corpus Study

Participants

50 participants (21 female, 29 male) took part in the study. Half of them participated as direction givers, the other half as direction followers, resulting in 25 dyads. Pairs of participants did not know each other prior to the study. All were recruited at Bielefeld University and received four to ten euros for participating.

Procedure

The setting of the corpus study (SaGA study, henceforth) consisted of two phases. In the first, the *stimulus presentation phase*, the direction givers were provided with knowledge to be communicated to the direction follower in the second phase of the study, the *dialogue phase*. At the beginning of the stimulus presentation phase, subjects participating as direction givers were equipped with virtual reality glasses and motion tracking markers at neck, hands and elbows. They were seated in an immersive virtual reality environment (‘Cave Automatic Virtual Environment’, CAVE) and experienced a ‘bus ride’ through a model of a town presented in virtual reality.

Instructions were to memorize the route, as well as to carefully familiarize oneself with the appearance of five ‘sights’—visually distinctive landmarks—at which the bus ride temporarily paused (see Figure 4.1). Participants decided on their own when the tour was to continue, so they could spend time viewing each of the five sights for as long as they liked.

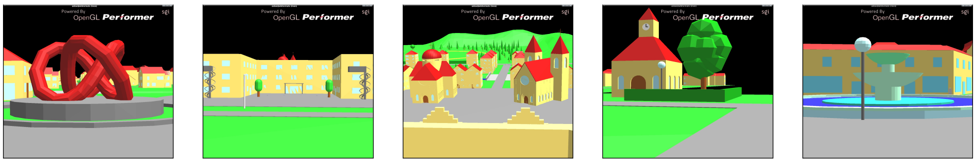


Figure 4.1: Virtual reality sights from the corpus study (from left to right): sculpture, townhall, churchsquare, chapel, and fountain.

Upon finishing the bus ride, the direction giver had to explain the route and describe the five sights to the direction follower, who would be tasked with navigating the same path through the virtual town afterwards. In doing so, The giver was asked to describe the five sights accurately enough for the follower to be able to notice possible discrepancies in appearance that might have been introduced afterwards.

4.1.2 Primary Data

The primary data of the SaGA corpus consists of 25 dialogs. Audio- and videotapes were taken of each dialog. For the videotape, three synchronized camera views were recorded (see Figure 4.2), as well as body movement data and eye-tracking data from the router. In total, the SaGA corpus consists of 280 minutes of audio- and video material.

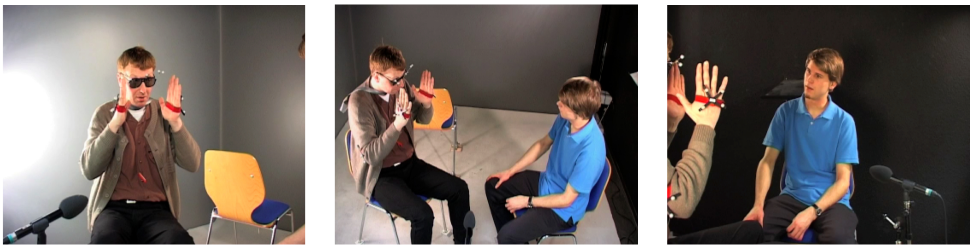


Figure 4.2: Dialogue phase of the corpus study: three camera views. Focus on the direction giver (left), focus on the direction follower (right), both participants (middle).

4.1.3 Secondary Data

The data was completely and systematically annotated based on an annotation grid developed according to theoretical considerations and refined in pilot annotation sessions. In the following, the basic categories and values will be presented. Detailed annotation guidelines can be found in Bergmann et al. (2007) and Bergmann et al. (2008).

The grid basically consists of three parts as summarized in Table 4.1. It comprises, first, a segmentation and classification of gesture including gesture representation techniques and morphological gesture features. Second, the spoken words are transcribed, tagged with part-of-speech information, parsed for their syntactical structure, and coded for their dialogue context. Third, a subpart of the corpus (only sight descriptions) has been coded for the gestures' referent objects and their spatio-geometrical properties (dimensionality, extents, symmetries, profiles, etc.). All multimodal corpus data are stored, retrieved and transformed within the Ariadne system (Menke and Mehler, 2010).

Table 4.1: Coding scheme for gestures, their referents and their discourse context.

	Variable	Annotation Primitives
Gesture	Gesture Phase	preparation, stroke, retraction, pre-stroke hold, post-stroke hold
	Representation Technique	indexing, placing, shaping, drawing, posturing, sizing, counting, hedging
	Handedness	LH, RH, 2H
	Handshape	ASL handshapes, see Table 4.2
	Palm Orientation	PAB, PTB, PTL, PTR, PUP, PDN + combinations and sequences
	BoH Orientation	BAB, BTB, BTL, BTR, BUP, BDN + combinations and sequences
	Movement Type	linear, curved + sequences
	Movement Direction	MF, MB, ML, MR, MU, MD
Verbal Context	NP types	Tags according to the STTS tagset (Table 4.3)
Discourse Context	Thematization	theme, rheme
	Information State	private, shared
	Communicative Goal	lmIntro, lmDescrProp, lmDescrConstr, lmDescrPos
Referent Features	Subparts	1 or more, none
	Symmetry	sym, none
	Main Axis	x-axis, y-axis, z-axis, none
	Position	3D vector (left, middle, right)
	Shape Property	longish, cubic, squared, arc-shaped, spherical, round, etc.

Gesture Segmentation

As a first step in gesture annotation, the stream of hand movements was segmented to identify single gesture occurrences. In general, a gesture is delimited by two consecutive *resting positions* (McNeill, 1992; Kita et al., 1998). Although this resting position might vary from speaker to speaker, it is typically a comfortable posture of hands and arms so that only a minimum of muscular strength is necessary. For instance, when the speaker is sitting, the hands often lay in the lap. A gesture's internal structure is then identified by segmenting the movement into the following *gesture phases* as described in Section 2.1.2:

- **Preparation** Hands and arms are moved from the resting position into the start position of the stroke phase.
- **Stroke** Strokes can be either *dynamic* or *static*. In the static case, the hands

are held in a particular configuration, whereas in the dynamic case they are moved. For static strokes, the starting point in time is characterized by halting the movement. For dynamic strokes, the starting point in time is typically characterized by a change in direction or speed. Physiologically, these strokes are characterized by a high tonicity.

- **Retraction** The hands are brought back into a resting position.
- **Pre-hold** Phase of static position *before* starting a dynamic stroke.
- **Post-hold** Phase of static position *after* starting a dynamic stroke.

Notably, the movements between two resting positions may contain multiple strokes. In these cases, the optional phase to prepare for the next stroke is coded as a preparation (not as a retraction).

Representation Techniques

Based on sub-classification approaches of iconic gestures from the literature, a set of consensus categories was identified (Section 2.2.3) and supplemented with further representation techniques according to the focus on object descriptions in the present data. This resulted in a set of eight representation techniques:

- **Indexing** Pointing gestures referring to positions in gesture space (referents were not physically present). Indexing gestures convey only information about the relative position of their referents.
- **Placing** Holding the hands as though they hold or grasp an object which is then is placed or set down within gesture space. In contrast to shaping gestures, in which the hands also virtually 'touch' the referent object, the stroke is static in placing gestures. That is, the referent object is touched, but not sculpted.
- **Shaping** Sculpting or contouring movement of the hands. The meaning consists of a shape that is formed by the hands. This shape is typically three-dimensional.
- **Drawing** Tracing the outline of an object's shape. Typically, one finger (mostly the index finger) is used as an imaginary pencil, so that the finger itself does not stand in for the referent, but the outline which is drawn in the air or on some kind of surface (e.g. the other hand or a table).
- **Posturing** Forming a static configuration to stand as a model for the object shape *itself.*
- **Sizing** Using to display sizes, distances or diameters of objects. This is done with both hands or with one hand and refers to one axis of the very object, i.e., the extent in one dimension.
- **Counting** Displaying with outstretched fingers a number between one and ten.
- **Hedging** Wiggling or shrugging movements depicting uncertainty.

Gesture Form

Gesture form coding followed the widely accepted four-part description of a gesture's physical appearance, as agreed upon in the majority of gesture coding schemes (Section 2.1.2).

Handedness For each gesture, whether it was performed with the right hand (***RH***), with the left hand (***LH***), or with both hands (***2H***) was coded.

Handshape Regarding the annotation of handshapes, there are two prevalent description methods, both of which originated in sign language description: the notation of hand configurations with labels of the American Sign Language (ASL), and the coding symbols from HamNoSys, the 'Hamburg notation system for sign language' (Prillwitz et al., 1989). The two description methods differ from each other in the way that ASL labels provide a comprehensive set of handshapes, whereas HamNoSys describes the hand configuration compositionally by 12 standard hand shapes (flat, fist, pointing index finger, etc.), and a set of modifications that can be applied to them, changing the bending of individual fingers or the thumb.

Concerning the question of which scheme to use, Kimbara (2008) conducted a study in which handshapes were used as an analytical target for coding. Kimbara concluded that, "ASL handshapes provided useful coding labels to broadly categorize the handshapes of speech-accompanying gestures based on their similarity in form." (Kimbara, 2008, p. 128f.). Given the fact that "nearly all gestures are deficient from an ASL point of view". McNeill (1992, p. 86) suggests to use the "ASL shape that the gesture mostly resembles". To systemize variations of the basic ASL handshapes, a set of modifiers was employed as in Kopp et al. (2007):

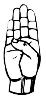
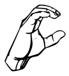
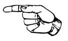
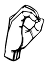
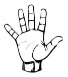
- ***bent***: if the fingers are not stretched
- ***loose***: a relatively relaxed realization of the very handshape
- ***spread***: if the fingers are not held close together
- ***small/large***: size modifiers for the handshape ASL-C

For a sample of handshape labels and modifiers used for the present data, see Table 4.2.

Hand Orientation Coding schemes basically agree on the option to code the hand orientation with two values: palm and finger (or back of hand) orientation, typically coded in terms of six speaker-centric, base- or half-axes (McNeill, 1992; Kopp et al., 2007)[1]

[1] Kendon (2004) proposed to annotate palm orientation more by arm position, rotational position of the forearm, and orientation of the metacarpal. This system, however, "seems not to allow the same accurateness in the description and expandability of categories" (Bressem, 2008, p. 15).

Table 4.2: ASL categories employed to annotate handshape in the SaGA corpus.

ASL label		Description	Modifiers
ASL-B		Fingers are held together and extended, forming a flat plane with the palm. The thumb is tucked in against the palm.	bent, loose, spread
ASL-C		The fingers are together and curved. The thumb is opposed and curved so the thumb and index finger resemble the letter C.	bent, large, loose, small
ASL-G		The index finger is bent at the base joint but otherwise extended, the thumb is rotated outward from the palm at the base and extended so that it is parallel with the index finger, and the other three fingers are curled into the palm,	bent, loose
ASL-O		The fingers are extended, together, and curled, and the tip of the thumb is touching the tip of the index and middle fingers to form a circle.	loose
ASL-5		All four fingers are extended and spread apart. The thumb is unopposed and extended.	bent, loose

The orientation of hands in the SaGA corpus was, therefore, defined in terms of two values: palm orientation and back of hand (BoH) orientation. **Palm orientation** was 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): palm up (***PUP***), palm down (***PDN***), palm to the left (***PTL***), palm to the right (***PTR***), palm away from the speaker's body (***PAB***), and palm towards the speaker's body (***PTB***). Up to three of these basic values were combined to encode diagonal or mixed directions, e.g., 'up/right' ('PUP/PTR') or 'up/right/forward' ('PUP/PTR/PAB'). This reults in a total of 120 basic palm orientation values. In order to capture dynamic palm orientations, it was possible to build a temporal sequence of these basic values by means of the '>'-operator. The value 'PUP>PDN', for instance, denotes an upwards-downwards movement sequence.

BoH orientation was coded along the same lines with the following six basic values: BoH up (***BUP***), BoH down (***BDN***), BoH to the left (***BTL***), BoH to the right (***BTR***), BoH away from the speaker's body (***BAB***), and BoH towards the speaker's body (***BTB***). The coded BoH orientation represents the direction of the fingers as if they were extended. Again, combinations of these basic values were used to code diagonal or mixed directions. Two examples for palm and BoH orientations are given in Figure 4.3.

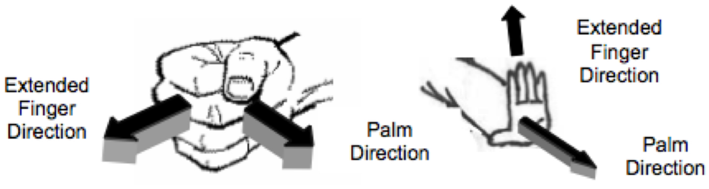


Figure 4.3: Hand Orientation defined in terms of palm and back of hand (BoH) orientation (reprinted from Kopp et al. (2007)). Assuming that left hand would be held straight out in front of the body, it would be coded as the palm facing left ('PTL') and the BoH facing away from the speaker's body ('BAB'). In the case of the right example, the coded palm direction would be 'PAB' (away from the body) and the BoH orientation would be up ('BUP').

Wrist Movement In most coding schemes dynamic aspects of gesturing are described using several attributes, whereby two values are quite prevalent: direction and trajectory. The movement direction is typically coded in terms of 'right', 'left', 'up', and 'down'. Trajectory coding aims to describe the shape of the motion. [2]

Along the same lines, in the SaGA corpus, the **movement direction** for dynamic gestures was annotated in terms of the six cardinal directions in space: movement upwards (***MU***), movement downwards (***MD***), movement to the left (***ML***), movement to the right (***MR***), movement forwards (***MF***), movement backwards (***MB***). As already described for palm and BoH orientation, combinations and sequences of the categories were used to describe directions in between the six basic values as well as temporal sequences.

To further classify the **type of movement** trajectory, two categories are distinguished: ***linear*** movements (movement in a straight line) and ***curved*** movements (movement along an arc or curve). Assume, for instance, the sequence of orientations 'MU>MR>MD>ML'. If it is performed linearly, the resulting trajectory is a square

[2] Some schemes provide further details about gestural movement. Müller (1998), e.g., focuses on the characterization of gestural motion patterns drawing attention to how gestural movements are modulated employing annotation values like 'stressed beginning' or 'middle phrase extended'. The CoGesT scheme (Gibbon et al., 2004) also contains descriptions of direction and trajectory, and additional modifiers for repetition, size and speed.

Table 4.3: Tags for noun phrase patterns according to the Stuttgart-Tübingen Tagset (STTS) for German.

(1)	ADJA NN	Adjective, Noun
(2)	ART ADJA	Article, Adjective
(3)	ART ADJA NN	Article, Adjective, Noun
(4)	ART	Article
(5)	ART NN	Article, Noun
(6)	CARD	Cardinal
(7)	CARD NN	Cardinal, Noun
(8)	NE	Proper Noun
(9)	NN	Noun
(10)	PDAT NN	Attributing Demonstrative Pronoun
(11)	PDS	Substituting Demonstrative Pronoun
(12)	PIAT NN	Attributing Indefinite Pronoun without Determiner
(13)	PIS	Substituting Indefinite Pronoun
(14)	PPER	Irreflexive Personal Pronoun
(15)	PRF	Reflexive Personal Pronoun

whereas it would be a circle if the same sequence would be performed in a curved fashion.

Linguistic and Discourse Context

Transcription The interlocutor's words were transcribed orthographically, i.e., they were put into the correct form of German spelling. Spoken language, however, is characterized by some facts that do not follow the rules of German orthography. First, there are cases of contractions of two words into one word. Some of these are even lemmatized, e.g., combinations of article and preposition as in 'zum' or 'am'. Second, there are interjections occurring in spoken language for which no orthographic rules exist, e.g., 'äh', 'mhm', 'tz', or 'boah'. These phenomena occur quite frequently in the SaGA corpus and, thus, are marked as interjection in the transcripts.

Syntax Tagging In the scope of the work reported here, we concentrate on noun phrases which were identified in 15 corpus transcripts by automatic Part-of-Speech-Tagging (POS) using the TreeTagger (Schmidt, 1994). A list of the most common noun phrase patterns found is given in Table 4.3.

Communicative Goals The transcription of the interlocutor's words is enriched with further information about the overall discourse context. For this purpose, the utterance is broken down into clauses, each of which representing a proposition. For each clause, we annotate its communicative goal. Denis (1997) developed several categories of communicative goals that can be distinguished in route directions due to the focus on object descriptions, which were revised and refined into four categories:

Landmark Introduction A landmark is mentioned without further exploration, e.g., 'there is a chapel'.

Landmark property description The properties of an object, in terms of shape, color, material etc., are described as in 'the town hall is U-shaped'.

Landmark construction description An object's construction is described, e.g., 'the church has two towers'.

Landmark Position Description The description localizes the object as in 'there is a tree in front of the building'.

Information Structure Clauses are further divided into two smaller units of thematization partitioning of the content of a sentence according to its relation to the discourse context. The structuring of utterances into a topic part and a comment part is a pervasive phenomenon in human language and there are numerous theoretical approaches describing thematization and its semantics (cf. Kruijff-Korbayova and Steedman (2003)). Following Halliday (1967) it is distinguished between thematization and information focus.

Thematization is coded in terms of ***theme*** (what the sentence is about) and ***rheme*** (what is being said about the theme), according to the 'before verb-complex' and 'after verb-complex' heuristics as described in Hiyakumoto et al. (1997). For example, in the utterance 'the church has two towers', the first noun phrase ('the church') is the theme and the second noun phrase is the rheme.

Focusing on noun phrases and their accompanying gestures, to which the annotation of information structure is restricted, information focus was annotated following Stone et al. (2003) in using the terms 'information state' and distinguishing straightforwardly between ***private*** (a discourse referent which lacks an antecedent in the previous discourse) and ***shared*** (a referent (or referent feature) already mentioned in the previous discourse) knowledge. For instance, in the utterance 'the church has a dome-shaped roof' the first noun phrase ('the church') is shared since the must haven been introduced into the discourse before (use of definite article). The second noun phrase ('a dome-shaped roof'), on the contrary, is private because the object (feature) is discourse-new.

Notably, as suggested in Ritz et al. (2008), thematization and information focus are annotated independently as different dimensions of information structure, assuming no prior relation between them. In particular, rhematic information is not always private, as, for instance, when content is repeated for better comprehension or in reply to interposed questions.

4.1.4 Referent Annotation

All gestures used in the object descriptions were further coded for their referent and some of their spatio-geometrical properties. These object features are drawn from an IDT representation of the VR stimulus of the study (cf. Section 3.1). For each gesture, a node ID of the IDT representation is coded from which the following properties can be inferred: (1) number of subparts (i.e., childnodes), (2) number of symmetrical axes, (3) main axis, (4) position, and (5) shape properties.

4.1.5 Coding Reliability

Annotation-based data might be problematic as they are based on subjective judgements of the coders. The reliability of the annotation is, therefore, of vital importance for the significance of results. It has to be shown that different people agree with respect to the coding judgements on which statistical analyses are based to make research results replicable (Carletta, 1996). To make a database as sound as possible, it is necessary to evaluate coding decisions with respect to their reliability. Two kinds of agreement are distinguished depending on whether one coder re-codes the same data (*intra-rater* agreement) or several coders annotate the same data (*inter-rater* agreement). The latter captures the stability of the annotation and is the more severe kind of reliability as it goes beyond intra-coder inconsistencies (Krippendorff, 1980).

Reliability Measures

Depending on the type and scale of the data, different statistical methods are available. A qualitative distinction has to be made between *Type I* and *Type II* ratings (Gwet, 2001). Type I measurements are those where the degree to which a rating is subject to human interpretation is well-understood and the outcome easily interpretable. As an example, Stegmann and Lücking (2005) cited the measuring of a patient's blood pressure by a doctor. The outcome displayed on the blood pressure gauge reflects the actual level of the patient's blood pressure (or at least approximates it in a sufficient way). Other doctors will come to the same result. Type II data, in contrast, are subject to interpretation by the coder. Here Stegmann and Lücking (2005) gave the example of a classification task in which, on the basis of data from a psychological questionnaire, raters have to determine the satisfaction level of various subjects assigning them to categories of emotion such as 'happy' or 'sad'. This difference between Type I data and Type II data has to be considered in evaluations of respective annotations as Type II ratings must be adjusted for chance-based agreements, whereas this is not necessary for Type I ratings. The SaGA corpus comprises both types of annotation. The classification of gestures in terms of representation techniques, reference objects and dialogue context information is interpretive and therefore of Type II. The respective annotation labels are categories on a nominal scale. Descriptions of gesture form make up data of Type I. With one exception (handshape, see below), the labels for annotating a gesture performance are ordered on an ordinal scale. Accordingly, different methods are employed to evaluate annotations of representation techniques and context information on the one hand, and annotations of gesture form on the other hand.

As a chance-corrected coefficient determining the level of agreement to be found in Type II data, the first order agreement coefficient AC_1 developed by Gwet (2001) was chosen since the widely used Kappa coefficient (Cohen, 1960) is often criticized on grounds of delivering anti-intuitive results under certain configurations (kappa paradoxes; for a discussion see Stegmann and Lücking (2005)).

Regarding the interpretation of agreement coefficients different quality thresholds exist. Values above which the agreement of raters is judged as acceptable range from 0.4 to more stringent conventions of 0.8 (cf. Artstein and Poesio, 2008). A frequently employed agreement level, also applied to the SaGA data, is 0.7 with an α-error of 0.05 and a β-error of 0.85 for Type II annotations.

In addition, to assess the extent of association between annotations of the Type I gesture morphology, an approach based on angle measures was employed for codings of directions and orientations. As the disagreement between, e.g., ‘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. This problem was addressed by translating the annotation labels into angular measures which can be analyzed in terms of numeric differences.

Movement directions, palm and BoH orientation were compared by calculating the angle between the two orientation vectors. For instance, there is an angle of 90° between ‘PTL’ and ‘PUP’, and an angle of 45° between ‘PTL’ and ‘PTL/PUP’. A maximal angle of 180° is present if the two vectors are opposing each other (e.g. ‘PTL’ and ‘PTR’) and can be considered as the worst match.

Reliability Results

Inter-rater agreement was calculated based on a sample of 477 gestures ($\sim$10% of the data) which have been classified independently by four annotators.

Type II data The first-order agreement coefficient AC_1 for gestures’ representation technique rating was 0.784 with a confidence interval of (0.758, 0.81). The sample of representation technique coding was classified independently by four annotators. The proportion of agreement on gestures’ representation techniques, given that the agreement was not due to chance, was significantly greater than 0.7. In particular, this result complied with the reliability level that was initially demanded.

The degree of reliability of the annotations of reference objects and context information was rated by two independent annotators. The agreement coefficient AC_1 for the classification of reference objects was 0.91, for information structure 0.95, for information state 0.86, and for communicative goals 0.88. All values are collected in

Table 7.10. In sum, the highly interpretive Type II data showed a reasonable degree of inter-rater reliability.

Type I data The annotations that make up Type I data of the SaGA corpus transcribe orientations and movement directions as they have a clear spatial interpretation. The reliability of this data was assessed by angle-based measures. The smallest angular deviation is 2.36° for the movement direction of hand shapes, and the largest is 46.16° for BoH orientation. On average, the angular difference 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°.

Evaluating the annotation of handshapes required a special treatment, since the categories developed to classify the handshape observed comprise both Type I and Type II shares. On the one hand, 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’. Therefore, all modified handshapes were mapped onto their basic type and treated them as Type I data. As a result, it was found that the four annotators agreed on 83% (AC_1=0.9, to give the Type II statistics for comparison) of the handshapes within the reliability sample of gestures.

Table 4.4: Reliability results for the annotation of the SaGA corpus.

		AC$_1$	**Angular Deviation (SD)**
Gesture	Representation Technique	0.78	
	Handedness	0.92	
	Handshape	0.90	
	Palm Orientation		19.14° (1.92)
	BoH Orientation		20.66° (2.47)
	Wrist Movement Direction		37.08° (6.5)
Discourse Context	Thematization	0.95	
	Information State	0.86	
	Communicative Goal	0.88	
Referent Object		0.91	

In sum, the evaluation of the secondary data of the SaGA corpus revealed a satisfactory degree of reliability. Chance-corrected agreement on Type II data surpassed the threshold of 0.7. Observed inter-rater agreement on Type I data resulted 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.

4.2 Statistical Analysis

The statistical analyses of the SaGA data were conducted in three steps. First, it was investigated *whether* gestures are employed and by which factors is gesture use modulated. Second, the use of representation techniques was analyzed accordingly. And finally, the degree to which form features are combined in the representation techniques was examined, and, again, by which factors this is modulated. Due to the focus on object description in this thesis, the majority of analyses will be based on a sub-corpus of object descriptions of four sights (town hall, church square, chapel, and fountain).

Methodologically, the statistical investigation was based on frequency distribution analyses and Pearson's chi-square tests to measure the association between two categorical variables (cf. Bortz, 2005). The latter compares an observed frequency distribution with an expected frequency distribution and measures the level of mismatch between the expected and observed frequencies over levels of categories. Note that, using the chi-square tests is inappropriate if any value for expected frequencies is below one or if the expected frequency is less than five in more than 20% of the contingency table cells. Therefore, the test was only applied where this requirement was given. With regard to inter-individual differences, only a limited number of speakers was considered who produced a considerable amount of gestures respectively. For reasons of clarity, statistical tests which failed to reject the null hypothesis at a significance level of $p<.05$ will not be reported. Results are marked with * for $p<.05$ with ** for $p<.01$, and with *** for $p<.001$.

4.2.1 Gesture Rate—To gesture or not to gesture?

A total of 4382 gestures was used by the 25 speakers (direction givers). Another 579 gestures in the corpus were produced by the addressees (direction followers), resulting in a total of 4961 gestures in the SaGA corpus. To quantify the rate of gesturing there are generally two methods to be employed: the number of gestures divided by speaking time, or alternatively, the rate of gestures per 100 words which adjusts for differences in speaking rate. Values for both measures are reported for the SaGA corpus in Table 4.5.

Inter-individual Differences

Concerning the number of gestures divided by speaking time, the analysis revealed a mean rate of 15.8 gesture strokes per minute. Notably, there was a considerable degree of inter-individual variation across speakers which became apparent in the standard deviation of 7.1. Gesture rates varied between a minimum of 3.0 for participant P2 and a maximum of 33.0 for participant P16.

A similar picture resulted from the analysis of gesture rate per 100 words. Here, the mean rate was 10.3 with a standard deviation of 4.0 which is approximately half of the mean. The minimum rate was 2.3 for speaker P2 and the maximum was observed for speaker P16, who produced 21.24 gestures per 100 words. That is, these two speakers turned out to gesture least/most as measured by both kinds of quantification.

Table 4.5: Gesture rates for the 25 speakers in the SaGA corpus.

Participant	**Number of gestures per 100 words**	**Number of gestures per minute**
P1	11.8	18.0
P2	2.3	3.0
P3	10.2	11.6
P4	10.9	19.4
P5	16.7	24.4
P6	11.8	21.1
P7	13.0	25.3
P8	13.4	19.7
P9	14.3	19.9
P10	11.8	19.5
P11	11.2	21.0
P12	6.8	10.2
P13	3.1	3.7
P14	9.5	13.8
P15	9.0	12.5
P16	21.2	33.0
P17	5.5	6.8
P18	5.9	7.8
P19	11.4	20.6
P20	11.6	20.6
P21	8.2	12.3
P22	8.5	10.9
P23	9.6	12.5
P24	9.0	13.5
P25	10.3	15.2
Total	*M*=10.3 *SD*=4.0	*M*=15.8 *SD*=7.1

Modulating Factors

The choice[3] to produce a gesture was decisively influenced by variables of four kinds as displayed in Table 4.6: discourse context, linguistic context, referent features, and the previous gesturing behavior. To also investigate in how far individual speakers differ with regard to gesture use, the interrelations were surveyed based on a set of

[3] The term ‘choice’ is not meant to imply a conscious process here.

Table 4.6: Interrelation of gesture occurrence and influencing variables. Parenthetical values are expected occurrences. For each interrelation the data of individual speakers was analyzed for significance (p<.05) and whether there is a similar distribution as in the combined data.

		Gesture (y/n)		**Individuals**	
		no gesture	**gesture**	**Signif.**	**Sim. Distr.**
Thematization	**rheme**	103 (142.8) ***	248 (208.2) **	4/5	5/5
	theme	96 (56.2) ***	42 (81.8) ***		
InfoState	**private**	83 (102.1)	168 (148.9)	2/5	5/5
	shared	116 (96.9)	122 (141.1)		
CommGoal	**lmIntro**	17 (10.6) *	9 (15.4)	2/5	5/5
	lmDescrProp	67 (65.1)	93 (94.9)		
	lmDescrConstr	54 (49.6)	68 (72.4)		
	lmDescrPos	61 (73.7)	120 (107.3)		
MainAxis	**none**	44 (47.2)	72 (68.8)	3/5	4/5
	width	44 (43.5)	63 (63.5)		
	height	100 (87.9)	116 (128.1)		
	depth	11 (20.3) *	39 (29.7)		
Subparts	**none**	68 (92.8)**	160 (135.2)*	2/5	5/5
	1 or more	131 (106.2) *	130 (154.8) *		
SymAxes	**none**	97 (74.5) **	86 (108.5) *	2/5	5/5
	sym	102 (124.5) *	204 (181.5)		
PrevGesture	**gesture**	68 (89.1)*	166 (140.9)	3/5	3/5
	no gesture	118 (92.9)*	137 (147.1)		

489 noun phrases was considered, including 290 gestures from five speakers who gestured at relatively high rates (P1, P5, P7, P8, and P15).

- **Discourse Context**
  For the discourse context, thematization was found to be decisive insofar as rhematic information is significantly more likely to be accompanied by a gesture (χ^2=66.396, *df*=1, *p*<.001). Individuals shared this relationship: although the relation was not significant for any one speaker, all five speakers agreed on the distribution, i.e., they tended to use gestures for rhematic information whereas for thematic information, gestures were less likely to occur. Regarding the information state, people were more likely to produce gestures for entities whose information state was private (χ^2=12.432, *df*=1, *p*<.001). This is in line with the view that new information is introduced into the discourse by gesture McNeill (1992). Again, all individuals shared the same distribution, although the relation was only significant for two of five speakers. So it seems as if this link between information state and gesture occurrence is not as strong as the link between thematization and gesture occurrence. Moreover, the communicative goal had an impact on the question of whether or not to gesture (χ^2 =10.970, *df*=3, *p*=.012). When a landmark was just mentioned (lmIntro), this utterance was not very likely to be accompanied by a gesture. This dependence between variables,

however, was only significant for two individuals, although all five agreed on the distribution by trend. That is, they used fewer gestures than expected for landmarks which had just been mentioned without further elaboration of any kind.

- **Referent Features**
  As concerns the influence of referent features, three features appeared to be decisive. First, there was a significant relationship between the choice to gesture and the referent's main axis: if from the speaker's point of view of an object's main axis was its depth (e.g., a tunnel into which one is looking) a gesture was more likely to be produced than in other cases (χ^2=10.424, *df*=3, *p*=.015). For three of the five speakers, this relation was significant, and only one speaker did not share the trend. Moreover, the complexity of the object (part) was influential. Utterances referring to leaf nodes of the IDT representation were more often accompanied by gestures than utterances referring to inner nodes of the tree representation (χ^2=20.916, *df*=1, $p<.001$). All individuals shared this kind of distribution, however, it was only significant for two of them. Furthermore, for objects which have at least one symmetry axis, gestures are more likely to occur than for completely asymmetrial objects (χ^2=18.363, *df*=1, $p<.001$). Again, all speakers shared this kind of distribution, but it was only significant for two of them.

- **Previous Gesture**
  Another factor found to be influential for gesture use was whether the speaker performed a gesture beforehand, or whether the hands were in a rest position. This relationship was highly significant (χ^2=19.09, *df*=1, $p<.001$) and based on the fact that speakers were more likely to gesture when they had gestured immediately before. In contrast, they were less likely to gesture when being in a rest position prior to that. This relation was observed in three of the five speakers under consideration.

- **Linguistic Context**
  For the analysis of gesture use and NP patterns, data of 4156 noun phrases was taken into account. The analysis aimed to correlate NP patterns with gesture use, whereby 37.8% of the NPs were accompanied by gestures (see Table 4.7). There was a highly significant correlation between the use of particular syntactic NPs and the use of gestures (χ^2 =248.89, *df*=14, $p<.001$). This relationship was due to the fact that some syntactic constructions, namely 'ART ADJA NN', 'ART NN' were significantly more likely to be accompanied by a gesture. On the contrary, other constructions, such as 'PIS' or 'PPER' were significantly less likely to occur with a co-speech gesture.

Table 4.7: Interrelation of gesture occurrence and NP type. Parenthetical values are expected occurrences.

NP type	no gesture	gesture
ADJA NN	54 (61.5)	45 (37.5)
ART ADJA NN	88 (130.5)***	122 (79.5)***
ART ADJA	45 (55.9)	45 (34.1)
ART NN	410 (500.9)***	396 (305.1)***
PDAT NN	37 (47.2)	39 (28.8)
CARD	19 (25.5)	62 (49.6)
CARD NN	19 (25.5)	40 (34.1)
ART	154 (152.3)	91 (92.7)
NN	137 (137.4)	84 (83.6)
PRF	52 (54.7)	36 (33.3)
PDS	212 (200.7)	111 (122.3)
PIAT NN	36 (30.5)	13 (18.5)
PIS	134 (115.6)	52 (70.4)*
PPER	857 (678.1)***	234 (412.9)***
Other	298 (336.2)*	243 (204.8)**

In summary, the decision as to whether to gesture was influenced by four variable kinds: (1) the discourse context, (2) referent features, (3) the previous gesturing behavior, and (4) the linguistic context. These findings were mostly systematic, i.e., uncontroversial among the five speakers we looked at. However, a significance of the very correlation was not given for all individuals. In other words, speakers varied particularly in how strong the link between particular variables was.

4.2.2 Gestural Representation Techniques

The use of gestural representation techniques and the influence of modulating factors on their choice will, as in the previous analysis of gesture use in general, take inter-individual differences into account. Accordingly the analysis will be based on the sub-corpus of 290 gestures from five speakers.

Inter-individual Differences

The use of gestural representation techniques was subject to major inter-individual differences (χ^2=85.34, *df*=16, $p<.001$). Table 4.8 shows the varying distributions of representation techniques for five speakers (P1, P5, P7, P8, P15), who gestured a relatively high rates, in the sub-corpus of object descriptions. For speaker P1, e.g., a significantly increased number of drawing gestures was observed, whereas the number of placing gestures was significantly below the expected amount. Speaker P15, by contrast, produced no posturing gestures at all, but instead used more abstract indexing gestures than expected. The distribution of representation techniques for speaker P5 is characterized by a high number of shaping and posturing gestures, whereas the number of drawing and abstract indexing gestures was lower as expected.

Table 4.8: Comparison of five speakers with regard to inter-individual differences in the use of gestural representation techniques. Parenthetical values are expected occurrences.

	Individuals				
	P1	**P5**	**P7**	**P8**	**P15**
Abstract Indexing	7 (4.4)	6 (11.9)	6 (10.1)	4 (6.6)	17 (7.0)***
Placing	1 (7.1)*	17 (19.0)	25 (16.1) *	14 (10.6)	7 (11.3)
Shaping	7 (12.9)	42 (34.7)	28 (29.5)	24 (19.4)	16 (20.6)
Drawing	15 (4.7)***	6 (12.8)*	6 (10.8)	5 (7.1)	11 (7.6)
Posturing	2 (2.9)	15 (7.7)**	8 (6.5)	1 (4.3)	0 (4.6)**

Modulating Factors

- **Referent Features**

  The first analysis aimed to correlate the use of representation techniques with the spatio-geometrical properties of the objects described. It turned out that there was a significant difference between object shapes which can be decomposed into detailed subparts (part-whole relations) and objects without any subparts (χ^2=32.39, *df*=4, $p<.001$). For objects without subparts, the number of shaping, drawing and posturing gestures was increased, whereas the rate of placing gestures was significantly decreased. For objects which have at least one subpart, placing gestures occurred significantly more often than expected, while shaping, drawing and posturing gestures occurred less often (Table 4.9). Thus, if an object was minimally complex in the sense that it had no subparts, and therefore seems more amenable to gestural reconstruction, depicting gestures were preferred. For more complex objects, with at least one subpart, participants preferred placing gestures.

  Another way to assess shape complexity of an object is in terms of its inherent symmetry: the more symmetric axes an object has, the less complex one's perception of it. Thus, the correlation between representation technique and existence of symmetry in the reference object was investigated. Again, a significant relationship (χ^2=26.90, *df*=4, $p<.001$) was found: objects which had no symmetrical axis, i.e., were more complex, indexing and placing gestures are used relatively often, while drawing, shaping, and posturing gestures were used less often than expected (see Tab. 4.9). In contrast, if an object had at least one symmetrical axis, the number of drawing, shaping, and posturing gestures increased, whereas the number of indexing and placing gestures decreased. This is in line with the above finding, that complex objects were likely to be positioned gesturally, while less complex objects were more likely to be depicted by gesture.

- **Discourse Context**
  A further analysis investigated the correlation of representation technique and dialog context, whereby a significant relationship was found between the communicative goal and the use of representation techniques (χ^2=81.206, *df*=8, $p<.001$). Descriptions came along with significantly more depicting gestures (shaping, drawing, posturing), while the spatial arrangement of entities was accompanied by indexing and placing gestures in the majority of cases.

- **Linguistic Context**
  There was a significant relationship found between the use of representation techniques and NP type (χ^2=160.784, *df*=70, p < .001). For this analysis, data of 4156 NPs was taken into account (see Table 4.10). On closer inspection, different gestural representation techniques co-occurred with certain NP patterns in a significant way. For patterns (1)-(4), i.e., those consisting of determiners, adjectives, and nouns, the number of shaping gestures was significantly increased in comparison with expectation. NPs consisting of determiner and noun (5), co-occurred with posturing gestures significantly more often than expected. Moreover, for cardinals (NP patterns (6)/(7)), placing gestures, and demonstrative pronouns, (11) posturing gestures were used more often than expected. Further, indexing gestures were frequently used along with personal pronouns (14), and for reflexive personal pronouns (15) the number of shaping gestures was significantly increased.

- **Previous Gesture**
  Regarding the influence of the previously performed gesture, there was also a significant relationship (χ^2=67.39, *df*=16, p < .001). This is due to the fact that speakers tend to stay in the same technique: Drawing gestures tend to be followed by drawing gestures, again. The same holds for posturing gestures. In addition, shaping gestures tend to follow placing gestures. However, these relations were not significant for single individuals.

Table 4.9: Interrelation of gestural representation techniques and modulating factors. Parenthetical values are expected occurrences. For each interrelation the data of individual speakers was analyzed for significance (p<.05) and whether there is a similar distribution as in the combined data.

		Representation Techniques					Individuals	
		Indexing	**Placing**	**Shaping**	**Drawing**	**Posturing**	**Signif.**	**Sim. Distr.**
Thematization	**rheme**	26 (34.2)	50 (54.7)	112 (100.1)	42 (36.8)	18 (22.2)	5/5	3/5
	theme	14 (5.8)***	14 (9.3)	5 (16.9)**	1 (6.2) *	8 (3.8)**		
InfoState	**private**	7 (23.2)***	41 (37.1)	71 (67.8)	34 (24.9)	15 (15.1)	3/5	5/5
	shared	33 (16.8)***	23 (26.9)	46 (49.2)	9 (18.1)*	11 (10.9)		
CommGoal	**lmIntro**	2 (1.2)	5 (2.0)*	2 (3.6)	0 (1.3)	0 (0.8)	2/5	5/5
	lmDescrProp	7 (12.8)	16 (20.5)	44 (37.5)	18 (13.8)	8 (8.3)		
	lmDescrConstr	3 (9.4)*	8 (15.0)	38 (27.4)*	10 (10.1)	9 (6.1)		
	lmDescrPos	28 (16.6)**	35 (26.5)	33 (48.4)*	15 (17.8)	9 (10.8)		
MainAxis	**none**	3 (9.9)*	6 (15.9)*	42 (29.0)*	9 (10.7)	12 (6.5)*	3/5	4/5
	width	1 (8.7)*	9 (13.9)	53 (25.4)	16 (9.3)*	2 (5.6)		
	height	31 (16.0)***	43 (25.6)***	18 (46.8)***	14 (17.2)	10 (10.4)		
	depth	5 (5.4)	6 (8.6)	22 (15.7)	4 (5.8)	2 (3.5)		
Subparts	**none**	19 (22.1)	18 (35.3)**	72 (64.6)	30 (23.7)	21 (14.3)	4/5	5/5
	1 or more	21 (17.9)	46 (28.7)**	45 (52.4)	13 (19.3)	5 (11.7)		
SymAxes	**none**	17 (11.9)	32 (19.0)**	24 (34.7)	11 (12.8)	2 (7.7)*	1/5	5/5
	sym	23 (28.1)	32 (45.0)	93 (82.3)	32 (30.2)	24 (18.3)		
PrevGesture	**Indexing**	10 (7.4)	2 (9.9)*	21 (17.9)	11 (7.1)	2 (3.9)	0/5	1/5
	Placing	2 (9.9)**	13 (10.8)	31 (21.2)*	4 (9.3)	6 (4.8)		
	Shaping	21 (17.9)	25 (19.5)	39 (38.2)	12 (16.8)	4 (8.6)		
	Drawing	11 (7.1)	5 (7.7)	7 (15.1)*	14 (6.6)**	3 (3.4)		
	Posturing	2 (3.9)	1 (4.2)	9 (8.3)	2 (3.7)	8 (1.9)***		

Table 4.10: Noun phrase patterns in relation with gestural representation techniques (part-of-speech tags according to the Stuttgart-Tübingen Tagset (STTS) for German).

	NP patterns	Frequency (%)	Representation technique
(1)	ADJA NN	2.4	shaping **
(2)	ART ADJA	2.2	
(3)	ART ADJA NN	5.1	
(4)	ART	5.9	
(5)	ART NN	19.4	posturing **
(6)	CARD	1.0	placing ***
(7)	CARD NN	2.2	
(8)	NE	1.2	posturing *
(9)	NN	5.3	
(10)	PDAT NN	1.8	
(11)	PDS	7.8	
(12)	PIAT NN	1.2	
(13)	PIS	4.5	
(14)	PPER	26.3	indexing *
(15)	PRF	2.1	shaping *

4.2.3 Form Feature Analysis of Representation Techniques

In order to further systemize the combination of form features in gestures, the data was analyzed for the use of gesture form feature values. To this end, it was convenient to consider the sub-categorization of iconic gestures into representation techniques. So, in the following, the sub-corpus of 1087 gestures from 25 speakers will be explored separately for each representation technique. In each case, the characteristics of the very technique will be broken down in terms of common technique-specific patterns as well as residual degrees of freedom which might be sensitive to referent characteristics and inter-individual differences. The total of 1087 object description gestures subdivided into 20.1% abstract indexing gestures, 11.7% placing gestures, 28.0% shaping gestures, 11.2% drawing gestures, and 6.3% posturing gestures. A proportion of 22.7% were of other types (sizing, counting, hedging), or combinations of several representation techniques.

To prevent from narrowing results to very specific value interrelations, handshape categories were assigned to five major categories: ‘ASL-B’, ‘ASL-C’, ‘ASL-G’, ‘ASL-O’ and ‘bent-5’. That is, values like ‘ASL-B-spread’ or ‘ASL-B-loose’ fall into the general category ‘ASL-B’. Regarding palm and BoH orientations, combined categories were split into their parts, e.g., ‘PAB/PDN’ into ‘PAB’ and ‘PDN’. Furthermore, ‘PTR’/‘BTR’ and ‘PTL’/‘BTL’ were merged into a common category ‘PTS’/‘BTS’ (for sideways palm/BoH orientation) to abstract from handedness.

Abstract Indexing Gestures

219 gestures in the sub-corpus of object descriptions were indexing gestures. The characteristics of these gestures are summarized in Table 4.11 with regard to what they have in common as opposed to gestures of other representation techniques. In the following, the degree to which the form features of indexing gestures are technique-specific and how variant gesture form features are shaped by modulating factors will be investigated. Inter-individual differences were tested for a subset of four from the 25 speakers who produced a sufficient number (greater or equal than 15) of indexing gestures (P9, P11, P20, P24).

Table 4.11: Gesture form feature analysis in indexing gestures (N=219).

Variable	Value	Relative Frequency	Number of observed (expected) occurrences
Handedness	LH	42.0%	92 (61.7)***
	RH	53.0%	116 (103.0)
	2H	5.0%	11 (54.4)***
Handshape	ASL-B	32.4%	71 (59.2)
	ASL-G	30.1%	66 (40.9)***
	ASL-C	7.8%	17 (34.3)**
	ASL-O	2.3%	5 (2.6)
	ASL-bent-5	1.4%	3 (17.1)***
Palm Orientation	PAB	23.3%	41 (35.5)**
	PTB	10.0%	22 (12.5)**
	PTS	45.7%	100 (80.6)*
	PUP	0.9%	2 (10.7)**
	PDN	33.3%	73 (58.6)
BoH Orientation	BAB	59.8%	131 (99.1)**
	BTB	–	0 (0.6)
	BTS	36.1%	79 (46.7)***
	BUP	26.5%	58 (48.0)
	BDN	3.2%	7 (4.2)
Wrist Movement	no movement	88.1%	193 (127.1)***

- **Handedness**
  A common feature of indexing gestures was that they were performed with one hand, either the right hand (53.0%) or the left (42.0%). However, although handedness was restricted to one-handed gestures in this representation technique, which hand to use, the left or the right, remained open. This choice was actually found to be constrained by two different factors: the referent's position as well as inter-individual differences. With regard to referent **position**, the use of right-handed gestures was positively correlated with object position (χ^2=113.70, *df*=4, $p<.001$): objects located on the right hand side were significantly more often than expected localized with right-handed indexing gestures.

Along the same lines, there was a also a significant relationship for left-handed gestures and objects located at the left. In addition, **inter-individual differences** of handedness (χ^2=74.87, *df*=6, $p<.001$) were due to the fact that one speaker (P24) preferred two-handed gestures, another speaker (P20) preferred left-handed gestures, and two other speakers (P09, P11) had increased numbers of right-handed gestures as compared to expectations.

- **Handshape**
  Handshape choice was basically limited into two major categories: for 71 gestures (32.4%) of the gestures ASL-B was used, and for another 66 gestures (30.1%) the typical pointing handshape ASL-G was employed. Which of these was actually used was subject to significant **inter-individual differences** (χ^2=10.43, *df*=4, *p*=.034). Although in general, the proportions of ASL-B use and ASL-G use were nearly equal, there was one individual (P9) in the test set who obviously preferred flat hand indexing while, of the others, three preferred pointed indexing. When taking all 25 speakers into account (most of them with quite low data cases, however), there was approximately one third preferring indexing gestures with ASL-B, one third preferring indexing gestures with ASL-G, and another third without a preference.

- **Palm Orientation**
  In a relatively large proportion of indexing gestures, sideways palm orientation occurred as a component (45.7%). Particularly, the number of rightwards oriented palms was larger than expected. In addition, the palm orientation component 'PDN' was contained in 33.3% of the indexing gestures, and 'PAB' was still found in 23.3%, which was significantly above expectation.

- **BoH Orientation**
  With regard to BoH orientation, the most frequently occurring orientation was away from the speaker's body. In 59.8% of the indexing gestures, the orientation 'BAB' was at least a component of diagonal or mixed orientations. Moreover, sideways BoH orientations ('BTL' and 'BTR') were observed in 36.1% of indexing gestures, which was more frequent than expected.

- **Movement Features**
  The majority of indexing gestures had in common that their strokes were predominantly static, i.e., hands are just brought to a particular position in the preparation phase of the gesture and they were not moved until the retraction phase started. At total of 193 indexing gestures (88.1%) had static strokes.

So far, the form features of abstract indexing gestures have been analyzed separately. Next, it will be investigated if there are any technique-specific patterns of co-occurence. Actually, handshape use divides the set of indexing gestures into two classes: *flat hand* *indexing*, realized with the flat handshape ASL-B, and *pointed indexing*, realized with the pointing handshape ASL-G.

These two sub-types differed from each other particularly with regard to palm orientation (χ^2=12.21, *df*=2, *p*=.002): as visualized in Figure 4.4, pointed indexing gestures were characterized by a relatively large proportion of gestures with downwards palm orientation: in 48.5% of the gestures, the orientation 'PDN' was at least one component of diagonal or mixed orientations (observed: 32, expected: 23.1, *). In contrast, the downwards-oriented palm was observed less often than expected in flat hand indexing gestures (observed: 16, expected: 24.9, *).

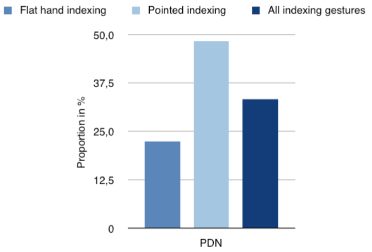


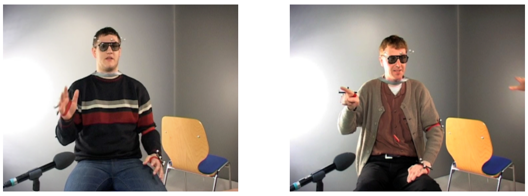
Figure 4.4: Flat hand indexing and pointed indexing differ with regard to the palm orientation category 'PDN'.

See Figure 4.5 for examples of flat hand indexing and pointed indexing. Figure 4.5(a) is an example of flat hand indexing: a right-handed indexing gesture with handshape ASL-B in which palm is oriented leftwards and the BoH away from the speaker's body. This gesture is referring to the townhall's left staircase[4]. A typical example of pointed indexing is given in Figure 4.5(b): a right-handed indexing gesture with handshape ASL-G. As in most gestures of this type, the palm is oriented downwards and the BoH directed away and to the right, indicating the location of the 'right church' the speaker is referring to.

Placing Gestures

A total of 127 of the gestures under consideration were placing gestures used to place or set down objects in gesture space. Regarding the question of how abstract indexing gestures appear and how their form differs from other gestures in the corpus, a summary of gesture form features occurring in these gestures is given in Table 4.12. The first step in the analysis of placing gestures was an examination of form features with regard to technique-specific characteristics and modulating factors. The analysis

[4] Here the speaker mixed up left and right: although verbally referring to the left staircase, he is pointing to the right.



(a) Right-handed indexing gesture with handshape ASL-B accompanying “fire scape to the left”.

(b) Right-handed indexing gesture with handshape ASL-G for “the right church”.

Figure 4.5: Examples of gestures from the two classes of indexing gestures: flat hand indexing and pointed indexing.

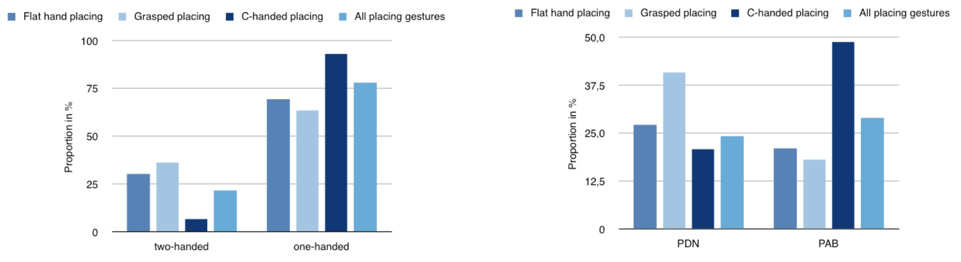
of inter-individual differences was based on a comparison of three speakers in whose data the number of placing gestures was greater or equal to ten (P5, P9, P11, P19).

- **Handedness**
  With regard to handedness, placing gestures did not differ significantly from other gestures. There were two significantly positive correlations of handedness with referent features, namely with the referent’s **main axis** (χ^2=23.26, *df*=6, *p*=.001) and its **position** (χ^2=22.39, *df*=4, $p<.001$). The former was due to the fact that objects with the x-axis (width) as their main axis were more often than expected referred to with two-handed and right-handed placing gestures, whereas objects with the y-axis (height) and z-axis (depth) as their main axis were more often than expected referred to with left-handed gestures. The statistically significant relationship between handedness and position was based on the fact that two-handed gestures tend to be used pre-dominantly for objects located in the middle, left-handed for objects at the left hand side, and right-handed gestures for objects at the right hand side.
  In addition, handedness was found to be subject to **inter-individual differences** (χ^2=35.17, *df*=6, $p<.001$) due to the fact that two speakers (P11, P19) preferred two-handed gestures, another speaker (P5) preferred left-handed gestures, and speaker P9 had an increased number of right-handed placing gestures.

- **Handshape**
  Concerning handshape use, ASL-B and ASL-C were the most often used handshapes. Whereas the number of gestures with ASL-B was basically in line with the expected number, the number of gestures with ASL-C was significantly increased as compared to the expected number. Similarly, the observed number of gestures with ASL-bent-5 was significantly greater than expected, although these gestures make up only 15.0% of all placing gestures.

Handshape use was also subject to **inter-individual differences** (χ^2=18.11, *df*=6, *p*<.006). In fact, the three speakers under consideration had different preferences each: speaker P9 had an increased number of handshape ASL-B, speakers P9 and P11 produced more gestures with handshape ASL-bent-5 than expected, and for speaker P19, the number of ASL-C handshapes was greater.

- **Palm orientation**
  Placing gestures showed a relatively high proportion of sideways palm orientations (54.3%). The second most frequent palm orientation component was 'PAB' (29.1%). For both categories, the observed number of placing gestures with this palm orientation component was significantly increased relative to expectations.
- **BoH Orientation**
  BoH orientations in placing gestures were characterized by a high proportion of 'BAB' (44.3%), whereby the observed frequency was significantly increased in comparison with the expected frequency. Furthermore, there was a large proportion of gestures with upwards oriented BoH ('BUP', 30.7%) and sideways oriented palm ('BTS', 29.9%). Both proportions were significantly higher than expected.
- **Movement Features**
  Similar to abstract indexing gestures, almost all placing gestures had in common that their strokes were not dynamic, i.e., hands were brought to a particular position in the preparation phase of the gesture and were not moved until the retraction phase started.



(a) (b)

Figure 4.6: Comparison of flat hand placing, grasped placing, and c-handed placing gestures with regard to handedness (a) and palm orientation (b).

Were there any technique-specific patterns of how form features co-occur in placing gestures? Generally, placing gestures sub-divided into gestures in which the

Table 4.12: Gesture form feature analysis in placing gestures (N=127).

Variable	Value	Relative Frequency	Number of observed (expected) occurrences
Handedness	LH	34.6%	44 (35.8)
	RH	43.3%	55 (59.7)
	2H	22.0%	28 (31.5)
Handshape	ASL-B	26.0%	33 (34.3)
	ASL-G	1.6%	2 (23.7)***
	ASL-C	33.9%	43 (19.9)***
	ASL-O	0.0%	0 (1.5)
	ASL-bent-5	17.3%	22 (9.9)***
Palm Orientation	PAB	29.1%	37 (20.6)***
	PTB	6.3%	8 (7.3)
	PTS	54.3%	69 (46.7)***
	PUP	6.3%	8 (6.2)
	PDN	24.4%	31 (34.0)
BoH Orientation	BAB	44.3%	67 (57.5)
	BTB	1.6%	2 (0.4)
	BTS	29.9%	38 (27.1)*
	BUP	30.7%	39 (27.8)*
	BDN	1.6%	2 (2.5)
Wrist Movement	no movement	62.8%	107 (73.7)***

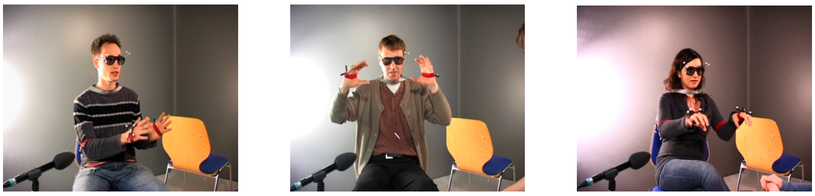
handshape ASL-B was used (*flat hand placing*), gestures with handshape ASL-C (*c-handed placing*) and gestures with the grasping handshape ASL-bent-5 (*grasped placing*). These three types were characterized by differences in the co-occurrence of handshape use and palm orientation (see Figure 4.6) as well as handedness (see Figure 4.6(a)).

Flat hand placing gestures with handshape ASL-B were typically realized with both hands (observed: 10, expected: 7.1). In Figure 4.7(a), a typical example is given for this type of two-handed placing gesture, with the right palm oriented to the left and the BoH oriented away from the body.

Grasped placing gestures were also more often than expected realized with both hands (observed: 8, expected: 4.7) and by a downwards-oriented palm (observed: 9, expected: 6.1). In Figure 4.7(c), an example of a typical two-handed grasped placing gesture is given.

C-handed placing gestures were characterized by the use of handshape ASL-C. These gestures were, in contrast to the other types, often performed with one hand, i.e., the proportion of two-handed gestures was rather low (observed: 3, expected: 9.2, *). Additionally, these were characterized by a relatively large proportion of palm orientations away from the speaker's body (observed: 21, expected: 14, *). Furthermore, there was a trend, although not significant, showing an increased proportion of the BoH orientation 'BUP' in c-handed placing gestures (observed: 21, expected: 16.2)

See Figure 4.7(b) for an example of two one-handed placing gestures with handshape ASL-C in which the palm of both hands is oriented away/sideways and the BoH upwards.



(a) Typical two-handed placing gesture with handshape ASL-B accompanying “there’s a church to the left”.

(b) Two one-handed placing gesture with handshape ASL-C for for the two churches.

(c) Two-handed placing gestures with handshape ASL-bent-5 for the two churches.

Figure 4.7: Examples of placing gestures.

Shaping Gestures A total of 304 gestures in the data were shaping gestures. Their form features will be analyzed in the following with regard to technique-specific characteristics and modulating factors. The analysis of inter-individual differences was based on a comparison of six speakers in whose data the number of shaping gestures was greater or equal to 15 (P5, P6, P8, P9, P11, P15).

- **Handedness** Shaping gestures were characterized by a high number of two-handed gestures (42.1%), which was significantly above expectations. The observed number of one-handed, and in particular left-handed, gestures was decreased as compared to expectations. Handedness was found to be subject to significant **inter-individual differences** (χ^2=45.81, df =10, p<.001) due to the fact that one speaker (P08) preferred two-handed shaping gestures, two other speakers (P5, P15) preferred left-handed gestures, and another, speaker P9, had an increased number of right-handed shaping gestures.

- **Handshape**
According to handshape use, 41.1% of shaping gestures were realized with the flat handshape ASL-B. This observed frequency of appearance of these gestures was significantly above their expected number of occurrences. Further, the handshape ASL-C was employed in 19.1% (more or less in line with the expected proportion), and ASL-bent-5 was used in 12.8% which was significantly above the expected proportion. Examples for shaping gestures with the two most frequent handshapes are given in Figure 4.8.

Table 4.13: Gesture form feature analysis in shaping gestures (N=304).

Variable	Value	Relative Frequency	Number of observed (expected) occurrences
Handedness	LH	17.1%	52 (85.6)***
	RH	40.8%	124 (142.9)
	2H	42.1%	128 (75.5)***
Handshape	ASL-B	41.1%	125 (82.2)***
	ASL-G	6.3%	19 (56.8)***
	ASL-C	19.1%	58 (47.5)
	ASL-O	0.7%	2 (3.6)
	ASL-bent-5	12.8%	39 (23.8)**
Palm Orientation	PAB	17.9%	36 (49.2)
	PTB	3.0%	9 (17.3)*
	PTS	34.9%	106 (111.9)
	PUP	3.9%	12 (14.8)
	PDN	23.0%	70 (81.4)
	Dynamic	34.2%	104 (50.1)***
Finger Orientation	BAB	42.8%	130 (137.6)
	BTB	0.3%	1 (0.8)
	BTS	14.1%	43 (64.9)*
	BUP	18.8%	57 (66.6)
	BDN	1.3%	4 (5.9)
	Dynamic	32.9%	100 (57.9)***
Wrist Movement	curved	29.6%	90 (44.7)***
	linear	41.6%	140 (72.2)***
	linear+curved	3.9%	12 (8.4)
	no movement	19.7%	60 (176.5)***

Handshape choice was found to be positively correlated with the referent's **main axis** (χ^2=20.75, *df*=6, *p*=.002): objects without a main axis were more often than expected depicted by gestures with handshape ASL-bent-5, whereas for referent objects with a main axis (either y-axis or z-axis), the number of gestures with handshape ASL-C was greater than expectation.

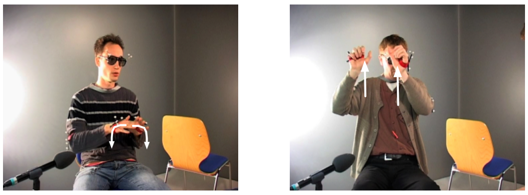
Handshape use in shaping gestures was also subject to **inter-individual differences** (χ^2=62.23, *df*=10, $p<.001$). This was due to the fact that some speakers had strong preferences for particular handshapes. For instance, P5 preferred ASL-C (observed: 8, expected: 3.2, **), or P11 preferred ASL-bent-5 (observed: 29, expected: 9.0, ***).

– **Palm Orientation**

Palm orientation in shaping gestures was characterized by a large proportion of dynamics. In 34.2% of the gestures, the palm orientation was modified during the gesture stroke, which was significantly above the expected number of occurrences. In gestures with a static palm orientation, sideways palm orientations were frequently used (34.9%), as well as downwards-oriented palms (23.0%).

- **BoH Orientation**
  BoH orientations were also characterized by a high degree of dynamism. With 32.9% of shaping gestures, the proportion of strokes with changing BoH orientation was significantly above the expected number. Other frequently occurring BoH orientation components were ‘BAB’, with 42.8% and ‘BUP’, with 18.8% of the shaping gestures.

- **Wrist Movement**
  Additonally, shaping gestures were characterized by a relatively high degree of dynamics in wrist location. Only in 19.7% of the gestures were executed without wrist movement. Apparently, the referent shape and the movement trajectory of shaping gestures were related to each other. The relationship of referent shape and the movement features of shaping gestures became apparent in significantly positive correlation between the referent’s **shape property** and the type of movement (χ^2=31.96, *df*=6, *p*<.001): roundish objects (shape properties ‘arc-shaped’, ‘round’, ‘spherical’, ‘cup’) were almost exclusively depicted by curved movements (observed: 24, expected: 14.2, **) and straight (‘cubic’, ‘longish’, ‘squared’, ‘flat’) referents by linear movements (observed: 58, expected: 39.6, **). Figure 4.8 gives two examples illustrating how this representation technique was used to depict particular shapes.



(a) Two-handed shaping gesture with handshape ASL-B and curved trajectory accompanying “a dome-shaped roof”.

(b) Two shaping gestures with handshape ASL-C and linear trajectory referring to “two towers”.

Figure 4.8: Examples of shaping gestures.

Drawing Gestures A total 124 gestures in the data were drawing gestures. Drawing gestures were found to be subject to relatively strong technique-specific patterns as summarized in Table 4.15 and described in the following. Inter-individual differences were analyzed on the basis of data from four speakers (P1, P11, P15, P20).

- **Handedness**
  With regard to handedness, drawing gestures were mostly performed with one

Table 4.14: Gesture form feature analysis in drawing gestures (N=124).

Variable	Value	Relative Frequency	Number of observed (expected) occurrences
Handedness	LH	23.8%	29 (34.3)
	RH	63.9%	78 (57.4)**
	2H	12.3%	15 (30.3)**
Handshape	ASL-B	0.0%	0 (33.0)***
	ASL-G	76.2%	93 (22.8)***
	ASL-C	3.2%	5 (19.1)**
	ASL-O	0.0%	0 (1.5)
	ASL-bent-5	0.8%	1 (9.5)**
Palm Orientation	PAB	4.8%	6 (19.8)**
	PTB	8.2%	10 (7.0)
	PTS	12.3%	15 (45.9)***
	PUP	0.8%	1 (5.9)*
	PDN	50.8%	64 (32.7)***
	Dynamic	28.7%	35 (20.1)
Finger Orientation	BAB	43.4%	53 (55.2)
	BTB	0.0%	0 (0.3)
	BTS	18.0%	22 (26.0)
	BUP	7.4%	9 (26.7)***
	BDN	3.3%	4 (2.4)
	Dynamic	40.2%	49 (23.2)***
Wrist Movement	curved	33.6%	41 (18.0)***
	linear	30.3%	37 (29.0)
	linear+curved	13.9%	17 (3.4)
	no movement	22.1%	27 (70.8)***

hand only. The number of two-handed drawing gestures was significantly below the expected proportion. The number of right-handed drawing gestures (63.9%) was significantly larger than expected.

Handedness in drawing gestures was found to be subject to significant **inter-individual differences** (χ^2=52.35, *df*=6, $p<.001$): while P20 preferred left-handed, and P1 preferred right-handed drawing gestures, there were more two-handed drawing gestures observed for P11 and P15 than expected.

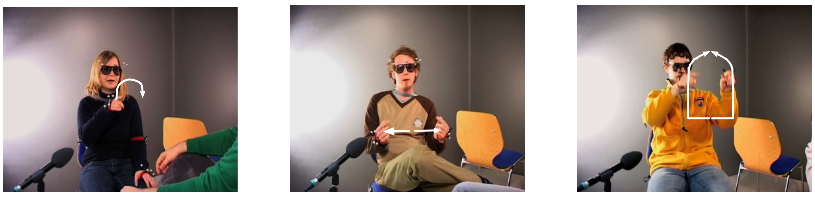
- **Handshape**
  Drawing gestures were characterized by the prevalent use of the pointing handshape ASL-G with an outstretched index finger (76.2%).

- **Palm Orientation**
  The most frequently occurring palm orientation in drawing gestures was the downwards orientation (50.8%). This observed number of 'PDN' occurrences was significantly above the number to be expected. In another 16% of the drawing gestures, the palm orientation contained the feature 'away', i.e., the palm is oriented either 'away/left', 'away/right' or 'away/down'. Other palm

orientation categories were observed only rarely. 36.3% of the gestures had dynamic palm orientations.

- **BoH Orientation**
  BoH orientations were dynamic to a relatively large extent, too (40.2%). The most frequent static category component was ‘BAB’ which occurred in 43.4% of the drawing gestures.

- **Movement Features**
  To trace the outline of an object, the gesture stroke typically contained a movement of the wrist. Only in 22.1% of the gestures there was no wrist movement (in these gestures, only the index finger was moved to trace something in the air). The relationship of referent shape and the movement features of drawing gestures became apparent in a significantly positive correlation between the referent’s **shape property** and the type of movement (χ^2=36.00, *df*=66, *p*<.001): roundish objects (shape properties ‘arc-shaped’, ‘round’, ‘spherical’, ‘cup’) were almost exclusively depicted by curved movements (observed: 20, expected: 13.2, *) and straight (‘cubic’, ‘longish’, ‘squared’, ‘flat’) referents by linear movements (observed: 13, expected: 7.5, *). Figure 4.9 gives three examples to illustrate how drawing gestures were used to depict particular shapes.



(a) Curved right-handed drawing gesture for the copula roof of a church.

(b) Linear two-handed gesture for a hedge.

(c) Two-handed gesture for a church window that combines curved and linear movement segments.

Figure 4.9: Examples of drawing gestures.

Posturing Gestures There was a total of 63 posturing gestures in the data. These gestures will be analyzed with regard to their form features and modulating factors in the following. Inter-individual differences were analyzed on the basis of two speakers’ (P5, P24) data.

Table 4.15: Gesture form feature analysis in posturing gestures (N=63).

Variable	Value	Relative Frequency	Number of observed (expected) occurrences
Handedness	LH	31.9%	22 (19.1)
	RH	36.2%	25 (32.4)
	2H	31.9%	22 (17.1)
Handshape	ASL-B	21.7%	15 (18.7)
	ASL-G	13.0%	9 (12.9)
	ASL-C	11.1%	8 (10.8)
	ASL-O	4.3%	3 (0.8)*
	ASL-bent-5	18.8%	13 (5.4)***
Palm Orientation	PAB	18.8%	13 (11.2)
	PTB	10.1%	7 (3.9)*
	PTS	56.5%	39 (25.4)**
	PUP	29.0%	20 (3.4)***
	PDN	18.8%	13 (18.5)
	Dynamic	0.0%	0 (11.4)***
Finger Orientation	BAB	59.4%	41 (31.2)
	BTB	0.0%	0 (0.2)
	BTS	27.5%	19 (14.7)
	BUP	36.2%	25 (15.1)*
	BDN	2.9%	2 (1.3)
	Dynamic	1.4%	1 (13.1)***
Wrist Movement	no movement	89.9%	62 (40.1)***

– **Handedness**
Handedness in posturing gestures does not significantly from other gesture types: approximately one third were two-handed gestures (36.2%); the others were one-handed (32.4% right-handed, 19.4% left-handed).

– **Handshape**
The most frequently used handshape in posturing gestures was ASL-B (21.7%), although this observed proportion was not significantly different from that in the other representation techniques. The second most utilized handshape in posturing gestures was ASL-bent-5 (18.8%).

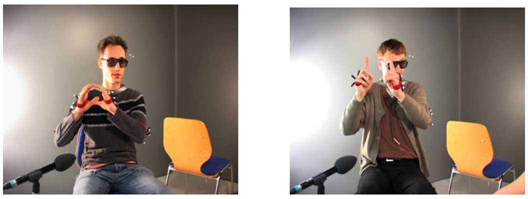
– **Palm Orientation**
The most frequently observed palm orientation component was ‘PTS’ (56.5%), which was significantly above the expected proportion as compared to other representation techniques. The second most frequently observed category was ‘PUP’ (29.0%). Notably, the number of observed gestures with this palm orientation component was significantly above the expected number (20 observed cases instead of 3.1 expected cases).

– **BoH Orientation**
In general, BoH orientation in posturing gestures was static and not dynamic.

59.4% of the posturing gestures, BoH was oriented away from the speaker's body. The second most often observed category was 'BUP' (36.2%), and in 36.5% of the posturing gestures the BoH orientation 'BTS' was used.

- **Movement Features**
  The majority of posturing gestures was characterized by static strokes (89.9%). The few dynamic strokes can be explained by the fact that iconic gestures are sometimes combined with beats (cf. McNeill, 2005).

Figure 4.10 gives two examples to illustrate how posturing gestures were used to depict particular shapes.



(a) Two-handed posturing gesture with handshape ASL-C depicting a "round window".

(b) Two posturing gestures with handshape ASL-G for "two towers".

Figure 4.10: Examples of posturing gestures.

4.3 Summary and Discussion

Building a gesture generation model under consideration of iconicity, contextual constraints, and intersubjective differences following the cyclic design methodology, requires an adequate empirical basis. The SaGA corpus provides such a data collection of 25 dyads engaged in a spatial communication task. The annotation comprises a classification for gesture representation techniques, coding of gesture form features and gestures' referent objects, as well as a transcription of the spoken words and annotation of contextual information.

The corpus consists of 4961 gesture instances. In the context of gesture research the SaGA corpus is, to the author's knowledge, by far the largest and most comprehensive collection of naturalistic, yet controlled, and systematically annotated speech-gesture data currently available. This is due to the great effort required by detailed, form-feature based gesture annotations need. Lacking the existence of sound automatic annotation tools, most existing gesture corpora are either limited to gesture segmentation and classification regarding isolated aspects of gesture use, or they are only of marginal size. The direction-giving corpus developed and employed by Kopp et al. (2007), though it consists of $\sim$1000 gestures, manually annotated for gesture form features, contains a variety of gestures referring to both actions ('turn left') and landmarks, which reduces the amount of data within single categories of interest (e.g., the number of gestures in one particular representation techniques). That is, although not the whole corpus was analyzed in detail in the scope of this thesis, the employed subsets (e.g., 1087 gestures for the form feature analysis) provide still a substantial and unique amount of detailed gesture data.

Nevertherless, the power of statistical methods was limited due to sample size: when looking at interrelations of modulating factors, e.g., analyzing the relation of referent shape and gesture form features like palm or BoH orientation given a particular representation technique (there was, e.g., only a total of 63 posturing gestures). Statistical analyses could accordingly only be applied when the sample size requirements of statistical tests were met. That is, although the statistical analysis revealed novel and detailed insights into gesture production, a (much) larger amount of data would be necessary for a complete empirical investigation of gesture use, modulating factors, inter-individual differences and their interrelations. With regard to computational modeling of gesture production, therefore, the model to be developed has to account for the **sparse data** situation.

The first analysis addressed gesture frequency. The mean rate of 15.8 gestures per minute is in agreement with gesture frequencies as reported in the literature (see Bavelas et al. (2008) for an overview). That is, the gesture rate in the SaGA study of direction giving and object descriptions from a VR stimulus was in line with gesture frequencies reported in studies in which different stimuli were described.

Findings from the analysis of gesture use in general can be summerized in three points. First, individuals were found to differ obviously in how much they gestured, ranging from $\sim$2 gestures per minute up to $\sim$30 gestures per minute. Second, whether a gesture is produced or not was found to be subject to a number of variables: referent-features, variables characterizing the linguistic and discourse context as well as previous gesturing behavior. And third, some of the correlations were found among individuals, whereas for others there was considerable variance among individuals.

The second analysis was applied to the use of gestural representation techniques and revealed similar results. First, speakers differed considerably in their preferences for particular techniques of representation. Second, the use of gestural representation techniques was found to be subject to characteristics of the referent, the linguistics/discourse-contextual situation, and previous gesture use. And third, the found correlations were not always present in the data of all speakers under consideration.

Two major implications can be drawn from these results for the design of a gesture generation model. On the one hand, the model has to be able to deal with

multiple infuencing factors. On the other hand, it has to be able to account for **inter-individual differences** and that on two levels: at the surface of gestural behavior and in how strong particular influencing relations are.

In the third analysis, representation techniques were analyzed for their form features. This investigation revealed novel and corpus-based insights into the structure of gesture techniques which can be summarized in three points: First, it turned out, that techniques were 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. Second, techniques were further characterized by being sensitive to *referent characteristics*. These features, notably, varied from technique to technique. In indexing gestures, for instance, handedness turned out to be sensitive to the position of the gesture's referent, while other form features had 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. Third, *inter-individual differences* were also found with regard to gesture form features, namely handedness and handshape choice.

That is, each technique was found to be characterized by particular conventional aspects as well as iconic aspects. In addition, the choice of handedness and handshape was shaped by the individual preferences of the speaker. The conclusion to be made regarding computational simulation, therefore, is that the generation model should be organized in a **feature-based** way to cover both conventional patterns as well as sensitivity to referent features and individuality of single gesture form features.

CHAPTER **5**

A Model of Gesture Generation

The previous chapter provided an empirical view on gesture use in the application scenario of this thesis. This chapter will now sketch a generation model of gestural behavior that makes use of the empirical insights as well as the reviews of literature and related work in chapters 2 and 3. Based on results obtained from the corpus analysis requirements for a gesture generation model will be inferred in Section 5.1 and an adequate method identified (Section 5.3.3). Section 5.4 will describe in detail how the formalism can be applied to tackle the problem of iconic gesture generation in the proposed *Generation Network for Iconic Gestures* (GNetIc).

Details of the generation model have been published in Bergmann and Kopp (2008, 2009a,b).

5.1 Requirements for a Computational Model

The empirical corpus analysis in Chapter 4 as well as the review of literature and related work in Chapters 2 and 3 revealed a number of insights which entail requirements for a gesture generation account:

1. One of the two main motivations to build a model of gesture generation is to gain insights into the gesture production process in humans. A computational model should, therefore, not only be able to achieve reasonable simulation results. It should also do this in an understandable and **interpretable** way.
2. Inter-individual differences have been found in the data on two levels: first, on the level of particular gesture features (either form features, representation techniques or gesture rate), and second, on the level of interrelations between gesture features and modulating factors. This fact calls for building speaker profiles in a **data-based** fashion which do not only cover the surface level of gestural behavior, but also the underlying contextual dependencies.

3. The approach has to be able to deal with **sparse data**. In terms of gesture research, the SaGA corpus is a rich database, for most machine learning techniques the amount of data is marginal. This is especially true for gesture form features of the particular representation techniques for which only 100–200 data cases are available. In combination with the previous point, the inter-individual differences, the number of data cases further decreases if one aims to account for speaker-specific patterns. As representation techniques are characterized by particular conventional aspects, a **rule-based** method lends itself to model the found patterns and to evaluate the generalizability of those characteristics.
4. Each representation technique is characterized by particular conventional aspects. The set of features affected by conventionality varies from technique to technique. In indexing, for instance, only the position of the referent is a variant feature, whereas other gesture form features like handshape or palm orientation are invariant technique-specific features. Generating all gestures under consideration adequately, therefore, calls for a **feature-based** approach to gesture generation.
5. Literature and corpus analytical results are in line regarding the fact that gesture use is subject to **multiple factors** such as the discourse context, linguistic factors, referent features, and previous gesturing behavior. A computational model, therefore, should be able to take these variables into account which necessitates to have access to the particular sources of information. In addition to the communicative goal on whose basis a particular gesture is to be planned, a discourse record has to be available providing information about what has been (successfully) communicated so far. The linguistic factor of thematization, further, demands that (an analysis of) concomitant speech is accessible. Referent characteristics in terms of complexity, symmetry, main axis or shape properties have to be available which necessitates either a feature-based representation of imagistic content or an analysis of imagistic models for these particular features. Finally, the previous gestures of the speaker and their detailed specifications have to be available in terms of a 'gesture history'.
   It is likely that the above-mentioned set of factors is incomplete. Research literature, rather, suggests to consider further parameters, including: the addressee in terms of recipient design, her/his gesturing behavior, characteristics of the physical environment etc. A computational account, thus, has to be easily **extensible** to be able to take further modulating factors into account at a later point in time.

Note that, requirements (3) and (4) seem to be conflicting—data-based modeling to account for inter-individual differences and rule-based modeling to deal with the fact of sparse data. This discrepancy, by the way, has already become obvious in the review of related work (Section 3.2), where common patterns of gesture use were found to be throughout modeled in a rule-based way, whereas individualization was realized either in a data-based fashion or by parameterization of pre-defined gestures. This was explained by the fact that, on the one hand, fine-grained annotated corpora of gesture use are not available in a size that is sufficient to adequately apply data-based techniques, while on the other hand, inter-individual differences are hardly investigated in a rule-based way. As a comprehensive account of gesture production, nevertheless, has to account for both, inter-individual differences and common patterns, only a **hybrid** approach would be able to meet both requirements.

5.2 Identifying an Adequate Formalism

Bayesian Decision Networks (BDNs) are such a hybrid method. They belong to the class of graphical models and provide a powerful modeling approach which is due to the fact that both probabilistic inference problems as well as deterministic decision making problems can be modeled and solved.

An integral part of BDNs are *Bayesian networks* realizing probabilistic modeling. They have been employed for modeling knowledge in a large variety of domains such as medicine, bioinformatics, information retrieval, and also in modeling human behavior. Ball and Breese (2000) employed a Bayesian network to relate emotion and personality to a variety of observable verbal and nonverbal behaviors, among others gestures[1]. Structure and parameters for the relation of personality and emotion with gesture use were defined on the basis of general trends reported in the literature. This way, gesture parameters like speed, size and repetitions were used to reflect a desired psychological state.

Pelachaud and Poggi (2002) employed Bayesian networks to link communicative functions with facial signals such as eyebrows, head direction, mouth shape, head movements, and gaze. Structure and parameters of the network were modeled based on results from empirical studies. That is, similar to Ball & Breese's modeling method, the network was empirically informed, but not directly learned from empirical data. The belief network was successfully employed to manage possible conflicts between different modalities.

More recently, in the CSG system (Section 3.2.2), Bayesian networks were employed to simulate culture-specific non-verbal behavior with virtual agents (Rehm et al., 2008). Based on the analysis of a cross-cultural corpus, Bayesian networks were created to account for culture-specific aspects in body posture and gestural expressivity. Hereby, particular parameters such as gesture speed, gesture size, power etc. were put in relation to a speaker's culture. The model was used bi-directionally, to infer the user's cultural background from his gestural behavior, and to determine a virtual agent's nonverbal behavior.

[1] Gestures were defined in a general way as body movements in their work.

The severalfold usage of Bayesian networks in the context of behavior modeling emphasizes the appropriateness of this method. Bayesian networks are advantageous for the following reasons:

- Accounting for requirement (1), learning the structure of Bayesian networks from a sample of data provides an appropriate method for behavior modeling in an understandable manner. Modeling results in terms of a Bayesian network structure are easily interpretable and intuitively meaningful since they directly represent connections between causes and effects.
- Bayesian networks are able to work with relatively sparse data. It could be shown that, in principle, Bayesian networks can have good prediction accuracy even with rather small sample sizes (Kontkanen et al., 1997). In addition, it is possible to improve prediction accuracy: structure learning algorithms can be chosen that are particularly suited for a limited amount of training data, e.g., the NPC algorithm. In addition, the search space of possible network structures can be reduced by bringing in a-priori information. It is, therefore, possible to employ a data-based method (requested by requirement (2)), even given the limited sample size of the SaGA corpus.
- Within the framework of Bayesian networks it is a simple matter to introduce new sources of information into the model (requirement (5)). It is likely that the currently available set of input variables provided by the SaGA corpus is not complete at all. Rather, further influencing factors like environmental constraints or addressee design, as discussed in Section 2.2, should also be integrated to provide a more complete picture of how gesture use is modulated by diverse factors. A Bayesian network can be easily be extended by introducing further variables, either annotated in the corpus, or inferred from that data.
- Bayesian networks provide several alternatives to deal with inter-individual differences. It is possible to learn networks from individual data of one particular speaker, to learn network structures from the combined data of several speakers, and also to apply adaptation algorithms which allow to adapt a general network structure with respect to individual patterns (Wittig, 2002).

In addition, Bayesian networks provide further advantages that do not directly follow from the set of requirements, but are nevertheless advantages in the context of gesture generation:

- Bayesian networks are able to deal with uncertainty, which is necessary to model non-deterministic connections between behavior and modulating factors. For instance, although a speaker prefers drawing gestures in particular contexts, she might at some points nevertheless use other techniques. As Bayesian networks are able to make predictions about the relative likelihood of different outcomes, they naturally capture uncertainty in human behavior.

- Bayesian networks are able to deal with uncertainty which is necessary to model non-deterministic connections between behavior and modulating factors. For instance, although a speaker prefers drawing gestures in particular contexts, she might at some points nevertheless use other techniques. As Bayesian networks are able to make predictions about the relative likelihood of different outcomes, they naturally capture uncertainty in human behavior. Moreover, Bayesian networks are able to deal with incomplete data. If not all input variables can be provided with evidence, the network is still able make predictions based on the given—incomplete—data. This feature becomes particularly important when the generation approach is integrated into a bi-directional dialogue system in which the user's gesture data would be analyzed on the fly to determine possible communicative reactions. Given the problems with current gesture recognition systems, a method which is able to deal with incomplete data is preferable.
- The same network can be used to calculate the likely consequences of causal node states (causal inference), as well as to diagnose the likely causes of a collection of dependent node values (diagnostic inference). In other words, either the gestural behavior of an agent can be generated, e.g., by propagating evidence about an object's properties. Or, given a particular gesture, the features of a referent object might be inferred from a user's gesture. This is of relevance for possible application of the model in a dialogue system where understanding gestures is also important.

Nevertheless, probabilistic approaches are at their limits for their application in the current gesture generation context. As stated in Section 5.1, particularly gesture form features such as hand orientation or movement features are constrained by several factors: representation technique-specific patterns come together with referent characteristics, inter-individual differences and other contextual constraints, respectively. This is why it is reasonable to apply an extension of Bayesian networks: BDNs are able to deal with rule-based decision making as requested by requirement (3). The additional rule-based part of the model also allows to employ the generally feature-based modeling method—requirement (4)—with representation technique specific patterns to ensure that variant gesture features are combined in a sensible way.

The formalism has proven itself in the simulation of human behavior; e.g., Yu and Terzopoulos (2007) used decision networks to simulate social interactions between pedestrians in urban settings.

Are there alternatives to BDNs? The application of classical machine learning techniques like artificial neural networks, inductive logical programming, or case-based reasoning are problematic, in particular, due to the sparse data problem. Other disadvantages of those approaches, in contrast to BDNs, are due to the lack of interpretability (artificial neural networks), or the possibility to combine the data-driven method with a priori knowledge or rule-based models as it is neccessary here (cf. Wittig, 2002).

There is, however, one method that is closely related to BDNs: decision trees (Quinlan, 1986) should be considered as an alternative. Decision trees are a classical way of representing decision problems. They explicitly represent all possible sequences of decisions and observations in the model. Just as in BDNs there are three types of nodes: chance nodes, decision nodes and utility nodes. Trees can be specified by experts or learned from data employing algorithms like CHAID, CHART, ID3, or C4.5. The main drawback of decision trees is that they grow exponentially with the number of variables. That is, large trees are required even to represent quite simple decision problems (cf. Jensen and Nielsen, 2007). BDNs provide a much more compact way to represent decision problems. Another disadvantage in comparison to BDNs is related to the problem of missing data: given a complete decision tree, decision making is impeded when some information is missing.

To conclude, BDNs are advantagous for a number of reasons given the data situation and further requirements like interpretability or the integration of rule-based modelling, as defined in Section 5.1. Accordingly, BDNs are chosen as a highly promising modelling method for the task of gesture generation.

5.3 Bayesian Decision Networks

In the following the concept of *Bayesian Networks* and their extension in terms of *Bayesian Decision Networks* will be introduced.

5.3.1 Bayesian Networks

Bayesian Networks are directed acyclic graphs in which the nodes represent variables[2], the edges signify the existence of direct causal dependencies between the linked variables, and the strengths of these dependencies are quantified by conditional probabilities (Pearl, 1985). If the variables are discrete, this can be represented as a table[3]. A BN, therefore, is a graphical representation of the probabilistic relationships among a group of variables. The graphical structure of a Bayesian network model describes, in an understandable visual manner, which variables have direct influence on other variables under consideration.

Formally, a Bayesian network consists of the following (Jensen and Nielsen, 2007, p. 33f.):

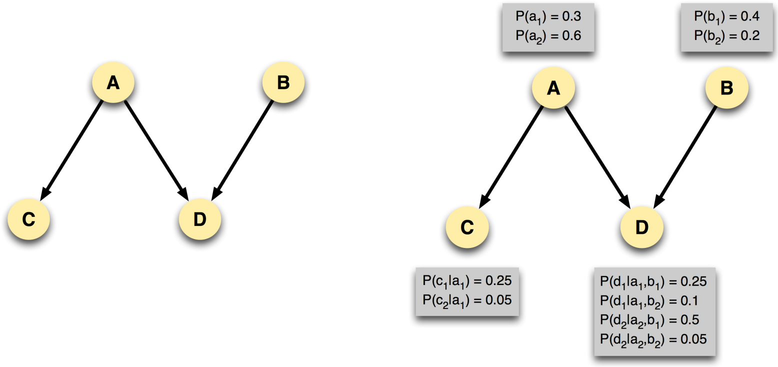
- A set of *variables* and a set of *directed edges* between the variables; each variable has a finite set of mutually exclusive states.

[2] The terms ‘variable’ and ‘node’ are used interchangeably.

[3] In this work only discrete variables will be considered.

- The variables together with the directed edges form an *acyclic directed graph*.
- To each variable A with parents $B_1, ...B_n$ a *conditional probability table* $P(A|B_1, ...B_n)$ is attached

For an example of a simple Bayesian network with discrete nodes see Figure 5.1 representing the situation that a variable C is depending on variable A, and variable D being dependent on two variables, namely A and B. Figure 5.1(a) displays the structure of these interrelations, whereas Figure 5.1(b) displays a full Bayesian network with marginal probability distributions for the nodes A and B which do not have any parents, and CPTs for nodes C and D which quantify the dependencies between the nodes.



(a) The structure of a Bayesian network: a directed acyclic graphs in which the nodes represent variables and the arcs signify the existence of direct causal dependencies between the linked variables.

(b) A Bayesian network in which the strengths of dependencies are quantified by marginal/conditional probabilities.

Figure 5.1: Bayesian network examples.

Inference The basis for the causal and inferential probabilistic modeling in Bayesian networks is *Bayes theorem* which provides a mathematical representation of how the conditional probability of event A given B is related to the converse conditional probability of B given A. In the case of discrete probability distributions of data, Bayes' theorem relates the conditional and marginal probabilities of events A and B, provided that the probability of B does not equal zero:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{5.1}$$

- P(A) is the prior probability of A. It is ‘prior’ in the sense that it does not take into account any information about B; however, the event B need not occur after event A.
- P(A|B) is the conditional probability of A given B.
- P(B|A) is the conditional probability of B given A.
- P(B) is the prior or marginal probability of B, and acts as a normalizing constant.

5.3.2 Building Bayesian Networks

Following the two-part design of Bayesian networks, learning Bayesian networks from a sample of data cases comprises two tasks: first, identifying an adequate structure of the DAG (*structure learning*), and second, estimating a set of corresponding parameters (*parameter estimation*).

Structure Learning

Specifying the structure of a network means to determine, for every possible edge in the network, whether to include edge in the final network and which direction to orient the edge. Even for a relatively small number of variables, however, structure specification is difficult and computationally non-trivial due to the fact that the number of possible network structures is super-exponential in the number of nodes. An exhaustive search is therefore not possible. That is, existing structure learning algorithms either solve a restricted problem (e.g., finding the best structure given a partial ordering of the variables) or employ search heuristics with either local or global search algorithms.

A major problem with which structure learning may be afflicted is the problem of sparse data. This may be due to the fact that the collection of adequate training data is expensive and time-consuming as is certainly the case for data of gesture use in context. This problem is even intensified when Bayesian networks are aimed to be employed to capture inter-individual differences, which obviously reduces the number of data cases from which a structure is learned. To deal with this problem, it is reasonable to support structure learning by bringing in a-priori knowledge. Jameson (2003) therefore differentiates between data-based and theory-based models. Accordingly, an integration of a-priori knowledge and machine learning techniques combines advantages of both alternatives. It is for instance possible that an expert provides the whole structure, or that an expert defines a partial structure which is used as the initial stage of the search in the space of all possible structures. Another sort of bias is providing expert information in order to reduce the search space. This can be done for the order of variables, stating that a variable can only be the parent of its predecessor, or, in the order, or, by making constraints on the direction of the edges. Notably, some learning algorithms explicitly require bringing in some kind of structural background information.

There are two very different approaches to learn the structures of Bayesian networks from given data: *score-based learning* and *constraint-based learning*. Both classes of structure learning algorithms will be introduced in the following with respect to their basic concepts and prevalently applied methods and algorithms. For a more profound presentation of algorithms see for instance Fast (2009).

Score-based Structure Learning The idea in score-based approaches is to define a global measure (score) which evaluates a given Bayesian network model as a function of the data. The problem is solved by searching in the space of possible Bayesian network models trying to find the network with optimal score. Concerning the scoring function, there are two popular choices. The first one, *Bayesian score*, measures the quality of a Bayesian network in terms of its posterior probability given the database (Cooper and Herskovits, 1992). It takes into account a prior probability and the marginal likelihood. The second one, the *Bayesian Information Criterion* (BIC), is a likelihood criterion which takes into account the estimated model parameters and a number of data cases (Schwarz, 1978). The BIC method has the advantage of not requiring a prior probability.

Even for a relatively small number of variables, however, the estimation of a DAG from given data is difficult and computationally non-trivial due to the super-exponential growth in the number of possible DAGs for each added node. This is why search heuristics are used with either *local* or *global* search algorithms.

The **K2** algorithm is a local, greedy search algorithm which is appropriate if the order of nodes is given (Cooper and Herskovits, 1992). Initially each node has no parents. The algorithm then adds incrementally that parent whose addition increases the score of the resulting structure most. When the addition of no single parent can increase the score, adding parents to the node is stopped. Note that, the algorithm requires a-priori knowledge in terms of a total order among variables that is used for reducing the search space of DAGs.

A prominent global search method to find adequate network structures is the **Markov Chain Monte Carlo (MCMC)** algorithm called Metropolis-Hastings (Madigan and York, 1995). The basic idea is to construct a Markov Chain whose state space is a set of DAGs. The idea is to sample from this chain for “long enough” (burn-in time) to ensure it has converged to its stationary distribution. Notably, the algorithm is non-deterministic; the chain can be run from multiple starting points with different values for the burn-in time whereby convergence is an open problem.

Constraint-based Structure Learning Constraint-based algorithms estimate from the data whether certain conditional independencies between the variables hold. The conditional independence constraints are propagated through the graph and the networks that are inconsistent with them are eliminated from further consideration. A sound strategy for performing conditional independence tests ultimately retains only the statistically equivalent networks consistent with the tests.

Constraint-based algorithms operate in two independent phases. The first phase, called *constraint identification*, uses a series of conditional hypothesis tests to identify an undirected skeleton, indicating the location, but not orientation of edges appearing in the final model. The second phase, called *edge orientation*, merges the learned independence constraints into a fully directed Bayesian network model.

Constraint-based skeleton algorithms employ conditional independence tests to determine which links to include in the skeleton. Classical hypothesis tests for categorical data typically utilize either the χ^2 or G^2 statistics. Both kinds of statistical independence tests are computed from a contingency table containing counts of variable values occurring in the data. The null hypothesis assumes the variables under consideration are independent. The alternative hypothesis is accepted when the probability of the observed statistic under the null hypothesis is below a specified significance threshold. Typically three levels of significance are utilized, the $p<0.05$ level, the $p<0.01$ level, and the $p<0.001$ level.

Given a particular independence test, each constraint identification algorithm applies the tests in a particular order. The predominant type of skeleton heuristics are local algorithms that consider each test independently of other decisions, whereby Fast Adjacency Search (FAS) algorithm is the most widely used ordering algorithm (Spirtes et al., 2000). The FAS algorithm operates in a breadth-first manner considering all pairwise tests followed by all tests conditioned on a single variable and so on until no more tests can be run. Alternatively, neighborhood models are another type of ordering heuristic that considers intermediate path information when determining whether to add an edge between the two variables under consideration. If a possible conditioning variable does not lie on a path between the two variables being tested then it should not be used to prove conditional independence based on the definition of d-separation in graphical models.

One of the most popular constraint-based algorithms is the **PC** algorithm (Spirtes and Glymour, 1991). It starts from a complete, undirected graph and recursively deletes edges on the basis of conditional independence decisions. Statistical tests for conditional independence are then performed for all pairs of variables using likelihood test statistic G^2 (Spirtes et al., 2000). An undirected link is added between each pair of variables for which no conditional independencies were found. Colliders are then identified, ensuring that no directed cycles occur. Next, directions are enforced for those links whose direction can be derived from the conditional independencies found and the colliders identified. One important thing to note about the PC algorithm is that, in general, it will not be able to derive the direction of all the links from data, and thus some links will be directed randomly. This means that the learned structure should be inspected, and if any links seem counterintuitive (e.g., gesture features causing object properties, instead of the other way around), one might consider going back and insert a constraint specifying the direction of the link. Correctness of the PC algorithm has been proven under the assumption of infinite data sets (Madsen et al., 2005).

Another popular constraint-based method is the **NPC** algorithm (Steck and Tresp, 1999) which is particularly useful when structure learning is based on a limited data set. The NPC algorithm is an extension of the PC algorithm which introduces the notion of a neighborhood model, the *Necessary Path Condition*: if, at some point, a conditional dependence is established, there must be a connecting path explaining this dependency. The NPC condition is necessary for the existence of a perfect map of the conditional dependence and independence statements derived by statistical tests. Thus, in order for an independence statement to be valid, a number of links are required to be present in the graph. Instead of randomly determining the directionality of the links that cannot be determined automatically from the data, the NPC algorithm relies on a-priori knowledge to determine the directionality of such links.

Score-based vs. constraint-based structure learning Both classes of algorithms were designed with a particular purpose in mind. Search-and-score techniques are generally very flexible and find high-likelihood structures. However, they do not enforce conditional independence relationships and often do not accurately reproduce the generating structure (Abellán et al., 2006; Teyssier and Koller, 2005). On the contrary, by constraining the space with conditional independence relations, constraint-based techniques are more efficient and more accuratee in recovering the structure of the generating distribution, but often do not achieve comparable likelihoods to search-and-score techniques (Tsamardinos et al., 2006).

In general, constraint-based algorithms show some advantages compared to score-based algorithms. First, score-based approaches consider a *global* measure for the entire network. Constraint-based methods, in contrast, remove all those edges from a network for which a conditional independence statement can be derived from the data. They do not take into account the structure of the network as a whole and can therefore be considered as *local*. That is, since knowledge discovery is a primary motivation of employing structure learning of Bayesian networks in the context of this thesis, constraint-based methods are to be preferred (Steck and Tresp, 1999). Second, constraint-based methods allow to vary the significance level which is used by the conditional independence tests. This way, it is possible to judge the *strength* of dependencies among variables. Third, constraint-based algorithms are better suited for learning accurate structure of Bayesian networks than search-and-score algorithms, which typically use a penalized likelihood score to choose model structures. Since training data are limited, the structure of the true model often contains more parameters than are supported by the data when optimizing for penalized likelihood. Optimizing for constraint satisfaction makes it possible to find structures with a large number of parameters, as the size of the model is not part of the model selection criterion (Fast, 2009). Finally, algorithms based on constraints have been proven sound in the sample limit (Spirtes et al., 2000).

Parameter Estimation

Parameter estimation is the second phase in learning Bayesian networks. A very prominent and prevalent method for learning the parameters of the network structure is the Expectation Maximization (EM) algorithm (Lauritzen, 1995). The algorithm searches for a set of maximum likelihood hypothesis (i.e. parameters) by repeatedly estimating values that are not present in the data. To do so, the EM algorithm is given initial values for parameters, and then iterates between an Expectation step and a Maximization step until successive parameter values are stable. During the Expectation step, the algorithm calculates the unobserved values of the data by using the current parameters. During the Maximization step, the algorithm re-calculates the maximum likelihood parameters by using the data set completed with the expected value.

Within the scope of this thesis, parameter learning is of lower importance in contrast to structure learning. Also, in the literature it is widely accepted that although having accurate parameters is important, they are completely useless if the structure is of bad quality. Druzdzel et al. (2000), for instance, stated that the graphical structure of a network is its most important part, as it reflects the independence and relevance relationships between the variables concerned.

5.3.3 Extending Bayesian Networks

The formalism of *Bayesian Decision Networks* (BDNs)—also termed *Influence Diagrams*—was introduced by Howard and Matheson (2005) as a “new form of description [...] that is at once both a formal description of the problem that can be treated by computers and a representation easily understood by people in all walks of life and degrees of technical proficiency.” The power of this representation is further due to the fact that it covers both deterministic as well as probabilistic cases.

A BDN is a directed acyclic graph which encodes three types of nodes:

- **Chance nodes**
  represent random variables each of which having a conditional probability table specifying the probability of the variable having a particular value given a combination of values of its parent nodes.
- **Decision nodes**
  represent decisions to be made. They can have both chance nodes and other decision nodes as parents indicating that the decision has to be made at a particular point in time, i.e., when all necessary evidence is available.

- **Utility nodes**
  represent a utility function. They have associated utility tables giving a value for each possible instantiation of its parent nodes.

The links between any of the nodes explicitly represent the relations between causes and effects, i.e., relationships among variables are encoded in the network. Edges between node pairs indicate influences of different types. Edges leading into a chance node show those variables on which the probability assignment of the chance node variable will be conditioned. These edges and the chance nodes constitute the 'Bayesian network part' of BDNs. Edges leading into a decision node show which variables will be known by the decision maker at the time that the decision is made. And edges leading into a utility node indicate onto which nodes a utility function is calculated.

See Figure 5.2 for a simple example of a BDN. It is based on the Bayesian network from Figure 5.1(a) and extended with a decision node and a utility node. The decision node takes the values of chances nodes C and D into account. The utility function is defined on the basis of the variable E.

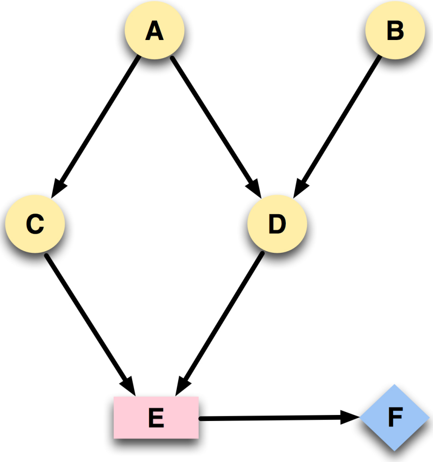


Figure 5.2: Example of a Bayesian Decision Network with chance nodes (drawn as ovals), a decision node (drawn as a rectangle), and a utility node (drawn as a diamond).

5.4 Using Bayesian Decision Networks for Iconic Gesture Generation

The process of constructing a BDN can be sub-divided into the following five steps based on the four step process of Bayesian network construction (cf. Wittig, 2002):

1. **Specification of variables**
   First of all, the variables in the domain have to be fixed. Input and outcome variables have to be precisely specified including their values. For the construction of BDNs outcome variables further have to be divided into chance variables, decision variables and utility variables.

2. **Structure specification**
   Learning the structure of a networks means to determine for every possible edge in the network, whether to include that edge in the final network and which direction to orient the edge.

3. **Parameter specification**
   Parameter estimation is concerned with learning the parameters of the network structure which represent the strengths of dependencies quantified by conditional probabilities.

4. **Rule definition**
   In this step, the Bayesian network build so far is enriched with decision and utility nodes. Rules and functions for these node types have to be defined.

5. **Application**
   Subsequent to the previous steps which were concerned with building a BDN, the model is applied in an application scenario. It has be ensured that the network is provided with all necessary kinds of evidential information.

5.4.1 Specification of Variables

The first step in building the model is to identify *input* and *outcome* variables. Outcome variables in the model are all those features of a gesture which are to be determined in the decision-making process. Factors which potentially contribute to these choices are considered as input variables. Variables and values in the model correspond to those covered by the coding scheme presented in Section 4.1.3. The set of values per variable has occasionally been narrowed to combined categories or to frequently observed values, i.e., values that occurred only rarely were excluded. For instance, in the case of handshapes ASL-B-spread, ASL-B-loose etc. are consolidated into one category ASL-B, or, in the case of representation techniques, only the five most frequent values were considered. Additionally, the previous gesture (features) are also taken into consideration. See Table 5.1 for an overview of variables and values.

A further sub-division is made for outcome variables into *chance* variables, *decision* variables and *utility* variables. The latter are disregarded at the moment, since the current conceptualization focuses on the BDN characteristics of combining determin-

Table 5.1: Variables in the generation model, specified by their values and their type.

	Variable	Values	Type
Gesture	Gesture Occurrence (G)	true, false	Outcome, Chance
	Representation Technique (RT)	indexing, placing, shaping, drawing, posturing	Outcome, Chance
	Handedness (H)	LH, RH, 2H	Outcome, Chance
	Handshape (HS)	ASL-B, ASL-C, ASL-G, ASL-O, ASL-5	Outcome, Chance
	Palm Orientation (PO)	PAB, PTB, PTL, PTR, PUP, PDN + combinations and sequences	Outcome, Decision
	BoH Orientation (FO)	BAB, BTB, BTL, BTR, BUP, BDN + combinations and sequences	Outcome, Decision
	Movement Type (MT)	linear, curved + sequences	Outcome, Decision
	Movement Direction (MD)	MF, MB, ML, MR, MU, MD	Outcome, Decision
Discourse and Verbal Context	Thematization (T)	theme, rheme	Input
	Information State (IS)	private, shared	Input
	Communicative Goal (CG)	lmIntro, lmDescrProp, lmDescrConstr, lmDescrPos	Input
	Noun Phrase Type (NP)	Tags according to the STTS tagset (Table 4.3)	Input
Referent Features	Subparts (CN)	1 or more, none	Input
	Symmetry (S)	sym, none	Input
	Main Axis (MA)	x-axis, y-axis, z-axis, none	Input
	Position (P)	left, middle, right	Input
	Shape Property (SP)	longish, cubic, squared, arc-shaped, spherical, round, etc.	Input
Previous Gesture	Gesture Occurrence (LG)	true, false	Input
	Representation Technique (LRT)	indexing, placing, shaping, drawing, posturing	Input
	Handedness (LH)	LH, RH, 2H	Input
	Handshape (LHS)	ASL-B, ASL-C, ASL-G, ASL-O, ASL-5	Input

istic and probabilistic decision making[4]. Note that, the distinction between chance variables and decision variables does not prevent detecting common patterns across individuals. These rather become apparent in identical or similar network structures.

The choice whether an outcome variable in the model is of one or the other kind is determined in accordance with the requirements of data-based and rule-based decision making as devised in Section 5.1. On the one hand, to capture inter-individual differences, it is reasonable to determine the values of outcome variables probabilistically. On the other hand, it is reasonable to determine outcome values in a deterministic way due to the sparse data problem: as some outcome variables have a large set of values (e.g., palm and BoH orientation or movement direction) which are, in addition, biased by several factors (e.g., representation technique-specific patterns and/or characteristic features of the referent), there are not enough data cases to learn the network structure and probability distributions from. This is reasonable, as the empirical analysis in Section 4.2) revealed representation technique-specific patterns and/or referent feature dependencies.

So each outcome variable has to be defined as being either a chance variable or a decision variable. Two decisions in the formation of a gesture which are inherently not subject to any representation technique-specific patterns, are the choice whether to gesture (or not) and the choice of representation technique. As these two variables are subject to significant inter-individual differences (cf. Sections 4.2.1 and 4.2.2, these two are chosen to be chance variables. The variables ‘handedness’ and ‘handshape’ are also significantly biased by inter-individual differences as shown in Section 4.2.3. These variables, therefore, are assigned to the set of chance variables in the model as well.

Gesture features having a relatively large set of values are palm and BoH orientations as well as movement direction. These variables are accordingly good candidates for decision variables. This choice receives further support by the fact that these gesture features are not only subject to representation technique-specific characteristics, but also make up the semantics of shaping, drawing and posturing gestures. In other words, these variables are strongly correlated with referent features. The second argument also holds for the variable ‘movement type’ which is therefore also modeled as a decision variable although it has only relatively few values.

In sum, the result of the first modeling step—identifying variables—is concluded with a collection of variables representing both gesture features and modulating factors. The variables are of three different types: input variables, chance outcome variables, and decision outcome variables. A visualization of the variables in the model is given in Figure 5.3.

[4] Utility variables to state configurations of the network can still be integrated into the network at any time

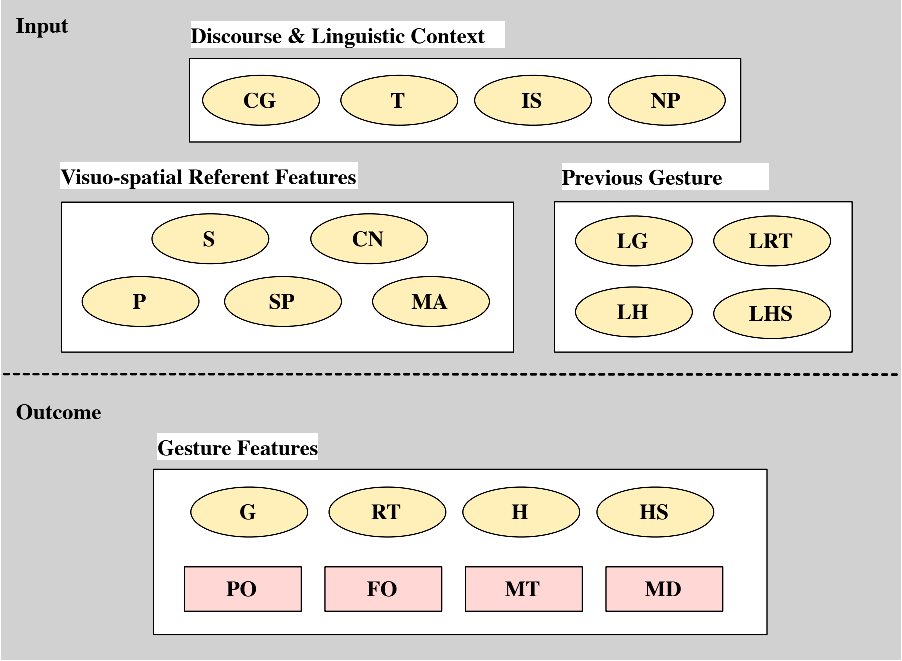


Figure 5.3: Variables to be considered in the generation network: chance variables are drawn as ovals, decision variables are drawn as rectangles.

5.4.2 Network Structure Specification

In the next step, the structure of the network has to be determined for the probabilistic part of the network (Bayesian network), i.e., for the outcome chance variables and their ingoing edges on which the probability assignment of the outcome node variable will be conditioned.

Choice of learning method To deal with the given problem of sparse data, a combination of data-based and theory-based structure specification is appropriate. That is, the application of learning algorithms is constrained by the following assumptions. First, the order of decisions is given so that the choice to produce a gesture or not is taken first, followed by the decision which representation technique to use. The choices for handedness and handshape are made thereafter. Second, the direction of edges is specified in a way that variables of gesturing behavior are influenced by other variables, i.e., it is assumed for instance that properties of the gesture referent have an impact on the gesture, and not the other way around. And third, no dependencies are assumed to exist among input variables characterizing the given situation in terms of referent features, discourse context, and the previously performed gesture. It is accepted that these assumptions might decrease the quality of the model in terms of prediction accuracy compared to empirical data, as the aim here is to elucidate by which factors gesture use is influenced and not, e.g., how any modulating factors are interrelated among each other.

Concerning the choice of structure learning algorithms, constraint-based algorithms have some advantages over score-based algorithms (see Section 5.3.2), providing support for algorithms like PC or NPC. It should be evaluated, nevertheless, if constraint-based learning actually outperforms score-based methods in the context of gesture generation. For this reason, different learning algorithms will be comparatively evaluated with regard to their prediction accuracy against the original SaGA data (Section 7).

Choice of database The result of structure specification is not only depending on the chosen learning algorithm and its (potential) combination with a-priori knowledge. Rather, the data from which the structure is learned is also decisive. In general, there are two possibilities of constructing a data set for learning the probabilistic part of the gesture generation network: either, building a speaker-specific, *individual* model from the data of one particular speaker, or learning the structure from the *combined* data of several speakers. In general, the problem of overfitting tends to occur when individualized models are learned from the (sparse) data of one individual speaker, since sparse data are often not able to represent all relevant relationships. In addition, it is sometimes the case that learning algorithms detect relations that are, on closer inspection of a larger data base, not typical for the speaker. Similarly, when considering models that are learned from the combined data of many speakers, it is possible that these models over-represent particular speakers or relationships. This is particularly the case when models are learned from few speakers or when interrelations are heterogenous in the data. That is, there are problems with both alternatives. And this is why both variants should be evaluated with regard to their prediction accuracy. A comparison of individual vs. combined networks further provides the chance to detect differences and commonalities between both kinds of models and data sets. A visualization of the model developed so far is given in Figure 5.4.

5.4.3 Parameter Estimation

To learn the parameters of the network structure, the prevalent EM algorithm is employed. As the focus of this thesis is on structure learning, variation of parameter learning was not performed. The fact that parameters are always learned in the same way provides the advantage that accuracy results are more readily comparable with regard to structure learning.

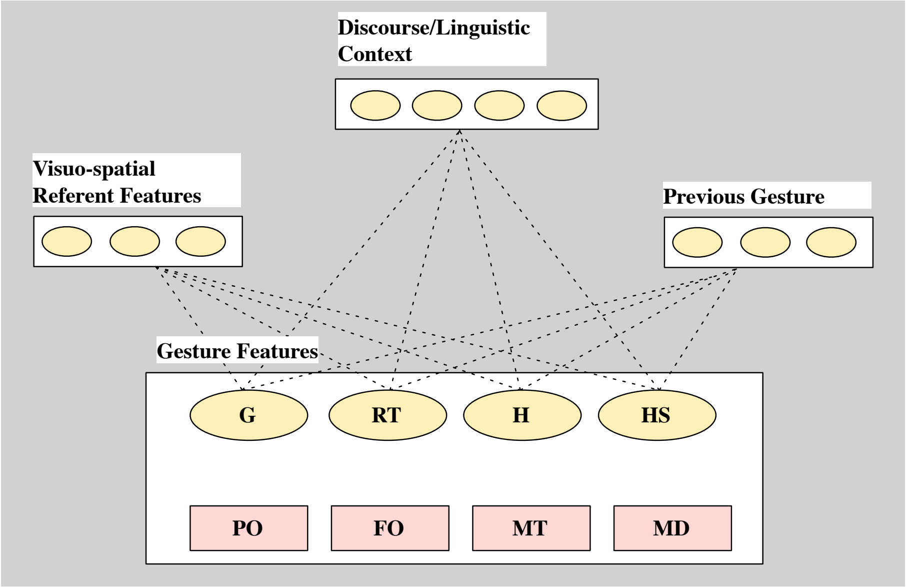


Figure 5.4: Schema of the generation network in which outcome chance nodes are potentially connected with input nodes, depending on the data set from which the network structure is learned.

5.4.4 Rule Definition

To extend the Bayesian networks which were built in the previous steps, decision nodes need to be specified. In order to make the definition of appropriate rules empirically sound, it is based on the corpus-based analysis of gestural representation techniques (see Section 4.2.3). Table 5.2 shows how gestural form features depend on other gesture features due to technique-specific patterns and/or referent characteristics.

The concrete representation technique-specific form-meaning mappings are outlined in the following. They are based on basic patterns found in the SaGA corpus. Note that, these mappings are not meant to be able to account for *every* single gesture occurrence. Speakers deviate from these patterns either due to the natural fuzziness of human gesturing, or by employing gestural patterns that are not frequent enough in the empirical data, e.g., due to the application domain or limited stimulus materials.

Indexing Indexing gestures were found to be relatively strongly shaped by representation technique-specific patterns.

- **Palm Orientation** As handshape and palm orientation were shown to be positively correlated, an adequate orientation of the palm is picked depending on the chosen handshape: ‘PDN’ for pointed indexing gestures (handshape ‘ASL-G’),

Table 5.2: Dependence of gestural form features on other gesture features due to technique-specific patterns and/or referent characteristics.

	Indexing	Placing	Shaping	Drawing	Posturing
Palm Orientation	Handedness, Handshape	Handedness, Handshape	Handedness, ShapeProperty, MainAxis	–	Handshape, Handedness, ShapeProperty, MainAxis
BoH Orientation	Position	Handshape	ShapeProperty, Main Axis	–	Handshape, Handedness, ShapeProperty, MainAxis
Movement Type	–	–	ShapeProperty	ShapeProperty	–
Movement Direction	–	–	Handedness, ShapeProperty, MainAxis	Handedness, ShapeProperty, MainAxis	–

and sideways palm orientation ('PTL' for right-handed gestures, 'PTR' for left-handed gestures) for flat hand indexing (handshape 'ASL-B').

- **BoH Orientation** The most frequent BoH orientation in indexing gestures was away from the speaker's body. This was, however, found to be slightly influenced by the referent position. Thus, the referent position is taken into account when choosing the BoH orientation in indexing gestures ('BAB/BTR' for referents at the right, 'BAB/BTL' for referent at the left)
- **Movement Features** Movement features in indexing gestures are generally set to be constant, since indexing gestures were found to be characterized by static strokes.

Placing Similar to indexing gestures, the form of placing gestures was also found to be strongly shaped by technique-specific patterns.

- **Palm Orientation** According to the chosen handshape, palm orientation was found to be different in the data. Therefore, in placing gestures with handshape ASL-B, the palm is assigned to be oriented sideways ('PTL' for right-handed gestures, 'PTR' for left-handed gestures). In placing gestures with handshape ASL-C, the palm is set to be oriented away from the speaker's body, and in placing gestures with handshape ASL-bent-5, the palm is assigned to be oriented downwards.
- **BoH Orientation** 'BAB' was empirically found to be the most frequent category and is, accordingly, set for all placing gestures with exception of c-handed placings for which 'BUP' was, by trend, found to be preferred.

- **Movement Features** Due to the fact that placing gestures were characterized by static strokes, movement features in placing gestures are generally assigned to be constant.

Shaping These gesture depict the shape of referents and are, accordingly, highly sensitive to characteristics of the referent.

- **Palm/BoH Orientation** For palm and BoH orientation there was neither a category present which was predominantly used due to technique specificity, nor was there any statistically significant sensitivity of palm orientation with regard to referent features. Therefore, rules for the choice of palm orientation were defined heuristically on the basis of corpus examples and depending on the referent's shape properties, main axis and the already chosen handedness.
- **Movement Type** Shaping gestures were shown to be characterized by a shape property specific wrist movement type. Accordingly, to depict the shape properties 'round' and 'arched' the movement type is set to be curved, whereas for longish, cubic and squared objects it is assigned to be linear.
- **Movement Direction** Rules for movement directions could not be based on statistically significant results due to low sample size and were rather defined heuristically on the basis of examples in the corpus: the foremost influential factor affecting the movement trajectory is the referent's shape, e.g., a circular trajectory for 'round' ('MR>MD>ML>MU') or a straight line for 'longish'. Concerning the latter, the referents' main axis has to be further considered so that high objects (main axis: y) are depicted by an upwards movement while rather broad objects (main axis: x) are set to be depicted by a sideways movement. Finally, handedness is considered, as for two-handed gestures the trajectory has to be different than for one-handed gestures (only half a circle instead of a full circle, for example).

Drawing Drawing gestures are shaped by both technique-specific patterns as well as referent characteristics with regard to movement features.

- **Palm/BoH Orientation** In line with the empirical data, palm and BoH orientation are assumed to be constant: palm oriented downwards, the BoH oriented away from the speaker's body.
- **Movement Type** Movement type was found to be dependent on the shape properties of the object to be depicted. Therefore, for round and arc-shaped objects, the type of movement is assigned to be curved, for longish, cubic and squared objects it is set to be linear.

- **Movement Direction** Similar to shaping gestures, the movement direction in drawing gestures is defined heuristically taking the referent's shape properties, its main axis as well as the chosen handedness into account.

Posturing Posturing gestures depict the shape of objects in a static way. Accordingly, non-movements have to be selected depending on referent characteristics.

- **Palm/BoH Orientation** Due to the fact that no statistically significant correlations could be found in the data for the choice of palm orientation and BoH orientation, these values are defined heuristically based on examples and tendencies in the data taking the previously chosen handedness and handshape into account, as well as the referent's shape and main axis.
- **Movement Features** As the strokes in posturing gestures are generally static, movement features do not have to be specified for these gestures.

With the definition node rules specified as described above, nodes in the gesture generation network appear to be connected as displayed in the network schema in Figure 5.5.

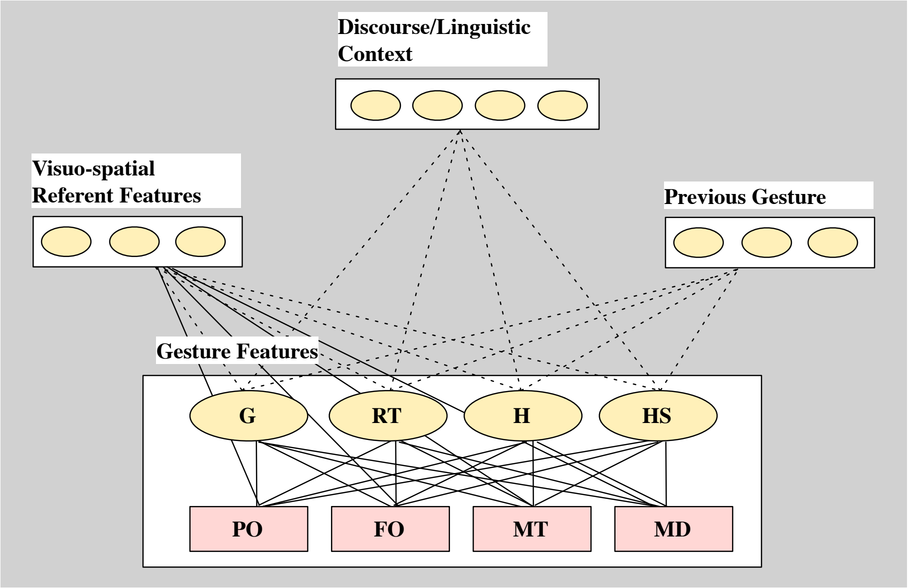


Figure 5.5: Schema of a gesture generation network in which gesture production choices are made either probabilistically (dotted lines, learned from corpus data) or rule-based (solid lines, defined in a set of if-then rules).

5.5 Integration into Overall Generation Architecture

Finally, after completing the previous steps of network construction, the BDN is to be employed in an application system. To make use of the generation network in a simulation account for virtual agents, the BDN is to be integrated into an overall speech and gesture production framework which was developed in the context of the sub-project B1 of CRC 673 at Bielefeld University (Kopp et al., 2008). Of particular importance is that the network is embedded in a way such that the input nodes of the gesture network are supplied with all necessary information. The requirements for an overall speech and gesture production architecture are summarized in the following:

- One key issue for both speech and gesture production is the representation of content to be uttered. Existing psycholinguistic production models agree on the fact that speech and gesture are derived from two kinds of representation (spatial and propositional). This is in line with Paivio's—empirically based—Dual Coding Theory (Paivio, 1986) which is characterized by two functionally independent systems, verbal memory and image memory, with associative links within each memory and possible referential links across the two systems. Imagery code is assumed to primarily represent shape and spatial and spatio-temporal relationships (rather than purely visual properties such as color or brightness). Verbal code was originally taken to be some form of inner speech in a natural language, which gave early rise to criticism and alternative suggestions to construe the two codes as imagery and conceptual ('mentalese') rather than imagery and English (Kieras, 1978). Along this line, elements of the imagery code represent what they are images of; elements of the verbal code represent what words mean.

  To realize such a dual coding account, content information has to be represented in these two formats. Further, an interface between the two kinds of representations is necessary to enable the coordination of modalities.

- **Cross-modal interaction** must not be limited to the level of content representation and content planning. To account for mutual influences of the two modalities, as presented in Section 2.2.4, a multimodal production model has to provide according mechanisms for mutual access and sensitivity of modality-specific processes. An example for cross-modal influences is realized in the GNetIc account which relies on information about the linguistic context in terms of thematization (thema, rhema). Cross-modal interaction is also needed for the coordination that enables temporal and semantic synchrony of the verbal part of the utterance and its accompanying gesture(s). At the lower levels of phonation and motor control interaction may account for compensation of potentially occuring timing lags during execution (e.g., pre-stroke holds in gesture).

- The Interface model (Section 2.3) proposed a **feedback mechanism**: the message generator receives information from the formulator whether a proposition is readily verbalizable or not. This accounts for evidence that gestures compensate for problems of verbal encoding (Bavelas et al., 2002). This idea should be implemented in a production model and could be extended such that not only encoding problems are effectual, but also the existence of primed linguistic constructions. Moreover, the flow of information should take place both in the speech as well as the gesture pathway to account for alignment of motor representations.
- To provide the GNetIc model with necessary information about the **discourse context**, a discourse record has to be available. In particular it is necessary, for both speech and gesture production processes, to know whether some information has already been uttered before or not. In other words, the information state has to be available to account for the mechanism of grounding. Speech generation has to differentiate (at least) between definite and indefinite articles in this regard, while for gesture generation the information state is one of the input variables. In addition, the evolving discourse context must comprise information not only about which meanings have been communicated, but also the words and gestures that have been employed for this.

In conclusion, a process model for the generation of multimodal utterances should be characterized by a high degree of interactivity, both across modalities as well as between two levels of planning. A speech and gesture production model that fulfills these requirements is displayed in Figure 5.6.

Computational approaches to producing multimodal behavior with artificial agents usually conceive of the generation problem in terms of three consecutive tasks (cf. Reiter and Dale, 2000): (1) figuring out what to convey (content planning), (2) figuring out how to encode it (micro-planning), and (3) realizing the behaviors (surface realization). Following this tripartition, the model is also inspired by, but also extends and substantiates the psycholinguistic models of speech and gesture production model in several ways (Kita and Özyürek, 2003; de Ruiter, 2000). Following Kita and Özyürek (2003), there are four processing modules involved in content planning and micro-planning of speech and gestures: *Image Generator, Preverbal Message Generator, Speech Formulator*, and *Gesture Formulator*. The idea of two functionally separable modules, one for activating and selecting features of visuo-spatial imagery (the Image Generator) and one for turning these features into gestural form (Gesture Formulator), is adopted (de Ruiter, 2000; Kopp et al., 2004). In addition, two dedicated modules (*Motor Control* and *Phonation*) are concerned with the realization of synchronized speech and gesture movements with the virtual human Max (Kopp and Wachsmuth, 2004). Further components are a discourse model and distinct long-term memories for imagery and propositional knowledge.

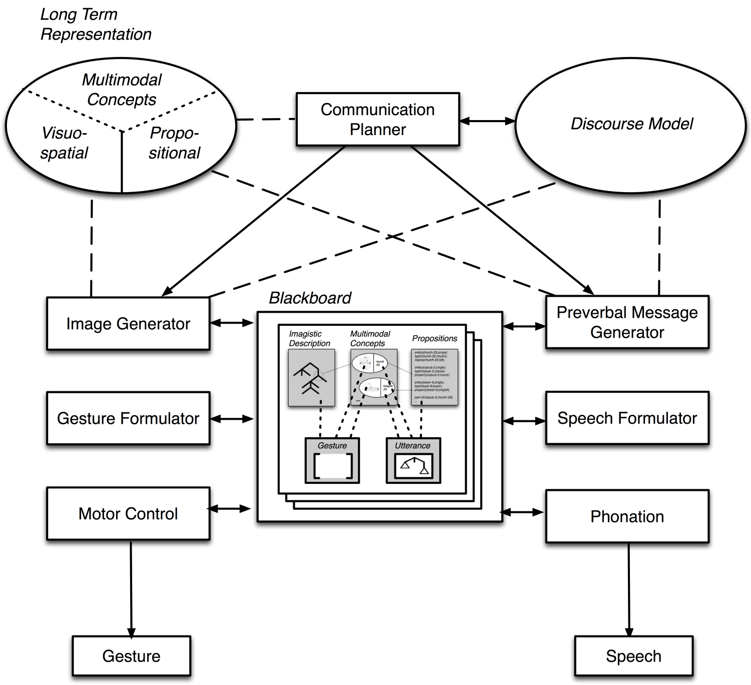


Figure 5.6: Overview of the architecture of the speech and gesture production model.

Content Representation Going beyond previous approaches in computational modeling of speech and gesture production, which were solely based on a propositional representation (Section 3), it is reasonable to adopt a dual coding perspective for content representation which differentiates between mental imagery and linguistic or conceptual knowledge. With regard to being either the basis for gesture generation or for speech generation, imagistic knowledge is assumed to primarily represent shape and spatial or spatio-temporal relationships (rather than purely visual properties such as color or brightness), and the verbal code is taken to conceptually represent what words mean (Kieras, 1978).

To interface between these two distinct representational formats, they are appropriately connected by an additional layer. The level of content representation underlying the production of multimodal utterances is, therefore, assumed to consist of three parts, (1) imagistic knowledge, (2) propositional knowledge, and (3) pairings of imagistic knowledge and propositions (see Figure 5.7). These three structures may be organized as a multimodal working memory, in terms of a globally accessible, structured blackboard (Hayes-Roth, 1985) on which all modules involved in the generation process operate concurrently.

A suitable model for the representation of **imagistic knowledge** are *Imagistic*

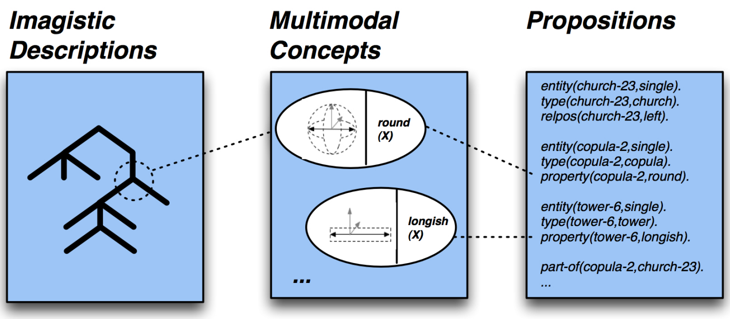


Figure 5.7: The content representational layer underlying the production of multimodal behavior consists of three parts, (1) the imagistic description, (2) propositional knowledge, and (3) pairings of imagistic knowledge and propositions.

Description Trees (IDTs, Section 3.1) that are required by the input nodes of GNetIc networks. It is easily possible to extract referent features such as object position, symmetry, main axis, and number of subparts from an IDT representation.

The second part of working memory is a **propositional knowledge base** consisting of logical formulae based on a formal ontology of domain knowledge (Witte and Weisemann, 2008). A representation of knowledge to be drawn upon by the speech formulation processes needs to be designed to fit the needs and affordances of natural language. As discussed above, this requires a proper representation of spatial knowledge as well as conceptual background knowledge about the considered entities. As is common in computational approaches to language semantics, we employ a propositional format for this, i.e., knowledge is encoded in terms of propositions over symbols that represent objects and relations, according to a given ontology. Due to the focus on object descriptions, the spatial knowledge to be captured pertains to objects, their geometrical properties, and the relations between them. The representation system thus consists of logical formulae based on a formal ontology.

To cover all relevant aspects of domain knowledge the ontology was built on the basis of three sources: (1) visual characteristics of the entities, (2) object descriptions given by the participants of the SaGA study, and (3) knowledge encoded in the IDT representation, e.g. in terms of part-whole relationships. World knowledge to be represented by the ontology consists of entities, attributes and relations:

- **Entities** are objects of the application domain with independent existence such as houses, streets, etc.
- **Attributes** are properties of entities, like proper name, color, quality, size, shape etc. Attributes can be single- or multi-valued.

- Entities are connected by different types of **relational links**, such as taxonomic (is-a), partonomic (part-of), or spatial relations (on-top-of, left-of).

The third part of the content representation are a number of **multimodal concepts** which are bindings of IMDs with corresponding propositional formulations (see Figure 5.8). The imagistic parts of these multimodal concepts are characterized by underspecification, i.e. they contain more than one alternative interpretation and thus represent abstract concepts. It is this underspecification which draws the distinction between the imagistic part of the multimodal concepts and the imagistic description of concrete spatial objects. For example, the property of being ‘longish’ can be represented as an underspecified IMD in which one axis dominates the other one or two axes, as well as in terms of a logical formula (e.g. ‘longish(X)’). Such an underspecified IMD can be matched with any other IMD by means of a formal graph unification algorithm as described in Sowa (2006). Currently, multimodal concepts for dimensional adjectives (longish, round, tall, etc.), stereotyped object shapes (box, circle, etc.), and basic spatial relations (right-of, side-by-side, above, etc.) are predefined as part of long-term memory. Matching of IMD-structures is realized by the unification procedure which also determines the necessary geometrical transformations for one IMD.

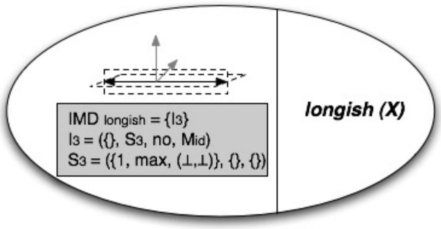


Figure 5.8: Multimodal concept corresponding to the property ‘longish’.

Image and Preverbal Message Generation Image generation and preverbal message generation are content planning processes which perform information selection and perspective assignment. This first step in translating a thought into speech and gesture is similar for speech and gesture production.

The production of a chunk of multimodal object description is assumed to start upon the arrival of a message from the Communication Planner, containing a new communicative intent. Such a communicative intent comprises a specification of the intended communicative goal, such as ‘introduce’, ‘describe’, ‘describe-position’, or ‘describe-construction’ (cf. Section 4.1.3) along with the entities to be referred to or the properties to be predicated.

The **Image Generator** has to access the imagistic representations stored in long-term memory and activates the imagistic descriptions of all objects involved in the communicative goal. This may appropriately be modeled by *activation spreading*, a mechanism from cognitive psychology for the retrieval of information from memory (Anderson, 1983; Collins and Loftus, 1975). The process of activation spreading is initiated by labeling a set of source nodes with weights and then iteratively propagating that activation to other nodes linked to the source nodes, whereby the weights decrease step by step. It is further possible that weights decay, as the activation propagates through the network. Activation values above a particular threshold may generally lead to import an IMD into the respective working memory structure. Likewise, the **Preverbal Message Generator** may apply the same mechanism of spreading activation to select knowledge from propositional long-term memory which are then asserted into working memory.

For IMDs with a significantly high activation, the Image Generator has to perform spatial perspective taking. That is, it has to determine which spatial perspective to adopt towards the object(s). This is based on the fact that direction-givers usually adopt either a route (1st person) or survey (3rd person) perspective (Levinson, 1996), with frequent switches between the two (Striegnitz et al., 2007).

The Image Generator tries to map the perspective IMD onto the imagistic parts of multimodal concepts in long-term memory. If this succeeds, the corresponding multimodal concept is added to the working memory unit. Likewise, multimodal concepts are asserted when they unify with propositions selected by the Preverbal Message Generator. Of particular importance to this process of cross-modal activation is the fact that neither the original IMD nor the original propositions have to be identical with the pole of a multimodal concept to match. Instead, a *similarity value* between 0 and 1 is calculated by comparison of the two IMDs, and a multimodal symbol is either selected or not depending on a customizable threshold of similarity.

Speech Formulation The Speech Formulator monitors the unit on the blackboard and carries out sentence planning for each set of propositions posted by the Preverbal Message Generator. As in related systems (Cassell et al., 2000a; Kopp et al., 2004), the SPUD system(Stone et al., 2003) is employed, a grammar-based micro-planner using a Lexicalized Tree Adjoining Grammar (LTAG) to generate natural language sentences. SPUD delivers back information on the semantics of each linguistic constituent which is brought in accordance with the entity or property the gesture aims to depict. The derived temporal constraints then mark the onset and end of the lexical affiliate on the word level and are asserted to the blackboard unit, too.

Gesture Formulation The readily constructed GNetIc decision networks, as described above, are used directly for gesture formulation. A few pre- and post-processing steps, however, are additionally necessary to complete the mapping from representations of imagistic semantics and discourse context onto an adequate speech-accompanying iconic gesture. Figure 5.9 gives a schematic of the formulation process.

- **Assessing referent characteristics—Analysis of imagistic knowledge** The initial situation for gesture formulation is an IDT representation of the object or property to be referred to. In a pre-processing step, this representation has to be analyzed to extract those referent features that are required as initial evidence for the network: (1) whether an object can be decomposed into subparts, (2) whether it has any symmetrical axes, (3) its main axis, (4) its position in the VR stimulus, and (5) its shape properties. Further information drawn upon by the decision network concerns the discourse context. It is provided by other modules in the overall generation process and can be accessed directly from the blackboard.

- **Making production choices** All evidence available is then propagated through the network resulting in a posterior probability distribution for the values in each chance node. To make the decision which value is finally chosen for realization, there are generally two alternatives, the *probability matching* strategy and the *maximization* strategy. Probability matching is a suboptimal decision strategy in which decisions are made proportional to the given probability distribution. Thus, if the posterior probability of a gesture occurrence is found to be 70% (i.e., the probability that no gesture occurs is 30%), probability-matching strategy would predict a gesture occurrence in 70% of the choices, and no gesture occurrence in 30% of the choices. The optimal maximization strategy maximizes the number of correct predictions (Duda et al., 2001). Following this strategy in the example means to always predict a gesture occurrence as it predicts the majority category in the absence of other information.
  The suboptimal probability-matching strategy is frequently employed by human subjects in decision and classification studies. Additionally, Foster and White (2007) compared both strategies in an evaluation study of automatically generated texts. Participants judged those texts as being better written and less repetitive that were generated while choosing from among the highest-scoring realization candidates instead of taking the single highest-scoring result. In other words, human users preferred variation.
  In the context of the evaluation studies to be carried out in this thesis it is, nevertheless, reasonable to employ maximization. With regard to the accuracy of simulation results compared to the human archetype modeled in the network, employing probability matching would decrease simulation accuracy. This would consequently conceal the model's prediction accuracy in a prediction-based evaluation, and it would bring an additional factor of uncertainty into play with regard to a user-based evaluation study. When bringing the genera-

tion model to application, however, probability-matching should be taken into consideration again.

- **Checking for Similarity** Finally, the obtained gesture specification is compared with the IDT representation it originates from by making use of the fact that the BDN formalism allows for two different types of inferences: causal inferences that follow the causal interactions from cause to effect and diagnostic inferences that allow for introducing evidence for effects and infer the most likely causes of these effects. The gesture feature matrix derived from the decision network is set as evidence for the output nodes of the BDN. A diagnostic inference then yields the most likely causes, i.e., the most likely referent properties and values of discourse contextual variables. The former are used to build an (underspecified) IDT representation of what the very gesture depicts which is to be compared with the initial referent-IDT.
  The IDT representation of gesture semantics, notably, allows the agent to 'know' what her interlocutor knows. This is pertinent in terms of grounding: assuming that the interlocutor acknowledges the utterance, this is exactly what becomes common ground. The measure of similarity is further beneficial for two issues. First, it allows to avoid gesture specifications with incompatible feature values. Such an incompatibility might occur especially in posturing gestures in which the static hand configuration carries the meaning of the gesture. Since the variable 'handshape' is, however, determined probabilistically, it might happen that its value cannot be adequately used to depict the intended meaning (e.g., handshape ASL-B for a round window). If a discrepancy is detected, the decision network is requested again with the additional constraint to either plan a gesture with a different technique or to return a feature matrix with a different handshape. Second, the similarity measure can also be employed as a decision-supporting measure for the case that an over-generation and selection strategy on the level of complete gestures (not just gesture features) is needed.

5.6 Summary

This chapter conceptually developed a gesture generation network. As a first step, requirements for such a model were formulated, and Bayesian Decision Networks (BDNs) have been identified as an an attractive method to model gestural behavior:

- The BDN formalism belongs to the class of graphical models which represent complex interrelations in an understandable, visual manner. It, therefore, simulates gesture use in an understandable and **interpretable** way, accounting for one of the two main motivations behind the aim to build a model of gesture generation, namely to gain insights into the gesture production process in humans.

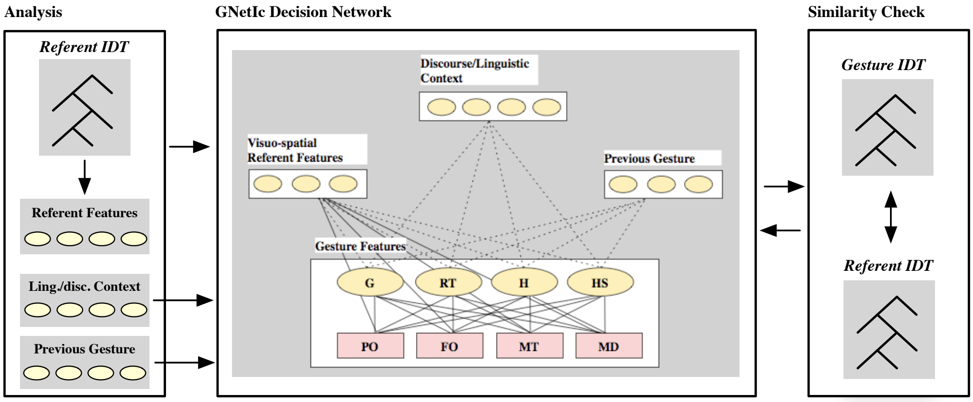


Figure 5.9: Schematic of the gesture formulation process: the referent IDT is analyzed for characterizing features required by the GNetIc model. These and other variables (linguistic/discourse context, previous gesture) are propagated through the network, resulting in a gesture specification which is compared with the referent IDT representation it originates from.

- BDNs provide a hybrid method covering both probabilistic inference problems and deterministic decision making problems. The formalism, accordingly, allows consideration of inter-individual differences in a **data-based** fashion. In addition, the BDNs are able to deal with sparse data, utilizing **rule-based** modeling.
- Gesture features can be realized as single nodes in the BDN model, providing a **feature-based** generation account. This is reasonable, since iconic gestures fall into different techniques of representation all of which are characterized by differing sets of variant and invariant features.
- Within the framework of BDNs, it is a simple matter to introduce new sources of information into the model. It is likely that the currently available set of input variables provided by the SaGA corpus is not complete at all. A BDN is easily **extensible** by introducing further variables, either annotated in the corpus, or inferred from that data.

In addition, Bayesian networks, making up the probabilistic part of BDNs, are able to deal with some problems that are likely to occur in the context of gesture generation. First, Bayesian networks are able to deal with uncertainty by making predictions about the relative likelihood of different outcomes. Second, they are able to handle incomplete data, which is of particular importance when the generation approach is integrated into a bi-directional dialogue system provided with data from gesture recognition systems. Third, also relevant for possible application in a dialogue system, the same network can be used to calculate the likely consequences of causal node states (causal inference), as well as to diagnose the likely causes of a collection of dependent node values (diagnostic inference).

The process of network construction, broken down into the following five steps (for an overview see Figure 5.10), was applied to the gesture generation problem:

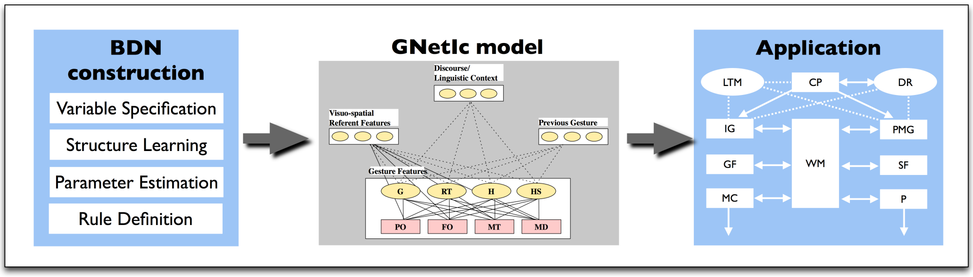


Figure 5.10: Concept of the gesture generation process: Bayesian Decision Networks are constructed in four steps combining probabilistic and data-based modeling. The resulting GNetIc model is integrated into an overall speech and gesture production architecture which automatically generates multimodal utterances.

1. **Specification of variables**
   First, the relevant variables were identified as being either input or outcome variables. The latter were further sub-divided 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').

2. **Structure learning**
   Learning network structures requires to choose an adequate learning algorithm and to decide on which data set the model is to be learned. Regarding the first issue, it has been argued that constraint-based algorithms are to be preferred for several reasons. Nevertheless, different learning methods will be evaluated regarding their prediction accuracy. Similarly, it was left for evaluation whether models should be learned from speaker-specific data or from the combined data of several speakers. In addition, some domain-specific constraints were specified to support structure learning by reducing the search space of potential networks.

3. **Parameter estimation**
   The prevalent EM algorithm is employed to learn the parameters of the network. Therefore, accuracy results are comparable with regard to structure learning.

4. **Rule definition**
   Decision nodes were specified to extend the probabilistic networks built in the previous steps. The definition of appropriate rules was based on the corpus-based analysis of gestural representation techniques. That is, depending on the very representation technique, gesture form features were defined to be subject to referent characteristics as well as other gesture form features.

5. **Application**
   Finally, after completing the previous steps of network construction, the BDN is to be applied in an overall speech and gesture production framework. The generation framework is based on key insights from psycholinguistic research of the speech and gesture production process in humans. Hallmarks of the overall speech and gesture production architecture are the following:

   - The representation of content is characterized by a dual coding perspective consisting of imagistic and propositional codes. Multimodal concepts are utilized to interface between both representational formats.
   - The production process for both speech and gestures is broken down into content planning (*Preverbal Message Generator* and *Image Generator*), microplanning (*Speech Formulator* and *Gesture Formulator*), and surface realization (*Phonation* and *Motor Control*). Further components are a discourse model and distinct long-term memories for imagery and propositional knowlege and multimodal concepts.
   - Interactivity between the modules is enabled through a central multimodal working memory, realized as a globally accessible blackboard upon which all modules operate concurrently. The overall production process thus evolves by each module's observing entries in the working memory, taking local action if necessary, and modifying existing entries or their activation or posting new entries.

   Complete GNetIc networks are accessed by the Gesture Formulator to conduct gesture planning. Input nodes of the network are provided with all necessary information about the referent, the linguistic/discourse context of the utterance to be produced, as well as the previous gestures of the speaker.

CHAPTER 6

Realization

This chapter is concerned with the realization of the previously developed concepts and methods for a virtual agent. The application scenario is based on the virtual world employed in the SaGA study (Section 4). The virtual agent Max is to be enabled to multimodally explain buildings of the virtual environment being equipped with adequate knowledge sources, i.e., communicative plans, lexicalized grammar, propositional, and imagistic knowledge about the world.

Section 6.1 explains how GNetIc models are realized, to be followed by Section 6.2, in which the implementation of the overall speech and gesture production architecture that incorporates these models is described. To illustrate the generation process, Section 6.3 details step by step how a multimodal utterance is generated. The chapter is concluded with a survey of modeling results (Section 6.4).

Details of the implementation have been published in Bergmann and Kopp (2009b,a).

6.1 Building GNetIc Models

The construction of Bayesian Decision networks for the purpose of gesture generation was realized with the HUGIN toolkit (Madsen et al., 2005), one of the oldest and best-known tools for Bayesian network construction and inference. It comes with a graphical user interface and the HUGIN Decision Engine for application development. The user interface contains a graphical editor, a compiler, and a runtime system for the construction, maintenance and usage of knowledge bases using Bayesian network technology. The Decision Engine contains all functionality related to handling and using knowledge bases in a programming environment. It is delivered with application program interfaces for major programming languages such as C, C++, Java, and .NET. The framework supports the construction of Bayesian Networks as well as Bayesian Decision Networks (called Influence Diagram Models in the HUGIN terminology).

With regard to building Bayesian networks, HUGIN supports the constraint-based structure learning algorithms PC and NPC (Section 5.3.2) and estimation of the conditional probability distributions using the EM algorithm (Section 5.3.2). It further allows the specification of domain expert knowledge in terms of structural constraints. This feature, notably, allows for the application of other structure learning algorithms and combining these with the parameter estimation and inference methods implemented in HUGIN.

Specification of Variables Nodes and values in the model correspond to those specified in Table 5.1. All are realized as discrete chance or decision nodes with a fixed set of values.

Structure Learning Data from the SaGA corpus was provided as input for learning the model structure. As score-based algorithms K2 and MCMC are employed, as implemented in the BNT toolkit (Murphy, 2001). Both algorithms are parameterized to use the BIC score. MCMC's starting position is a graph without edges and the burn-in time is set to *5n* whereby *n* is the number of nodes. Since MCMC is nondeterministic, the algorithm had to run five times per data set, which sometimes resulted (sometimes) in different network structures. For each data set, the network structure found most often by the algorithm was chosen. As constraint-based methods PC and NPC are employed, as implemented in HUGIN with different significance levels.

Parameter Estimation Once the structure of the network had been determined, its maximum likelihood estimates of parameters could be computed employing the EM algorithm as implemented in the HUGIN toolkit.

Rule Definition for Decision Nodes Decision nodes implement the empirically based dependence of gestural form features on other gesture features, due to technique-specific patterns and/or referent characteristics (Section 5.4.4). A set of if-then rules is specified in each decision node of the network. For our current domain of application, a set of 50-100 rules is defined in each node. In the following, the character of these rules is illustrated by way of example.

Consider, for instance, that the circular window from the introductory example (Section 1.1) is to be depicted and that the probabilistic part of the network has determined (1) that a gesture is to be used, (2) that the representation technique to be utilized is 'drawing', (3) that the gesture is to be realized with the right hand, and (4) that the pointing handshape 'ASL-G' is to be used. The definition of rules is based on the empirical corpus analysis in Section 4.2.3. With respect to representation technique-specificity, the rules account 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 property 'round2d' is considered in terms of a curved movement with a circle-shaped trajectory. The following set of rules determines the values for palm and BoH orientation, movement type, and movement trajectory:

```
if (and (Technique="drawing"), PO="PDN").

if (and (Technique="drawing"), BoH="BAB").

if (and (Technique="drawing",
ShapeProp="round2d"), MT="curved").

if (and ( Handedness="rh", Technique="drawing",
ShapeProp="round2d"), MD="MR>MD>ML>MU").
```

A second example illustrates how the (probabilistic) choice of handshape influences choices made in the decision nodes. Given that the probabilistic part of the network determined (1) that a gesture is to be used and (2) that the representation technique to be utilized is 'indexing', the rules determine the gesture's palm orientation depending on the chosen handshape, reflecting the empirical fact that indexing gestures divide into 'flat hand indexing' and 'pointed indexing' (Section 4.2.3). The flat hand indexing handshape ASL-B is combined with a sideways palm orientation, which makes the palm orientation further dependent on handedness: for left-handed gestures, it is oriented to the right, and for right-handed and two-handed gestures it oriented to the left[1]. For pointed indexing gestures, no such distinction is necessary—the palm is oriented downwards independently of handedness. The following set of rules determines palm orientation in indexing gestures accordingly:

```
if (and (Handshape="ASL-B", Handedness="lh",
Technique="indexing"), PO="PTR").

if (and (Handshape="ASL-B", Handedness="rh|2h",
Technique="indexing"), PO="PTL").

if (and (Handshape="ASL-G",
Technique="indexing"), PO="PDN").
```

6.1.1 GNetIc example model

An example of a GNetIc model is given in Figure 6.1. The probabilistic part was learned from the data of one particular speaker in the SaGA corpus (P15) using

[1] Two-handed gestures are principally treated like right-handed gestures. The difference is that in two-handed gestures, the gesture specification for the right hand is symmetrically mirrored with the left hand.

the NPC algorithm with a significance level of .001. From the network structure, it becomes apparent that the node determining gesture occurrence is connected with three input nodes: one from the set of referent features ('shape property'), one from the set of discourse and linguistic context variables ('thematization'), and one from the set of previous gesture features ('last gesture'). Similarly, the other three gesture nodes are connected with input nodes of different kinds. Decision are connected to referent features and previously determined gesture features as described in Section 5.4.4.

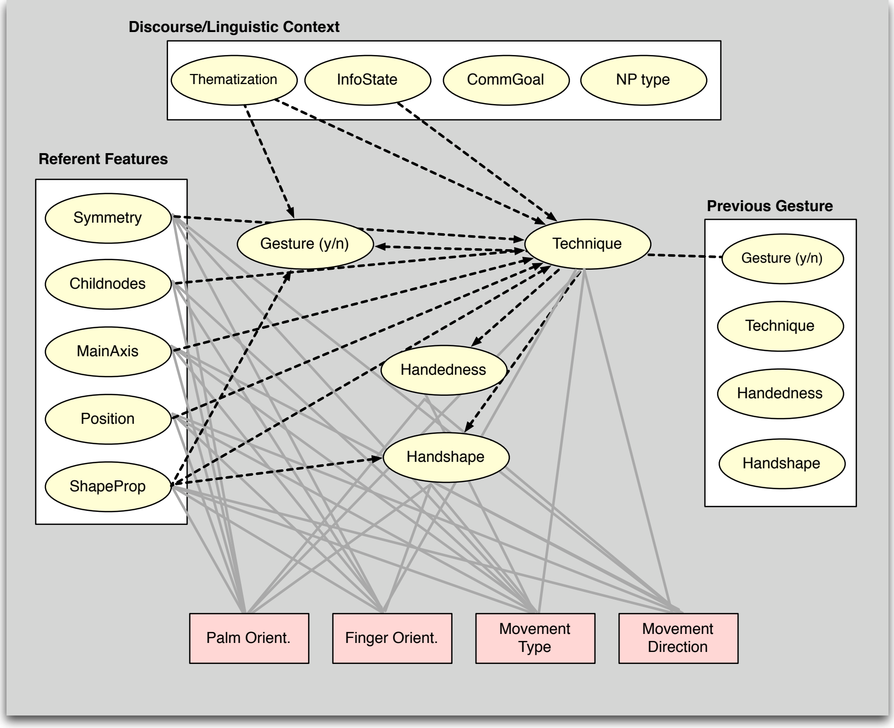


Figure 6.1: Example of a GNetIc model: the probabilistic part of the model is learned from the data of speaker P15 (black dotted links), the rule-based part is defined generally (gray links).

6.2 Realizing the Speech and Gesture Production Architecture

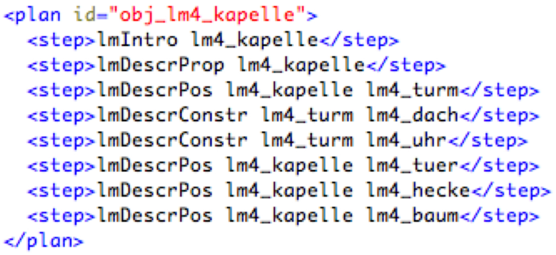
The modular production architecture (Section 5.5) is basically implemented in C++. Three major frameworks are incorporated, namely the IDT formalism (Section 3.1) for the representation of imagistic knowledge, the SPUD*lite* system for natural language generation (Stone et al., 2003), and the ACE engine to realize synchronized multimodal behaviors on the basis of MURML specifications for virtual agents (Section 6.2.5). The production process is treated as a multi-agent problem solving task. All modules are modeled as software agents that operate concurrently and proactively on a central working memory, realized as a globally accessible, structured blackboard.

As opposed to message-passing architectures, the overall production process thus evolves by each module observing entries in the working memory, taking local action if possible, and modifying or posting entries in response. In this way, interaction among the modules carries out content planning and micro-planning in an interleaved and interactive manner, and it enables bottom-up processes in both modalities.

6.2.1 Communication Planner

The Communication Planner is concerned with providing communicative goals which initiate the production process of multimodal utterances. The planner is equipped with *communicative plans* in an XML format to describe a couple of buildings from the virtual SaGA world. Each of these plans consists of a sequence of communicative goals built on the basis of descriptions as given in the SaGA study. See Figure 6.2 for an example plan. Each of the goals to be realized and communicated consecutively consists of the communicative intent ('lmIntro', 'lmDescrProp' etc.) along with the entire content.

This straightforward definition of how objects are described is intended to be substituted with a salience-based planning method as developed in Baake (2009). The algorithm computes communicative plans employing a salience ranking in terms of object size and shape, such that highly salient object parts are described first.



```
<plan id="obj_lm4_kapelle">
  <step>lmIntro lm4_kapelle</step>
  <step>lmDescrProp lm4_kapelle</step>
  <step>lmDescrPos lm4_kapelle lm4_turm</step>
  <step>lmDescrConstr lm4_turm lm4_dach</step>
  <step>lmDescrConstr lm4_turm lm4_uhr</step>
  <step>lmDescrPos lm4_kapelle lm4_tuer</step>
  <step>lmDescrPos lm4_kapelle lm4_hecke</step>
  <step>lmDescrPos lm4_kapelle lm4_baum</step>
</plan>
```

Figure 6.2: A communicative plan consisting of a sequence of communicative goals to describe the chapel from the virtual SaGA world (Figure 6.4). Each of the goals consist of an intent ('lmIntro', 'lmDescrProp' etc.) along with the entire content of the communicative intention.

The Communication Planner further acts as an interface between the generation system and the user. It, therefore, implements user control via a WiiMote which enables the human addressee to provide feedback regarding the previous utterance given by the virtual agent. Positive feedback (button '1') gives rise to publish the next communicative goal on the blackboard, in the case of negative feedback (button '2'), the previous utterance is repeated once again. Raw data from the WiiMote is accessed via Bluetooth and further processed by a server module implemented in C++. The Communication Planner incorporates a client that maps incoming data on the above-mentioned feedback events.

6.2.2 Content Representation and Blackboard

Imagistic knowledge is operationalized with the IDT formalism (Section 3.1). The application domain—buildings from virtual world in the SaGA study (Section 4)—is modeled in the XML-based description format.

Speech formulation draws upon propositional knowledge in terms of logical formulae. These are based on an ontology representing entities, attributes, and relations of the virtual SaGA world realized in Prolog, a declarative logic programming language. That is, the program logic is expressed in terms of relations, represented as facts and rules. While initially aimed at natural language processing, Prolog is used in different areas such as theorem proving, expert systems, ontologies etc. It is, therefore, well suited to code world knowledge with regard to generating verbal utterances. Prolog is used in the comprehensive implementation of SWI-Prolog[2] which is widely used in research and education as well as for commercial applications. It has a rich set of features and libraries for interfacing with Java and other programming languages.

Multimodal concepts are specified in the XML-based IDT description format. They aim to interface between imagistic and propositional knowledge.In general, there are two types of multimodal concepts. Unary concepts that take one IDT (node) specification are employed to specify for shape properties, such as 'longish', 'circular', 'spherical', and 'pointed'. Binary concepts are utilized for size features ('large', 'small', etc.) and spatial relations ('left-of', 'right-of').

An additional source of information is the discourse record, which contains two kinds of information about what has been communicated so far: (1) the set of propositional formulae that have already been communicated in the present discourse and (2) the surface structure of previous utterances in terms of gesture form feature matrices and natural language sentences.

A globally accessible, structured blackboard is realized as working memory, in which all representations accumulate as they arise in the development and use of speech and gesture. The blackboard is a shared repository of representations. All generation modules are able to access information provided by other modules and to publish their own contributions.

The blackboard is structured into several parts, each of which associated with a mutex variable to protect the shared resources:

[2] http://www.swi-prolog.org/

- Active imagistic knowledge
- Active propositional knowledge
- Matrix of previously uttered gesture form features
- Communicative goal(s)
- Matrix of gesture form features
- Tagged natural language sentence

6.2.3 Content Planning Modules

An integral part of *Image Generator* and *Preverbal Message Generator* is an activation function. Applied to the IDT representation, entities involved directly in the communicative goal receive full activation (1.0), which is propagated down the tree in a bisecting way such that child nodes receive half of their parent's activation value. For IDT nodes with a significantly high activation, the Image Generator performs spatial perspective-taking to determine how the objects appear from the particular point of view adopted and along that particular view direction. This operation is directly implemented as a transformation on object schemas using a standard view transformation from computer graphics. All occluded IDT nodes which are culled.

Once the operation is completed, the propositional knowledge base is updated to account for the adopted perspective. The activation function implemented in the Preverbal Message Generator assigns full activation to all propositions that are directly involved in the communicative goal, half of the activation to proposition which contain other predicates from the highly activated ones etc..

If the activation of IDTs and propositions exceeds a particular threshold (0.2) these objects are imported into the respective working memory structure.

6.2.4 Microplanning Modules

Speech Formulator The Speech Formulator employs SPUD*lite*[3] (Stone, 2002), which is a lightweight Prolog implementation of the original SPUD system implemented in the functional programming language 'Standard ML of New Jersey' (for a detailed overview on how far SPUD*lite* simplifies SPUD, see Buschmeier (2008)). The Prolog framework is interfaced via the bi-directional Java-Prolog interface JPL[4].

SPUD*lite* takes three types of input, (1) a lexicalized tree-adjoining grammar, (2) a knowledge base in the form of predicate logic formulae, and (3) the specification of communicative goals to be realized by the natural language sentence. SPUD*lite* carries out the different microplanning tasks (lexical and syntactic choice, referring expression generation, and aggregation) at once by treating microplanning as a search

[3] http://www.cs.rutgers.edu/∼mdstone/class/taglet/taglet.pl; retrieved 2011-01-26

[4] http://www.swi-prolog.org/packages/jpl/; retrieved 2011-01-26

problem. It tries to find an utterance which meets the constraints set by its input. This is done by exploring the search space, spanned by the linguistic grammar rules and the knowledge base, until a goal state is found. Non-goal states are preliminary utterances and are extended by one linguistic structure in each search step until a syntactically complete utterance which conveys the specified communicative goals is found.

Exploiting the capability of SPUD*lite* to connect sentence parts with the semantic information they convey, the result of speech formulation is a natural language sentence tagged with semantic information in terms of propositional formulae.

Gesture Formulator The Gesture Formulator brings the GNetIc models to application within the overall production framework via the HUGIN C++ interface. The HUGIN API provides methods to load existing networks, propagate information and access the posterior probabilities of nodes in the network.

To provide the GNetIc models with all relevant information, a set of analysis functions is implemented to extract the referent features as required to inform the networks' input nodes. The number of child nodes is directly available from the tree representation. Profile properties encoded in the IDT representation are accessed to determine the symmetry value. A comparison of axes extent reveals information about the main axis. The object's position is calculated from the IDT's transformation matrix. And, finally, the IDT is analyzed for shape properties by means of the multimodal concepts. A comparison method (Sowa, 2006) is employed to compare the IDT representation of the referent and the underspecified descriptions of multimodal concepts. For each concept, thus, a similarity value between 0.0 and 1.0 is calculated and the shape property with the highest score is determined to be the referent's shape property. To further inform the GNetIc input nodes, additional details are accessed from the blackboard, including information about the previously performed gesture, the information state of the very entity, and whether the noun phrase accompanying the gesture would fulfill exprssion of the sentence's theme or rheme.

All information is propagated through the GNetIc model via the HUGIN C++ interface. For the posterior probabilities inferred by the network, values with maximum score are selected following the maximization strategy.

Each gesture feature matrix derived from the decision network is then analyzed for its semantics by applying the GNetIc networks for diagnostic inferences. The resulting *gesture*-IDT is compared to the initial *referent*-IDT by means of formal graph unification. If a discrepancy between the representations is detected, the GNetIc model is requested again with the additional constraint of using either a different technique to plan a gesture or a different handshape to return a feature matrix.

Finally, the Gesture Formulator fills the gesture matrix on the blackboard to specify the form features of the readily formulated gesture.

6.2.5 Surface Realization

In the virtual human Max, the *Articulated Communicator Engine* (ACE) is employed for behavior realization. The ACE model is an incremental production model which aims to create synchronized multimodal behaviors, e.g., on the basis of MURML specifications (Kopp, 2003; Kopp and Wachsmuth, 2004).

MURML

The *Multimodal Utterance Representation Markup Language* (MURML) is an XML-based representation format that specifies all overt aspects of a communicative utterance (Kranstedt et al., 2002; Kopp, 2003). MURML descriptions assume an incremental process model that synthesizes continuous speech and gesture in successive chunks. A MURML specification consists of two major parts, namely (1) a textual definition of the verbal part of the utterance, i.e., the words to be spoken, and (2) specifications of para-verbal and non-verbal behaviors such as prosodic foci, gestures, and facial animations to be realized. MURML provides the possibility to specify behaviors that overlap temporally, either by an explicit specification of absolute times (start, end, duration) in relation to the start of the overall chunk, or implicitly by an affiliation with particular words.

Gesture specification in MURML is characterized by a high degree of flexibility. Gestures can be represented either as a parametric keyframe animation, or in terms of spatio-temporal form features (see Figure 6.3 as an example). In the latter case, a hand-arm configuration is specified in terms of four components: (1) wrist location, (2) handshape, (3) extended finger orientation, and (4) palm orientation. Values for each of these components are described either numerically or symbolically, building upon a notation system for sign languages (Prillwitz et al., 1989). A MURML gesture representation is a combination of these features. To specify constraints for a particular feature over a certain period of time, two types of constraints are available: *static* constraints to define that a feature is to be held constant for a certain period of time, and *dynamic* constraints to specify a significant movement within a feature.

Behavior Realization

The ACE model is a production system that creates multimodal behavior on-the-fly from a given MURML specification. It enables virtual agents to show synchronized verbal and non-verbal behaviors in a human-like flow of multimodal behavior.

The ACE model is based on the segmentation hypothesis McNeill (1992), according to which the co-production of continuous speech and gesture is organized in successive *chunks*, each of these expressing a single idea unit. With regard to gestures, these units correspond to gesture phrases (cf. Section 2.1.2). Concerning speech, the units relate to intonation phrases which are separated by significant pauses. Following

```xml
<specification>Der Turm hat<time id="t1"/>ein spitzes Dach<time id="t2"/></specification>

<behaviorspec id="gesture1">
   <gesture>
      <affiliate onset="t1" end="t2"/>
      <constraints>
         <symmetrical dominant="right_arm" symmetry="SymMS">
            <parallel>
               <static slot="HandShape" value="ASLg"/>
               <static slot="PalmOrientation" value="DirLD"/>
               <static slot="ExtFingerOrientation" value="DirLU"/>
               <dynamic slot="HandLocation">
                  <dynamicElement type="linear">
                     <value name="LocLowerChin LocCenterRight LocNorm" type="start"/>
                     <value name="DirD" type="direction"/>
                     <value name="24" type="distance"/>
                  </dynamicElement>
               </dynamic>
            </parallel>
         </symmetrical>
      </constraints>
   </gesture>
</behaviorspec>
```

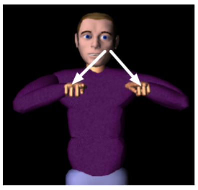


Figure 6.3: Example of a MURML specification and its realization with the virtual agent Max.

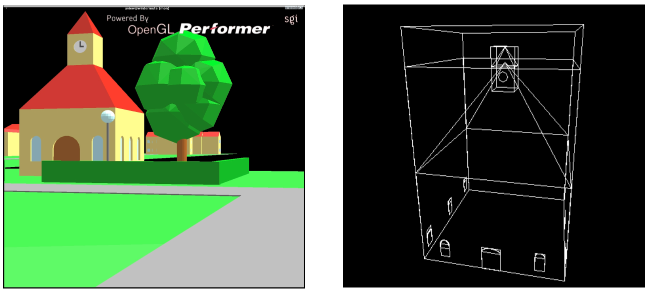
rather the semantical structure than the syntactical phrase structure, intonation phrases are further characterized by a meaningful pitch contour with exactly one primary pitch accent, the *nucleus*.

In order to achieve a natural and synchronized flow of speech and gesture across successive coherent chunks, the model employs a number of cross-modal adaptation mechanisms. The gesture stroke is planned to start in synchrony with the onset of its lexical affiliate and to span for the duration of the gesture's affiliated words. If necessary, a post-stroke hold is inserted after the stroke phase, or the stroke is repeated. Likewise, the duration of the silent pause between two intonation phrases may vary according to the required duration of the preparation for the next gesture.

The animation of co-verbal gesture is realized in a hierarchical way. A motor planner is provided with timed form features as described in the MURML specification. In the following, the problem of complex control is broken down into sub-problems to be solved by independent motor control modules for the arms, wrists, and hands.

6.3 Generation Example

To illustrate the production process of a multimodal utterance with the production system described so far, the generation of an example utterance to be realized with the virtual agent Max is described in the following. Let us assume that Max is engaged in describing the chapel following the communicative plan given in Figure 6.2. As a prerequisite, let us further assume that Max has chosen a point of view in front of the building, facing the chapel at a distance of 10m. Accordingly, the imagistic and propositional content representations have been adopted to that viewpoint.



(a) (b)

Figure 6.4: Chapel from the virtual SaGA world (a) and a visualization of its IDT representation (b).

Let us assume the communicative goal provided by the Communication Planner to be

```
lmDescConstr lm4_tower lm4_roof
```

The content planning modules continuously poll the blackboard for new communicative goals. As soon as the goal is published, they start to respectively select and activate imagistic propositional knowledge. The IMDs in Max's imagery labeled 'lm4_tower' and 'lm4_roof' receive an activation value of 1.0. This activation is propagated through the tree so that parent and child nodes still receive half of their parent's activation The activated part of the IDT is imported into working memory along with a set of significantly activated propositions which contain 'lm4_tower' or 'lm4_roof' as an argument (high activation) or are related to one of the referents, e.g., by a part-of relation (less highly activated):

```
shared(entity(lm4_tower, single)).
shared(instance_of(lm4_tower, tower)).
private(entity(lm4_roof, single)).
private(instance_of(lm4_roof, roof)).
private(property(lm4_roof, pointed)).
private(part_of(lm4_tower, lm4_roof)).
```

Now the formulator modules come into play, bringing production process to the microplanning stage. The Speech Formulator lets the SPUD*lite* system search for an adequate verbalization of the propositions on the blackboard. Figure 6.3 shows the resulting LTAG tree generated for the example utterance 'the tower has a tapered roof'. Each surface element is annotated with information about the propositional meaning it conveys.

The Gesture Formulator runs twice, once for each entity mentioned in the verbal utterance whose activation is high enough to place on the blackboard, namely ‘lm4_tower’ and ‘lm4_roof’. In the run for the entity ‘lm4_tower’, all facts available on the blackboard are entered into and propagated through the GNetIc model used by the Gesture Formulator. The corresponding noun phrase ‘the tower’ thus fulfills the role of the sentence’s theme, and no shape property exceeds the necessary threshold. These facts are entered into the network and yield a gesture production likelihood of .06. Following the maximization strategy, the Gesture Formulator consequently halts the generation of the gesture and moves on to the second entity, ‘lm4_roof’. Its referring expression ‘a tapered roof’ is the sentence’s ‘rheme’, and the shape property with the highest similarity value is ‘pointed’ (see Figure 6.5 for visualizations of imagistic representations of ‘lm4_roof’ and the schematic of ‘pointed’ which were compared here to analyze the imagistic representation of the referent for its shape properties). Propagated through the network, the resulting gesture production likelihood is 1.0.

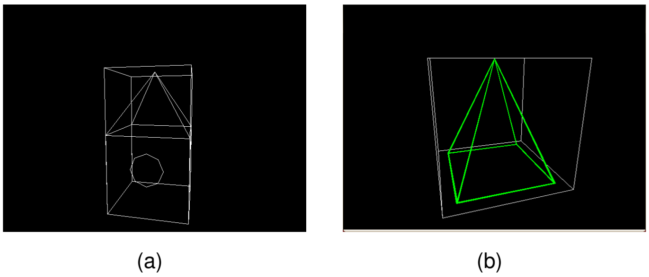
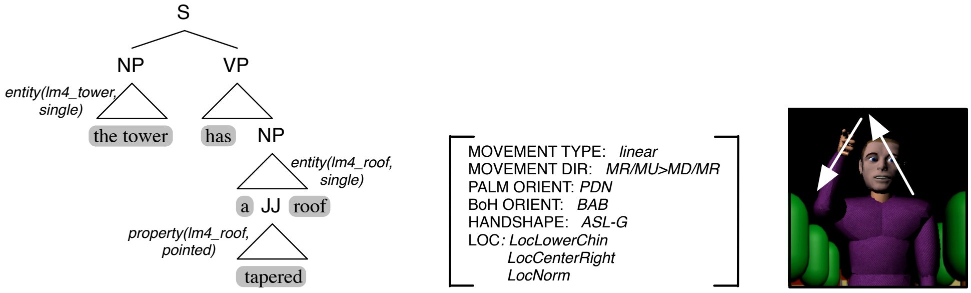


Figure 6.5: Visualization of imagistic representations for ‘lm4_roof’ (a) and the schematic of the shape feature ‘pointed’ (b).

The Gesture Formulator now enters facts required by the node ‘representation technique’: thematization (‘rheme’), information state (‘private’) and the shape property ‘pointed’. This results in the maximum probability of 1.0 for a drawing gesture. Accordingly, handedness and handshape are specified. The likelihood for a right-handed gesture is .83 given that the representation technique is drawing, and the likelihood for handshape ‘ASL-G’ is 1.0 based on evidence about the referent’s main axis (‘none’) and the representation technique.

The next step is the determination of values in the decision nodes. Evidence about the referent features and the previously determined gesture is propagated through the network revealing value specifications for the features ‘handedness’ (RH), ‘handshape’ (ASL-G), and ‘representation technique’ (drawing). These are provided for making further decisions in the decision nodes, resulting in a choice for the palm to be oriented downwards, the BoH oriented away from the speaker’s body, the movement type being linear and the trajectory being a sequence of an upwards/right movement and an downwards/right movement not unlike tracing an upside-down letter ‘V’ (‘MR/MU>MD/MR’).

The gesture matrix is completed by querying the gesture location from the IMD accounting for the referent’s position. The resultant feature matrix for the gesture is given in Figure 6.3. Now the gesture specification is planned: the entity ‘lm4_roof’, is set in synchrony with the verbal utterance segment referring to the same entity: the noun phrase ‘a tapered roof‘. Finally, both utterance parts are translated into a MURML description, which is then sent for realtime realization to the virtual agent Max. Figure 6.3 shows a screenshot of the resulting gesture.



(a) LTAG tree generated for the utterance ‘the tower has a tapered roof’ by the Speech Formulator.

(b) Gesture form feature matrix generated by the Gesture Formulator.

(c) Utterance realized with the virtual agent Max.

Figure 6.6: Generation results for an example utterance.

6.4 Modeling Results and Discussion

6.4.1 GNetIc Models

GNetIc models were learned by applying (1) different structure learning algorithms and (2) different data sets. Concerning the former, four different algorithms—two score-based (K2 and MCMC) and two constraint-based (PC and NPC)—were examined (Section 5.3.2). Regarding the second issue, the choice of training data, five training sets of individual speaker’s data were used, as well as one combined set of data from several speakers.

Figure 6.7 displays the results of all four algorithms for five training sets of individual speakers’ data (rows 1-5), and another dataset combining the data of all five speakers (row 6). The constraint-based algortithms PC and NPC were applied with a significance level of p=.01.

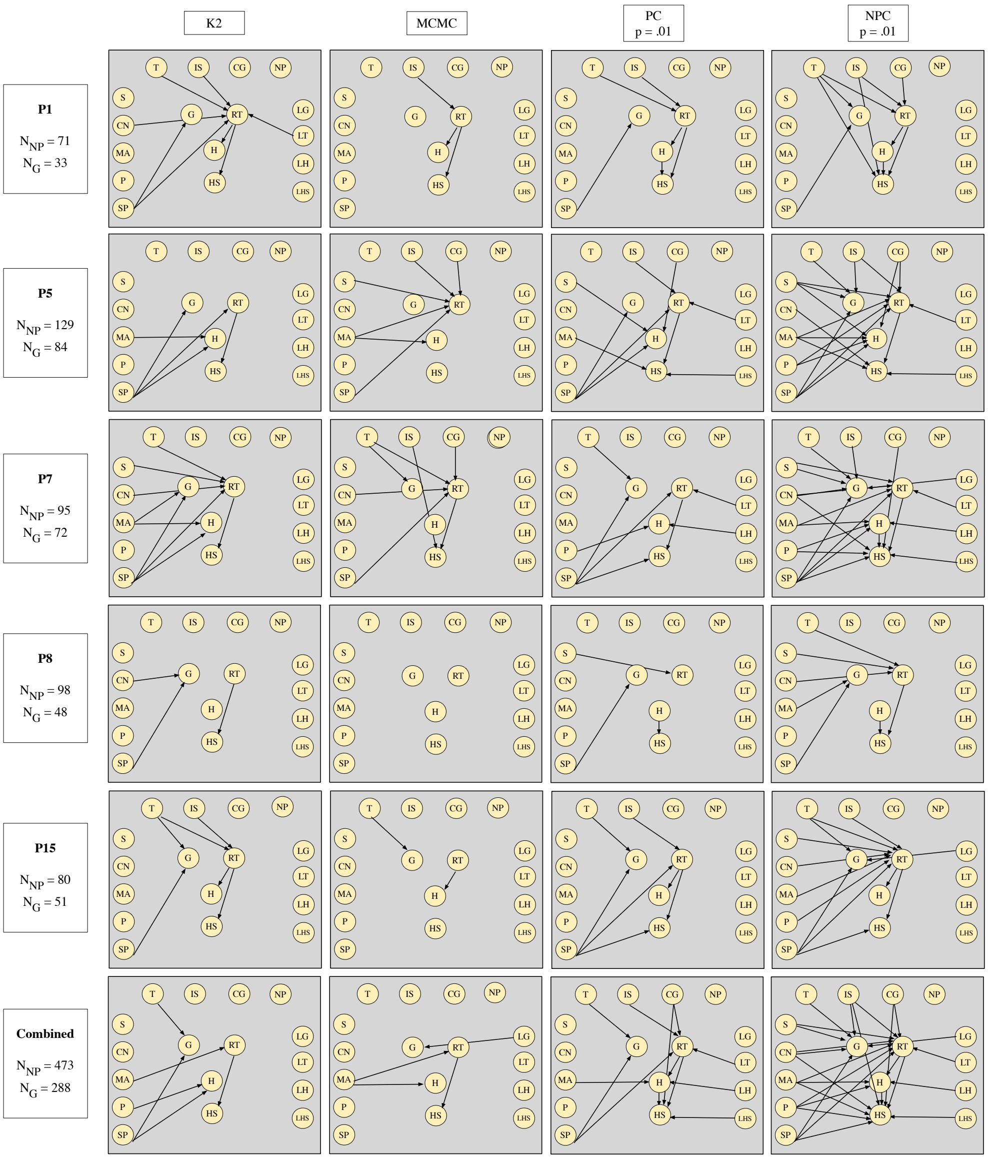


Figure 6.7: Network structures obtained with different algorithms (columns) for different data sets (rows). See Table 5.1 for the reading of the variables and their value sets..

Differing Network Structures

A comparison of the networks in each row reveals that learned structures are subject to (1) the different learning algorithms, and (2) the data set.

Comparison regarding Algorithms With regard to the former, network structures trained with MCMC typically show the lowest number of edges, whereas networks trained with NPC show various links among variables. As concerns the constraint-based algorithms PC and NPC, it is conspicuous that dependencies found by the PC algorithm are always also covered by the network structure found by the NPC algorithm. In most of the cases, NPC detects more edges than PC which is not surprising due to the necessary path condition. This result is in line with Steck and Tresp (1999) who found that the NPC algorithm learns more of theedges present in a dataset than the PC algorithm.

Further, it turns out that despite the different solutions yielding multiple edges in common, there are also some inconsistent edges, in which the structures differ. In this context, it is noteworthy that score-based approaches as K2 and MCMC consider a *global* measure for the entire network. Constraint-based methods, in contrast, remove all those edges from a network for which a conditional independence statement can be derived from the data. They do not take into account the structure of the network as a whole and are, therefore, considered *local*.

Comparison regarding Data Sets With regard to the second source of differences, data sets, a comparison of the different network structures in each column shows obvious differences. That is, different networks are resulting from different speakers' data. In the following, to elaborate on this, networks trained with one and the same algorithm, namely NPC, will be compared.

Take, for instance, the network of speaker P1: the production choices made are predominantly dependent on discourse context, i.e., neither referent features nor the previous gesture have an impact. At first glance, it seems implausible that fundamental choices in the planning of iconic gestures can be done without considering any aspects of the object. On closer inspection, however, speaker P1 has a very strong preference for drawing gestures (46.9% vs. 15.3% in the whole corpus) which goes along with a high proportion of handshape ASL-G typically used for drawing gestures. Note, iconicity of drawing gestures is mainly established by the movement trajectory determined in the decision nodes.

Another remarkable case is that of speaker P5; in this individual's data, a large number of dependencies were found. Every single generation choice in this network has predecessor nodes from all three variable sets. In other words, every inference process relies on evidence from at least three variables of different types. In fact, the choice of handshape depends on evidence from six variables ('Childnodes', 'Technique', 'Position', 'Symmetry', 'Main Axis', and 'Previous Handshape'). In contrast, in speaker P8 the same choice is only influenced by one predecessor node ('Technique'). Moreover, P5 is unique in the data in that variables linked to the previous gesture influence 'Technique' and 'Handshape'.

Examining the networks with particular attention to the impact of different influences types reveals that referent features and discourse context play the most significant role. The previous gesture has less influence and the linguistic context variable is entirely negligible at a significance value of .01.

Link Strength

In particular, how far a variation of the significance level impacted modeling results was tested. Not surprisingly, the greater the count of links present in the network structures, the lower the significance level.

In this way, the link strength becomes apparent by varying the significance level used in the conditional independence tests. That is, it is possible to judge the strength of the dependencies among variables. As an example, consider Figure 6.8, which displays two network structures. The two networks obviously differ with regard to the number of dependencies learned. Whereas gesture production choices in the left network are predominantly influenced by the discourse context (nodes 'T', 'IS', and 'CG'), gesture features in the right network are additionally influenced by all referent features (nodes 'G', 'CN', 'MA', 'P', and 'SP') as well as by the previously performed gesture (nodes 'LG', 'LT', 'LH', and 'LHS'). Moreover, only a part of the learned dependencies is highly significant. These connections are found at a significance level of 0.001, whereas other links are less strong, however, still significant (significance level of 0.01 or 0.05). In the left network, for instance, the decision to use a gesture (node 'G') is strongly influenced by the shape properties of the referent (node 'SP'), whereas link strength for the discourse factors thematization (node 'T') and number of subparts (node 'CN') are less strong. Depending on the significance level used for learning the network structure, the resulting network structures are different: all three connections would, e.g., be present in the resulting network for a significance level of 0.05, whereas for a significance level of 0.001 only the strongest link would be learned.

Applying GNetIc models trained with varying significance levels for generation choices revealed the following: if too many links were connected with outcome nodes, the HUGIN decision engine was unable to process these adequately. In these cases, an equal distribution of probabilities was returned, which can be attributed to the fact that these models learned from sparse data with heterogenous characteristics (cf. Wittig, 2002). Therefore, to be able to make appropriate inferences with the model (going beyond prior probabilities in the outcome nodes), the significance level has to be rigorous enough. Along the same lines, Abellán et al. (2006, p. 4) argue that,

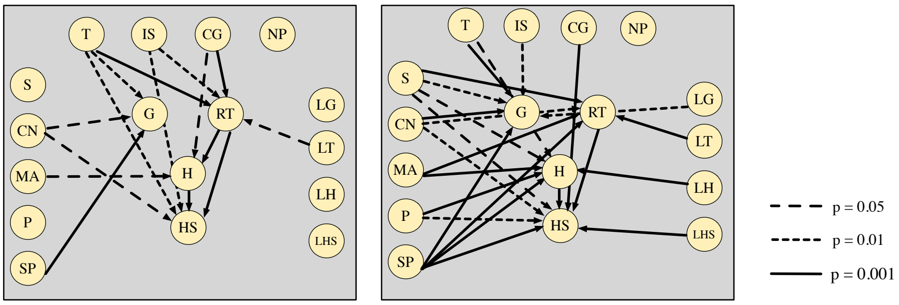


Figure 6.8: Network structures obtained with NPC algorithm for two different speakers (left, right): link strength corresponds to significance level.

"depending of the sample size, it can be more convenient to use a simpler graph than the true one."

This constraint led to a situation in which the linguistic constraint (NP type) was no longer influential, the case for only two networks when trained with a rather low significance level of .05. A fact that intensifies the nevertheless low influence of NP type on gesture features is that the NP variable has a large set of values. Larger sample sizes or an aggregation of values would, possibly, change the situation.

6.4.2 Generation with the Overall Production System

The resulting gesturing behavior for a particular referent in a respective discourse context varies in dependence on the decision network used for gesture formulation. In Figure 6.9, examples from five different simulations are given, each based an identical initial situation, i.e., all gestures are referring to the same referent (the round church window from the introductory example (Section 1.1)), and under identical contextual constraints. The resulting nonverbal behavior varies significantly depending on the decision network underlying the simulation: For P7, no gesture is produced at all, whereas for P5 and P8, posturing gestures are produced which, however, differ in their low-level morphology. For P5, the handshape ASL-C is employed using both hands, while in the simulation for P8 ASL-O is used with the right hand only. P1 and P15 both use drawing gestures which differentiate in their handedness.

That is, the different network structures learned from different speaker's data result in inter-individually different simulations of gesturing behavior. The conclusion to be taken is, therefore, that the GNetIc simulation approach is, first, valuable for an adequate simulation of speaker-specific gestures (as evaluated in the next section). Second, it reveals that gesture feature choices are influenced in inter-individually different ways, suggesting that different speakers actually employ different strategies in behavior production.

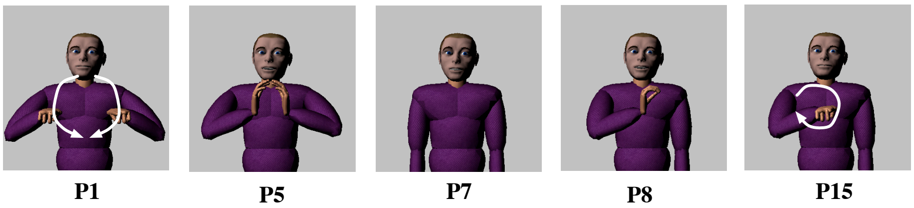


Figure 6.9: Example gestures from the simulation of different speakers accompanying the words 'the church has a round window'. Each utterance is produced for the same referent (a round window of a church) in the same initial situation.

6.5 Summary

This chapter was concerned with the realization of the previously developed concepts and methods in a virtual agent. It described how GNetIc models were built and how the speech and gesture production engine into which the gesture networks were integrated was implemented. Via this method, the virtual agent Max was enabled to explain buildings of the virtual SaGA environment to a human user—autonomously and in real time—being equipped with proper knowledge sources, i.e., communicative plans, lexicalized grammar, propositional, and imagistic knowledge about the world.

To explore GNetIc generation models, networks were trained using different learning algorithms and from different data sets. A comparison of the resulting networks revealed that learned structures differ with regard to both. The differences due to algorithm choice were obvious in the number of links, but also became apparent in inconsistent links—a result of the fact that score-based algorithms (K2, MCMC) employ a global measure, while constraint algorithms (PC, NPC) employ a local measure. Differences due to the data set on which networks were trained revealed that individual differences are not only present in the overt gestures, but also in the production process they originate from. Whereas gesture production in some individual networks was, e.g., predominantly influenced by visuo-spatial referent features, other networks revealed a stronger influence of discourse context.

The application of constraint-based algorithms provided the advantageous ability to choose between significance levels. In doing so, it was possible to judge the strength of dependencies among variables. However, applying constraint-based structure learning with rather low significance levels did not lead to appropriate inference results due to the heterogeneous character of limited data. Therefore, a higher level of significance is preferred.

CHAPTER 7

Evaluation

The previous chapter described the implementation of the GNetIc model and its application in a production framework for multimodal utterances. The final step which is still due now is an evaluation of the generation results. According to the two-fold research objective pursued in this thesis, results will be evaluated in two ways.

First, in a *prediction-based* evaluation, it will be investigated, to what extent the derived model enables a prediction of empirically observed gestural behavior. In general, this evaluation method aims to measure the model's prediction accuracy by computing how often the model's assessment is in agreement with the actual gesturing behavior of human speakers in the SaGA corpus. In particular, this corpus-based evaluation will be used to compare how different structure learning algorithms (score-based vs. constraint-based) and training sets (individual vs. combined) affect the prediction accuracy of the model.

Second, in a *perception-based* evaluation, it will be assessed if and how automatically generated gestures, realized with a virtual agent, are beneficial for human-agent interaction. This investigation comprises two major aspects, namely the *communicative* role of gesture use and the role of gestures in the *subjective impression* a human has of the interaction. With regard to the first point, the user's uptake of gestural information is explored, while the second issue is concerned with (1) the quality of the produced gestures as rated by human users; (2) whether an agent's gesturing behavior could systematically alter a user's perception of the agent's likeability, competence, and human-likeness; and (3) whether producing gestures like a particular *individual* or like the *average* speaker (i.e., a GNetIc model learned from the combined data of several speakers) is preferable.

Results of the prediction-based evaluation have been published in Bergmann and Kopp (2010a), results of the perception-based study in Bergmann et al. (2010).

7.1 Prediction-based Evaluation

To get the most complete picture possible of GNetIc's prediction accuracy, the networks will be evaluated with regard to each single decision which is made in the generation process. As part of the gesture features are determined probabilistically (with chance nodes) while others are determined in a rule-based way. These two issues will be distinguished in the evaluation as well.

7.1.1 Evaluation of Probabilistic Generation

The evaluation of generation choices made probabilistically with chance nodes in the GNetIc models compares (1) how different structure learning algorithms (K2, MCMC, PC, NPC) and (2) training sets (individual vs. combined) affect the prediction accuracy of the model.

Results

To investigate the quality of the different network structures, each model's prediction was compared with the empirically observed gesturing behavior from the SaGA corpus. For each network structure found, its maximum likelihood estimates of parameters were computed with the standard EM algorithm (Lauritzen, 1995).

In a leave-one-out cross-validation each model has been learned from $n - 1$ selected data cases (n is the number of cases in the data set), leaving out the data case i. Then the model was tested on that data case. For each model the procedure is repeated n times so that each data case was used for testing once. The reported results, as summarized in Table 7.1, are an average over all n runs.

Table 7.1: Prediction accuracy in single features for the networks learned using different structure learning algorithms from individual and combined speaker data, respectively.

		Accuracy (%)							
Generation Choices	**Chance Level Baseline**	K2		MCMC		PC		NPC	
		Indiv.	Comb.	Indiv.	Comb.	Indiv.	Comb.	Indiv.	Comb.
Gesture (y/n)	50.0	82.5	83.3	68.7	60.7	82.5	83.3	84.8	77.6
Technique	20.0	60.6	49.1	50.9	49.1	66.4	62.6	59.5	59.9
Handedness	33.3	60.6	65.4	63.7	59.2	62.3	69.9	61.2	65.1
Handshape	16.7	65.1	49.8	60.6	49.8	71.3	59.2	67.8	53.3
Total Accuracy		69.3	57.4	61.2	55.5	71.3	69.1	67.8	65.8

It turned out that for all generation choices the prediction accuracy values clearly outperform the chance level baseline. In total, the prediction accuracy achieved with individual networks is, by trend, better than the accuracy achieved with networks learned from non-speaker specific data. This holds for the results of all four learning methods.

A comparison of the different learning techniques shows that networks learned with the score-based MCMC algorithm result in the lowest accuracy values, whereas best results are achieved with the constraint-based PC algorithm. In general, higher accuracy values were achieved with constraint-based algorithms in comparison to score-based algorithms. That is, the general advantages of constrained-based algorithms (Section 5.3.2) receive further support for their application in the current context of gesture generation due to their higher prediction accuracy.

7.1.2 Evaluation of Rule-based Generation

In a second step, the performance of the four local generation choices made in decision nodes of GNetIc is evaluated. Note that decisions are made in a particular order, which has an impact on the validation results. If one of the earlier choices does not match the observed value in the test data, the following decisions typically cannot match the data either. Assume for instance that the predicted representation technique is 'indexing' although the representation technique in the test data is 'shaping'. The subsequent choices concerning morphological gesture features would accordingly be made under false premises. Results for the rule-based decisions are, therefore, validated locally, i.e., taking the test case data for previous decisions as a basis and evaluating the quality of our decision making rules directly.

Results

Detailed results are given in Table 7.2. Note that, the same measure is not applied for all four variables. Palm and finger orientation are compared by calculating the angle between the two orientation vectors (cf. Section 4.1.5). For instance, there is an angle of 90° between 'left' and 'up', and an angle of 45° between 'left' and 'left/up' . A maximum angle of 180° is present if the two vectors are opposing (e.g. 'left' and 'right'), and can be considered the worst match. Considering this, the mean deviation for palm orientation of 54.6° (SD = 16.1°) and the mean deviation for finger orientation of 37.4° (SD = 8.4°) are quite satisfactory with deviations which lie well within the fuzziness of natural gestures in humans.

Movement direction is distinguished between motions along the following three planes: (1) sagittal plane (forward, backward), (2) transversal plane (left, right), and (3) vertical plane (up, down). Each segment in the generated movement description is tested for co-occurrence with the annotated value, resulting in an accuracy measure between 0 (no agreement) and 1 (total agreement). For multi-segmented movements the mean accuracy is calculated, i.e., if a generated movement consists of two segments from which only one matches the annotation, its similarity is estimated with a value of 0.5. Evaluating GNetIc with this measure gives a mean similarity of .75 (SD = .09). For the movement type (linear or curved), the standard measure of accuracy is employed, i.e., it is determined whether the generated value exactly matches the annotated value. The mean accuracy for the movement type is 76.4% (SD=13.6).

Table 7.2: Evaluation results of generation choices assessed in GNetIc's decision nodes.

Generation Choices	P1	P5	P7	P8	P15
Palm Orientation	37.1°	61.9°	76.4°	57.1°	40.5°
BoH Orientation	29.0°	41.7°	41.9°	27.9°	46.6°
Movement	.69	.84	.84	.56	.89
Movement Direction	.82	.76	.82	.61	.76

7.2 Perception-based Evaluation

Until now, it was evaluated in how far the GNetIc model is able to simulate gesture use as observed in human speakers. Altogether, given the large range of potential values (or value combinations) for each of the variables, the results are satisfactory. Moreover, generated gestures whose features do not fully coincide with the original data may still be beneficial for human-agent interaction, either by communicating relevant information, or by improving the subjective impression of the interaction as well as the impression of the virtual agent itself. Both issues will be investigated in the following.

Information Uptake The question whether and how speech-accompanying gestures contribute to communication cannot be answered unambiguously. While some research has shown that listeners' comprehension of speech was not influenced or supplemented by gesture use, other studies showed that listeners do incorporate gesturally expressed information into their broader understanding of communicative behavior.

Evidence against the communicative function of gestures for addressees was reported by Krauss et al. (1991). In four experiments it was investigated whether conversational gestures actually communicate. The authors found that performance was better than chance but markedly inferior to performance when words were used as stimuli. Further evidence along the same lines comes from a set of experiments with stimulus descriptions of different kinds (abstract graphic designs, synthesized sounds, samples of tea). Addressees had to select the described object from a set of objects. In none of the experiments accuracy was enhanced when participants were allowed to see the speaker's gestures (Krauss et al., 1995).

In contrast to this, evidence for the communicative function of gestures from the *listener's* point of view comes from studies in which speech was ambiguous or presented in a noisy environment. In these cases listeners were found to rely on gestural cues (Thompson and Massaro, 1986; Rogers, 1978). Moreover, Cassell et al. (1999) reported that listeners do attend to information conveyed in gesture, when that information supplements or even contradicts the information conveyed by speech. In a more detailed analysis, Beattie and Shovelton (1999a,b) discovered that gesture use was particularly beneficial for addressees with respect to certain semantic categories, namely the relative position and size of objects. Holler et al. (2009) replicated this finding for face-to-face interaction. They found that in some cases gestures were even more effective at communicating position and size information in a face-to-face condition compared to the presentation of video stimuli. Further support for the communicative function of gestures as judged by recipients comes from neuropsychology. Kelly et al. (2004) and Habets et al. (2010) found N400 effects, a marker of semantic integration, providing evidence that the iconic content of gestures is picked up and processed by listeners.

For the case of human-agent communication there are hardly any findings regarding the question whether humans take up information from an agent's gestures. Krämer et al. (2003) compared a gesturing virtual agent to text and audio conditions: no supporting effect of gesture use on comprehension and recall was observed. For the special cases of names and technical terms it was even found that the text and audio conditions were found to be more advantageous. Buisine and Martin (2007) investigated whether speech-gesture cooperation, in terms of redundancy and complementarity, influences participants' recall from a multimedia presentation given by animated agents. 2D cartoon-like agents were employed to give technical presentations associated with an image displayed on a whiteboard wherein speech was combined with deictic and iconic gestures. Results showed the advantage of a redundant strategy in this context: multimodal redundancy improved recall of the verbal content. Overall, redundancy yielded a recall proportion of 49%, while 41% of the complementary information was recalled correctly. Notably, redundancy influenced verbal but not graphical recall.

Presentation Quality and Agent Perception The second major issue to be addressed concerns user acceptance of generated gestures. There is increasing evidence that endowing virtual agents with human-like, non-verbal behavior may lead to enhancements of the likeability of the agent, trust in the agent, satisfaction with the interaction, naturalness of interaction, ease of use, and efficiency of task completion (Bickmore and Cassell, 2005; Heylen et al., 2002).

With regards to the particular effects of co-speech gestures, Krämer and colleagues found no effect on agent perception when comparing a gesturing agent with a non-gesturing one. The agent displaying gestures was perceived just as likeable, competent, and relaxed as the agent that did not produce gestures. In contrast, Cassell and Thórisson (1999) reported that non-verbal behavior (including beat gestures) resulted in an increase of perceived language ability and life-likeness of the agent, as well as smoothness of interaction. A study by Rehm and André (2007) investigated whether a gesturing agent would change the perceived politeness tone compared to that of the textual utterances and whether the subjective rating is influenced by the type of gestures (abstract vs. concrete). Their studies revealed that the perception of politeness depends on the graphical quality of the employed gestures. In cases where the iconic gesture was rated as being of higher quality than the metaphoric gesture, they observed a positive effect on the perception of the agent's willingness to co-operate. In cases where the iconic gesture was rated as being of lower quality than the metaphoric gesture, they observed a negative effect on the perception of the agent's willingness to co-operate. Moreover, in the aforementioned study by Buisine and Martin (2007), effects of different types of speech-gesture cooperation in an agent's behavior were found: redundant gestures increased ratings of explanation quality, expressiveness of the agent, likeability and positive perception of the agent's personality.

7.2.1 Study Design

A second evaluation was designed to evaluate the GNetIc model with regard to the following three questions: First, how much information do participants take up from a speech-gestural presentation? Second, is it possible to achieve a reasonable quality in the iconic gestures automatically derived with GNetIc, as perceived by users? Third, is the user's perception of an agent in terms of likeability, competence, and human-likeness altered by the agent's gesturing behavior?

The prediction-based part of the evaluation revealed an advantage of training gesture generation networks from speaker-specific data with regard to prediction accuracy—does this advantage also hold for the subjective perception of users? All three issues mentioned above will be investigated under consideration of the overall question whether it is preferable to produce gestures like a particular individual or like the average speaker. A user study was set up with a between-subject design in order to compare how individual vs. combined GNetIc networks are perceived.

Independent Variables

In a between-subject design, participants were presented with a description of a church building given by the virtual human Max. All descriptions were produced fully autonomously at runtime using the speech and gesture production architecture into which GNetIc is integrated (Section 5.5). The gesturing behavior of the agent was manipulated, resulting in five different conditions in which Max, notably, received the identical communicative goals and produced identical verbal utterances throughout (see Table 7.3). Furthermore, all gestures were generated from the same knowledge base.

In two individual conditions, *ind-1* and *ind-2*, the GNetIc networks were learned from data of individual speakers from the SaGA corpus (subject P5 in *ind-1*, subject P7 in *ind-2*). These two speakers were chosen because both speakers gestured quite frequently and approximately at the same rate. In a *combined* condition, the GNetIc network was generated from data of five different speakers (P1, P5, P7, P8 and P15). These speakers' gesture styles were thus amalgamated in one network. As a consequence, Max' gesturing behavior was not as consistent as with individual networks with regard to the probabilistic choices in the model. Finally, two control conditions were added. In the first one, *no gestures* were produced at all, whereas in the second one, values in the four probabilistic nodes were determined by chance (*random*). The latter condition can result, for instance, in gestures occurring at atypical (e.g., thematic) positions in a sentence since the network was applied for every noun phrase in the verbal description.

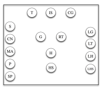
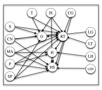
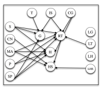
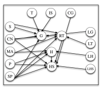
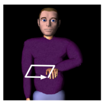
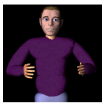
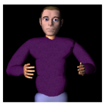
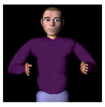
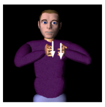
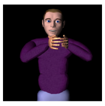
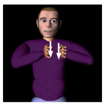
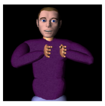
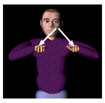
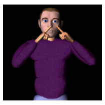
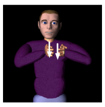
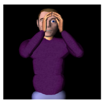
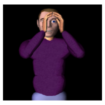
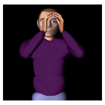
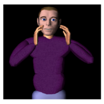
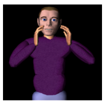
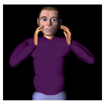
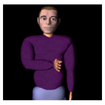
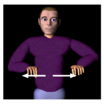
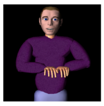
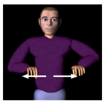
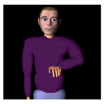
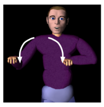
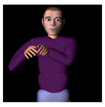
Overall, the virtual agent's verbal utterances were held constant and all gestures were created fully autonomously by the system. There was no within-condition variation, because choices in the Bayesian networks were not made via sampling, but by choosing the values with maximum a-posteriori probability. Furthermore, the values for the decision nodes were determined in the same rule-based way in all conditions, to ensure that no 'nonsense' gestures were produced throughout.

Table 7.3 shows the stimuli which resulted from the five different conditions. There is no wide difference across conditions in gesture frequency (either five, six or seven gestures in six sentences). However, the two individual GNetIc conditions are characterized by less variation in the production choices. In condition *ind-1* gestures are predominantly static ones while there are more dynamic shaping gestures in condition *ind-2*. Moreover, the gestures in condition *ind-1* are mostly performed with c-shaped hands, whereas in *ind-2* some gestures are performed with a flat handshape. In the *combined* GNetIc condition, a combination of different techniques is present. A similar mixture of techniques is observable in the *random* condition which is further characterized by inconsistency in handedness and handshapes. Moreover, gestures in this condition can occur at atypical positions in a sentence.

Dependent Variables

Following the three-part objective of the user-centered evaluation, dependent variables were assessed with a questionnaire grouped into three parts, (1) information recall, (2) presentation quality, and (3) agent perception.

Table 7.3: Stimuli presented in the five different conditions: verbal description given in each condition (left column; translated to English; gesture positions labelled with squared brackets); GNetIc networks from which the gesturing behavior were produced (top row); gestures produced (right columns).

		no gesture	*random*	*combined*	*ind-1*	*ind-2*
						
(1)	[The church is squared]...					
(2)	...and in the middle there is [a small spire.]					
(3)	[The spire]...					
	...has [a tapered roof].					
(4)	And [the spire]...					
	has [a clock].					
(5)	There is [a door] in front.					
(6)	And in front of the church there is [a low, green hedge].					
(7)	There is [a large deciduous tree] to the right of the church.					

Information Recall First, information uptake was judged in terms of the semantic information participants were able to recall. As in Buisine and Martin (2007), two types of recall were employed, *written* recall and *graphical* recall. An established micro-analytic coding method was employed using a set of semantic features (cf. Beattie and Shovelton, 1999a,b, 2001; Bergmann and Kopp, 2006; Holler and Beattie, 2002, 2003, 2004). The following set of semantic features was considered to capture the proportion of semantic information recalled from the object descriptions given by the virtual agent Max:

- *Entity*: This semantic category reflects whether or not the participant correctly specified a particular entity. The recalls were analyzed with respect to the following entities: 'church', 'tower', 'roof', 'clock', 'door', 'hedge', and 'tree'.
- *Shape*: This category reflects whether or not the participant correctly specified the shape of the particular entity.
- *Size*: This category reflects whether or not the participant correctly specified the size of the particular entity.
- *Relative Position*: This category reflects whether or not the participant correctly specified the relative position of the particular entity.

Presentation Quality To investigate how users perceive the quality of the overall presentation, and the agent's use of speech-accompanying gestures in particular, the following items were chosen to be judged on seven-point Likert scales:

- *Gesture Quantity*
  The amount of gestures was... [too few—too many]
- Gesture Quality, following (Hartmann et al., 2006):
  - *Spatial Extent*
    The spatial extent of gestures was ... [too small—too large]
  - *Temporal Extent*
    The temporal extent of gestures was ... [too slow—too fast]
  - *fluidity*
    Gesture execution was fluid. [not fluid—very fluid]
  - *Power*
    The execution of hand and arm movements was ... [weak—powerful]
- *Eloquence*
  [Max is not eloquent—Max is eloquent]
- *Overall Comprehension*
  The description given by Max was ... [not comprehensible—easily comprehensible]

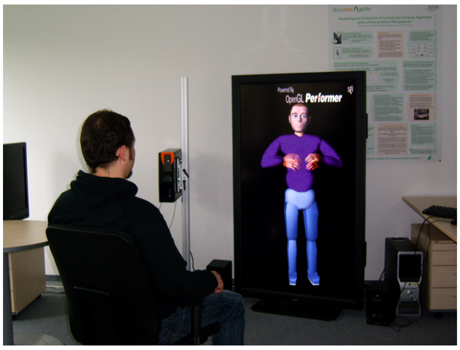


Figure 7.1: Set-up of the stimulus presentation phase.

- *Gesture’s Helpfulness for Comprehension*
  for the overall understanding gestures were ... [not helpful—very helpful]
- *Vividness of the agent’s mental image*
  Max had a vivid mental image from the things he described ... [not vivid—vivid]

Agent Perception Finally, participants had to report their perception of the virtual agent’s personality. To this end, 18 items were chosen (Fiske et al., 2006; Hoffmann et al., 2009): *active*, *affable*, *approachable*, *dedicated*, *expert*, *friendly*, *fun-loving*, *helpful*, *humanlike*, *intelligent*, *likeable*, *lively*, *organized*, *pleasant*, *sensitive*, *sociable*, *thorough*, *trustworthy* (translated from German) to assess the degree to which they attributed them to Max using a 7-point Likert scale.

Participants

A total of 110 participants (22 in each condition), aged from 16 to 60 years (M = 23.85, SD = 6.62), took part in the study. 44 participants were female and 66 were male. All of them were recruited at Bielefeld University and received 3 Euro for participating.

Procedure

Participants were instructed to carefully watch the presentation given by the virtual agent Max in order to be able to answer questions regarding content and subjective evaluation of the presentation afterwards. Figure 7.1 shows the setup used for stimulus presentation: Max was displayed on a 80 x 143 cm screen and thus appeared in life-size of 1.25 m. Life-sized projections have been shown to yield visual attention and fixation behavior towards gestures that is similar to behavior in face-to-face interactions (Gullberg and Holmqvist, 2006). Participants were seated 170 cm away from the screen and their heads were approximately leveled with Max’ head.

They were randomly assigned to one of the five conditions. The object description given by Max was preceded by a short introduction: Max introduced himself and repeated the instruction already given by the experimenter to allow participants to get used to the speech synthesis. The subsequent object description was always six sentences long and took 45 seconds. Each sentence was followed by a pause of three seconds. Participants were left alone for the stimulus presentation, and after receiving the questionnaire to complete it (neither experimenter nor Max were present during completion). The questionnaire consisted of the following parts:

1. Information Uptake
   - Written recall
     - General summary of the description
     - Detailed recall of objects and their properties
   - Graphical recall
2. Presentation quality
3. Agent perception

7.2.2 Analysis

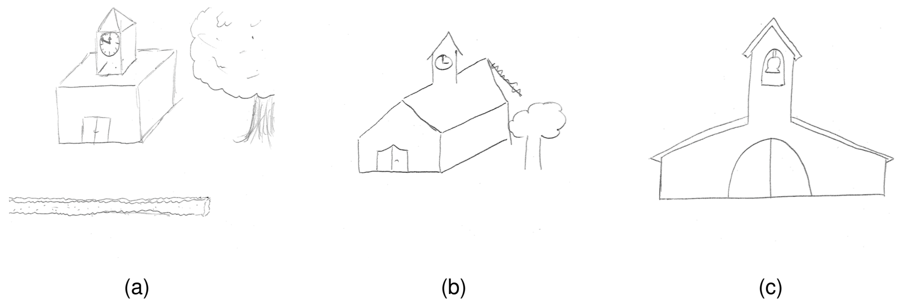
Scoring of semantic features in stimuli To investigate the amount of information participants take up from a given presentation, it is necessary to estimate the amount of information actually present in the stimuli. For this purpose two independent coders judged all five stimulus presentations with respect to the semantic features (SFs) communicated. The agreement was measured using the chance-based coefficients AC_1 (Gwet, 2001) and Cohen's Kappa (Cohen, 1960). Results, given in Table 7.4 show a high degree of inter-rater agreement (cf. Artstein and Poesio, 2008). The few discrepancies that did emerge were resolved through discussion.

Table 7.4: Agreement coefficients for scoring semantic features in the stimulus presentations.

		AC_1	Kappa
Experimental Conditions	GNetIc *ind-1*	0.96	0.94
	GNetIc *ind-2*	0.95	0.95
	GNetIc *combined*	0.86	0.84
	no gesture	1.00	1.00
	random gestures	0.95	0.95
Gestural SFs		0.90	0.84
Verbal SFs		1.00	1.00

Scoring of Written and Graphical Recall Subject of the analysis to be carried out is the accuracy of the participants' recall with respect to semantic features. Critical to this analysis is what is judged as a correct answer. This was accomplished by scoring all questionnaires independently by two raters who applied the following rules for written recall: the participant's answers relating to the semantic feature *entity* were scored as correct when the entity was either named with the label that was used in the stimulus presentation or with an apparent synonym (e.g., 'door' (German: 'Tür') and 'gate' (German: 'Tor')). Answers about *shape*, *size*, and *relative position* were judged as correct when it was explicitly stated that a particular entity was, e.g., 'tapered' (German: 'spitz') or 'round' (German: 'rund'). Again, apparent synonyms and paraphrases were also evaluated as correct (e.g. 'low' (German: 'niedrig') and 'small' (German: 'klein')). Answers, however, that contained only a general judgement, such as 'the tree is standing next to the church' (German: 'der Baum steht neben der Kirche') instead of 'right of the church' were not considered as correct answers due to a lack of precision in the participant's recall.

The same measures were employed when evaluating graphical recall. The semantic feature *entity* was only scored as correct when the entity was obviously present in the drawing. Semantic features *shape* and *relative position* were judged as correct when the particular property was apparently identifiable. Take Figure 7.2 as an example: the drawing in Figure 7.2 (a) the semantic feature 'squared' was judged as being present, whereas the drawing in Figure 7.2 (b) was evaluated as not containing the feature 'squared'. In figure 7.2 (c) it is not identifiable if the participant actually got that the church has a squared ground, and it is, therefore, not scored as being correct regarding the semantic feature *shape*.



(a) (b) (c)

Figure 7.2: Three examples of graphical recall in the evaluation of information uptake.

Both written and graphical recalls were scored independently by two judges. The overall reliability between the two judgers was found to be AC_1=0.94 and *Kappa*=0.94, showing a high degree of interrater agreement (cf. Landis and Koch, 1977), thus providing a solid basis for further analyses of the data. There are no obvious differences between (1) the different experimental conditions, (2) the types of recall, and (3) the types of semantic features as summarized in Table 7.5. Further, there is no obvious difference between AC_1 and Kappa coefficient for any of the features.

Table 7.5: Agreement coefficients for scoring semantic features from written and graphical recall.

		AC_1	Kappa
Experimental Conditions	GNetIc *ind-1*	0.93	0.93
	GNetIc *ind-2*	0.96	0.96
	GNetIc *combined*	0.93	0.93
	no gesture	0.94	0.94
	random gestures	0.95	0.94
Type of recall	Overall summary	0.94	0.94
	Verbal recall	0.95	0.95
	Salient objects/properties	0.97	0.91
	Graphical recall	0.94	0.90
Overall		0.94	0.94

7.2.3 Results—Information Uptake

To test the effect of experimental conditions on the dependent variables, analyses of univariate variance (ANOVA) and paired-sample *t*-tests for pairwise comparisons between condition means were conducted. In the following, results are reported with respect to the effect of experimental conditions on the accuracy proportion participants achieved in verbal and graphical recall.

General Recall Table 7.6 summarizes the mean scores and standard deviations. The mean of the accuracy proportion for verbal recall was 53.6%. There was no significant main effect for experimental conditions, however, there was a trend indicating that accuracy proportions are highest in the *no gesture* condition (60.84%), whereas lowest accuracy values were observed in the *random* (48.83%) and *combined* (49.06%) conditions. For graphical recall the mean of the overall recall proportion was 77.15%. Experimental conditions had no significant main effect on graphical recall, but recall accuracy in the individual GNetIc conditions and the *no gesture* condition were by trend higher than in the *combined* GNetIc condition and in the *random* condition.

Semantic Categories The different semantic categories employed in the analysis were then considered separately. In general, it turned out that—for verbal recall—the mean accuracy proportions differed obviously for the different semantic categories. Accuracy proportions ranged from 21.82% for the semantic feature *color*, up to 81.43% for the semantic feature *entity*. Analyses of univariate variance revealed that for only one of the semantic categories, namely *shape*, there was a main effect of experimental conditions ($f(4,105)=25.07$, $p<.001$). This was due to the fact that the mean accuracy proportions for both control conditions significantly outperformed the accuracy in the three GNetIc conditions: *no gesture/ind-1*: $t(42)=8.29$, $p<.001$; *no gesture/ind-2*: $t(42)=9.64$, $p<.001$; *no gesture/combined*: $t(42)=8.05$, $p<.001$; *random/ind-1*: $t(38)=4.63$, $p<.001$; *random/ind-2*: $t(36)=5.19$, $p<.001$; *random/combined*: $t(42)=5.27$, $p<.001$. For the other semantic categories *entity*, *relative position*, *size*, and *color* no significant main effect was found for experimental conditions in verbal recall.

In addition there was a tendency, although not significant, for the SF categories *entity* and *color* to be better recalled in the three GNetIc conditions as compared to the control conditions (means of 84% vs. 77.5% for *entity*; 27.3% vs. 13.5% for *size*). In both SF cases, the two individual GNetIc conditions received slightly higher recall accuracy than the *combined* GnetIc condition.

A separate analysis of the semantic categories for graphical recall revealed that entities were recalled with highest accuracy (M=83.3%) while the mean recall accuracy for *shape* information was 71.5% and for *relative position* 72.3%. There was no significant main effect for experimental conditions.

Table 7.6: Mean accuracy proportions of semantic features in the five experimental conditions (standard deviations in parentheses).

		Entity	*RelPos*	*Size*	*Shape*	*Color*	Overall
Verbal recall	*ind-1*	.86 (.14)	.64 (.32)	.35 (.27)	.60 (.18)	.27 (.46)	.56 (.17)
	ind-2	.85 (.17)	.60 (.29)	.34 (.25)	.58 (.16)	.32 (.48)	.54 (.17)
	combined	.81 (.18)	.61 (.24)	.30 (.24)	.51 (.25)	.23 (.43)	.49 (.18)
	no gestures	.79 (.17)	.72 (.26)	.38 (.32)	.98 (.11)	.18 (.39)	.61 (.16)
	random gestures	.76 (.18)	.51 (.27)	.17 (.16)	.91 (.25)	.09 (.29)	.49 (.13)
	Overall	.81 (.17)	.62 (.28)	.31 (.26)	.72 (.27)	.22 (.41)	.54 (.16)
Graphical recall	*ind-1*	.88 (.12)	.76 (.22)		.75 (.20)		.73 (.30)
	ind-2	.86 (.15)	.81 (.24)		.71 (.25)		.75 (.26)
	combined	.79 (.19)	.72 (.27)		.61 (.32)		.61 (.38)
	no gestures	.86 (.16)	.73 (.23)		.80 (.25)		.80 (.25)
	random gestures	.77 (.16)	.60 (.27)		.70 (.25)		.70 (.25)
	Overall	.83 (.16)	.72 (.25)		.72 (.26)		.72 (.29)

Table 7.7: Mean proportions of semantic features that were correctly recalled in percent (standard deviations in parentheses).

		Redundant SFs	Complementary SFs
Verbal recall	*ind-1*	.57 (.22)	.27 (.34)
	ind-2	.82 (.22)	.20 (.30)
	combined	.52 (.19)	.25 (.37)
	random gestures	.68 (.24)	—
	Overall	.65 (.24)	.24 (.33)
Graphical recall	*ind-1*	.77 (.23)	.77 (.30)
	ind-2	.74 (.26)	.67 (.40)
	combined	.65 (.24)	.61 (.41)
	random gestures	.48 (.27)	—
	Overall	.66 (.27)	.68 (.37)

Relation of Speech and Gestures Next, the role of information distribution across modalities was considered. That is, semantic features communicated by both speech and gesture (redundant SFs) were contrasted with semantic features communicated only gesturally (complementary SFs). For verbal recall, participants' information uptake differed considerably between redundantly and complementarily communicated SFs. The mean accuracy proportion for complementary SFs was 24.24%, whereas it was 64.58% for redundant SFs. For the latter there was a main effect for experimental conditions ($f(3,84)=8.35$, $p<.001$). Results of t-tests showed significant mean differences between one of the individual GNetIc conditions, namely *ind-2*, and the other conditions: *ind-2*/*ind-1*: $t(42)=3.80$, $p<.001$; *ind-2*/*combined*: $t(42)=3.80$, $p<.001$; *ind-2*/*random*: $t(42)=1.96$, $p=.057$. In addition, the recall of redundant SFs in the *random* condition significantly outperformed the recall in the *combined* GNetIc condition: *random*/*combined*: $t(40)=-2.54$, $p=.015$.

For graphical recall—notably—there was no such difference between redundant and complementary SFs. Here, the proportion of correctly recalled complementary SFs was 68.0% and therefore almost identical to the graphical recall proportion of redundant SFs.

Serial Position Since it was notable that redundant features were communicated particularly successfully in the *random* condition, the position of semantic features in the sequential presentations was further taken into consideration. Actually, the redundant features in the *random gestures* presentation were predominantly present in the first half of the description. In contrast, in the three GNetIc conditions, redundant features were distributed over the complete presentation. The same holds for verbal recall results from the SF-based analysis: here the SF *shape* was recalled most

Table 7.8: Mean proportions of semantic features that were correctly recalled in percent (standard deviations in parentheses).

		First SFs	Middle SFs	Last SFs
Verbal recall	*ind-1*	.73 (.15)	.66 (.16)	.52 (.27)
	ind-2	.75 (.15)	.58 (.19)	.55 (.28)
	combined	.65 (.17)	.58 (.19)	.47 (.26)
	no gestures	.74 (.14)	.72 (.26)	.54 (.26)
	random gestures	.69 (.12)	.68 (.27)	.32 (.23)
	Overall	.71 (.15)	.65 (.22)	.48 (.27)
Graphical recall	*ind-1*	.82 (.21)	.88 (.15)	.69 (.34)
	ind-2	.83 (.18)	.82 (.23)	.76 (.34)
	combined	.84 (.15)	.73 (.29)	.60 (.38)
	no gestures	.86 (.15)	.87 (.22)	.67 (.30)
	random gestures	.78 (.21)	.83 (.23)	.49 (.31)
	Overall	.83 (.18)	.83 (.23)	.64 (.34)

successfully in the two control conditions in which *shape* features were presented at the beginning of the description, while *shape* features were distributed over the complete description in the GNetIc conditions. This is why the position of gesturally communicated semantic features in the sequence of speech-gesture utterances was further considered comparing semantic features presented (1) in early (sentences 1—2), (2) in the middle (sentences 3—5), and (3) at the end (sentences 6—7) of the description.

The analysis revealed that there was a considerable difference between the three classes of SF position with regard to verbal recall accuracy: SFs presented at the beginning of the description were recalled with the highest accuracy rate of 73.3%. SFs communicated in the middle part of the description were recalled at a rate of 64.5%, whereas SFs from the end of the presentation only had a mean recall proportion of 48.1%. Analyses of univariate variance revealed that only for the latter there was a main effect for experimental conditions ($f(2,326)=2.87$, $p=.03$). Results of pairwise t-tests showed that this effect was due to the fact that recall accuracy in the *random* condition was significantly lower in comparison with the *no gesture* and the two individual GNetIc conditions: *random/no gesture*: $t(41)=-2.94$, $p=.01$; *random/ind-2*: $t(40)=-2.99$, $p=.01$; *random/ind-1*: $t(39)=-2.63$, $p=.01$.

Graphical recall was also decreased for SFs occuring at later positions in the presentations. In contrast to verbal recall, there was no difference between SFs in the first and the middle part of the description. A main effect of experimental conditions was not found for SF positions with regard to graphical recall accuracy.

Table 7.9: Mean values for the dependent variables of presentation quality in the five conditions (standard deviations in parentheses).

	ind-1	*ind-2*	*combined*	*no gestures*	*random*
Gesture Quantity	3.91 (1.15)	3.95 (0.95)	3.59 (0.91)	2.48 (1.21)	3.55 (1.22)
Spatial Extent	3.77 (0.87)	4.14 (0.83)	3.59 (1.05)	–	3.55 (1.05)
Temporal Extent	3.68 (0.83)	3.64 (0.66)	3.50 (1.01)	–	3.30 (0.87)
Fluidity	4.09 (1.48)	4.00 (1.57)	3.05 (1.32)	–	3.65 (1.53)
Power	3.59 (1.10)	4.09 (1.27)	3.91 (1.38)	–	3.90 (1.48)
Eloquence	3.50 (1.74)	4.91 (1.14)	3.05 (1.46)	3.69 (1.11)	3.25 (1.61)
Comprehension	5.18 (1.33)	5.27 (1.16)	4.68 (1.49)	4.95 (1.32)	4.18 (1.37)
Gestures helpful	5.68 (1.56)	5.82 (0.85)	4.70 (1.62)	1.82 (1.14)	4.10 (2.05)
Vividness	5.32 (1.62)	5.45 (1.13)	4.18 (1.81)	4.08 (1.32)	3.81 (1.80)

7.2.4 Results—Quality of Presentation

The perceived quality of presentation was investigated with regard to gestures, speech, and content. Participants were asked to evaluate each variable on a seven-point Likert scale. To test the effect of experimental conditions on the dependent variables, analyses of univariate variance (ANOVA) were conducted and paired-sample *t*-tests with 95% confidence intervals (CI) for these pairwise comparisons between condition means. Mean values and standard deviations are summarized in Table 7.9 and visualized in figure 7.3 for dependent variables with significant main effects.

Gesture Quantity With regard to gesture quantity, the overall mean value for the four gesture conditions was M=3.75 (SD=1.06) on a seven-point Likert scale (too few—too many). There was no significant main effect for experimental conditions. That is, participants were quite satisfied with the gesture rate. For the *no gesture* condition, participants rated gesture quantity as rather too low (M=2.48, SD=1.21).

Gesture Quality No main effect for experimental conditions was obtained for the four attributes characterizing gesture quality: spatial extent (too small—too large, M=3.77, SD=0.97), temporal extent (too slow—too fast, M=3.53, SD=0.85), fluidity (not fluid—very fluid, M=3.70, SD=1.51), and power (weak—powerful, M=3.87, SD=1.30). In all four gesture conditions the four quality attributes were rated with mean values between 3.0 and 4.0 on a seven-point Likert scale.

Eloquence With regard to perceived eloquence of the virtual agent (Max is not eloquent—Max is eloquent), there was a significant main effect (*f*(4,79)=3.12, *p*=.02). This is due to the fact that the mean of condition *ind-2* differed from all other conditions (*ind-2*/*no gesture*: *t*(21)=2.64, *p*=.02, CI=[0.26;2.17]; *ind-2*/*random*: *t*(25)=2.94, *p*=.01, CI=[0.50;2.82]; *ind-2/combined*: *t*(25)=4.02, *p*=.001, CI=[0.91;2.82]; *ind-2/ind-1*: *t*(31)=2.43, *p*=.02, CI=[0.23;2.59]). That is, gestures produced with a suitable individual gesture network have the potential to increase the perceived eloquence (recall that the verbal explanations were identical in all conditions).

Overall Comprehension Another variable of interest was the comprehensibility of the overall description (not comprehensible—easily comprehensible). Although the ANOVA marginally failed to reach significance (*f*(4,105)=2.37, *p*=.057), simple effects for experimental conditions were analyzed. The means for both individual GNetIc conditions significantly outperformed the mean of the *random* gesture condition (*ind-1/random*: *t*(42)=2.46, *p*=.018, CI=[0.18; 1.82]; *ind-2/random*: *t*(41)=2.85, *p*=.007, CI= [0.32;1.86]). By trend, the *no gesture* mean differed from the *random* mean. That is, participants reported greater comprehension of the presentation when the agent produced no, rather than random gestures.

Gesture's Helpfulness for Comprehension With regard to perceived helpfulness of gesturing (not helpful—very helpful) a significant main effect (*f*(4,104)= 25.86, p<.001) was obtained. Not surprisingly, participants in the *no gesture* condition rated gesturing as less helpful than participants in the other conditions (*t*-test, p<.001 in each case). In addition, gestures in both individual conditions (*ind-1, ind-2*) were rated more helpful than in the *random* condition (*ind-1*: *t*(41)=2.87, *p*=.006, CI=[0.47;2.70]; *ind-2*: *t*(41)=3.63, *p*=.001, CI=[0.77;2.68]).

Vividness Furthermore, participants were asked to rate the vividness of the agent's conception of the presented content (not vivid—vivid). Random gesturing tended to hamper this impression even more than no gesturing and combined gesturing. Furthermore, the ANOVA revealed a significant main effect (*f*(4,79)=3.50, *p*=.01). Results of *t*-tests showed significant mean differences between both individual GNetIc conditions and the other three conditions (*ind-1/no gesture*: *t*(29)=2.47, *p*=.02, CI=[0.16;2.32]; *ind-1/random gestures*: *t*(30)=2.66; *p*=.01, CI=[0.38;2.63]; *ind-1/combined*: *t*(41)=2.19, *p*=.03, CI=[0.09;2.18]; *ind-2/no gesture*: *t*(22)=2.76, *p*=.01, CI=[0.33;2.43]; *ind-2/random gestures*: *t*(25)=2.91, *p*=.01, CI=[0.38;2.90]; *ind-2/combined*: *t*(31)=2.12, *p*=.04, CI=[0.05; 2.50]). That is, producing gestures with an individualized network helps a virtual agent to create the impression in human recipients of having a better idea of what is being described.

7.2.5 Results—Perception of the Virtual Agent

How Max was perceived was assessed using several items, such as 'pleasant', 'friendly', 'helpful' on seven-point Likert scales (not appropriate—very appropriate). To measure the reliability of these items, they were grouped into three scales 'likeability', 'compe-

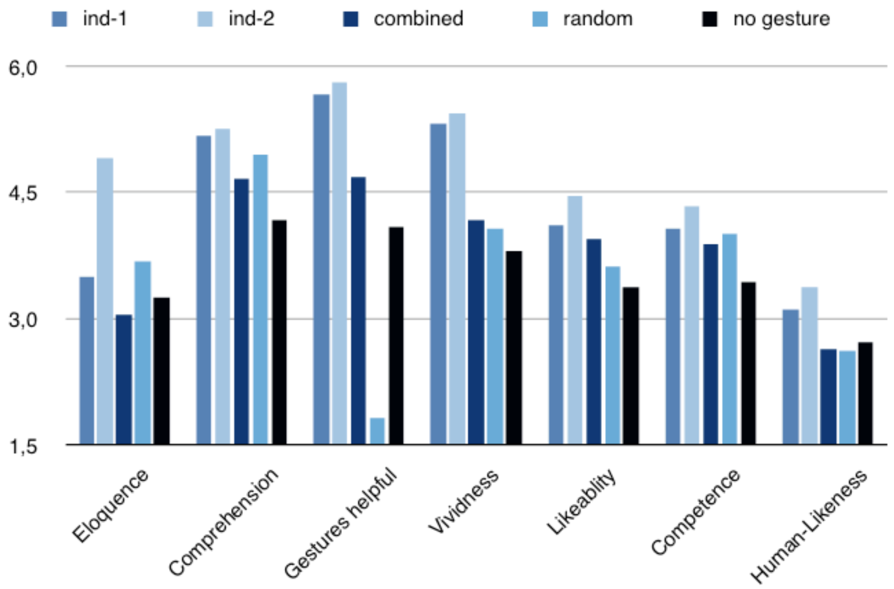


Figure 7.3: Mean values of the dependent variables in the five conditions (see Tables 7.9 and 7.11 for exact values and SDs).

tence', and 'human-likeness' (see Table 7.10), and Cronbach's alpha for the indeces was calculated. Alpha values for all three scales were above 0.7, which justifies combining these items into one mean value as a single index for this scale. The main effect for experimental conditions was analyzed by applying ANOVAs and the pattern of means further investigated by computing paired-samples *t*-tests with 95% confidence intervals (CI) for pairwise comparisons between condition means. Mean values and standard deviations are summarized in Table 7.11 and visualized in figure 7.3.

Table 7.10: Reliability analysis for the three scales 'likeability', 'competence', and 'human-likeness'.

Scale	Items	Cronbach's Alpha
Likeability	pleasant, sensitive, friendly, likeable, affable, approachable, sociable	.86
Competence	dedicated, trustworthy, thorough, helpful, intelligent, organized, expert	.84
Human-likeness	active, humanlike, fun-loving, lively	.79

Likeability Regarding likeability, a significant main effect for experimental conditions ($f(4,104)$=3.88, p=.01) was found. Mean ratings for the two individual GNetIc

Table 7.11: Mean values for the agent perception scales in the five different conditions (standard deviations in parentheses).

	ind-1	*ind-2*	*combined*	*no gestures*	*random*
Likeability	4.12 (1.18)	4.47 (0.81)	3.95 (0.87)	3.62 (1.24)	3.39 (1.14)
Competence	4.07 (1.11)	4.34 (0.55)	3.89 (0.84)	4.01 (1.09)	3.44 (1.07)
Humanlikeness	3.11 (1.29)	3.38 (1.07)	2.64 (1.01)	2.62 (1.00)	2.73 (0.98)

conditions were higher than in the other conditions. In particular, this relationship was significant when comparing the *ind-2* condition with *no gesture* (t(36)=2.68, p=.01, CI=[0.21;1.48]) and *random* conditions (t(38)=3.58, p=.001, CI=[0.47;1.67]). The mean difference between *ind-2* and the *combined* condition marginally failed to reach significance (t(40)=1.99, p=.054; CI=[-0.01;1.02]). For individual condition *ind-1*, the difference of mean evaluation of likeability in comparison with *random* gestures was significant (t(42)=2.08, p=.05, CI=[0.02;1.43]). In addition, means for the *combined* GNetIc condition were higher than in both control conditions. In other words, all three GNetIc conditions outperformed the control conditions, whereby best evaluations for likeability were obtained by participants in the individual GNetIc conditions.

Competence With regard to the evaluation of the agent's competence, a significant main effect (f(4,101)=2.65, p=.04) was also found. The GNetIc condition *ind-2* received higher mean evaluations than the *random* condition (t(42)=3.51, p=.001, CI=[0.38;1.42]). The *combined* GNetIc condition also received a higher mean evaluation than the *random* condition which is, however, not significant. Notably, there were no significant differences between the GNetIc conditions and the *no gesture* condition.

Human-likeness Finally, the analysis of ratings for human-likeness revealed a main effect (f(4,104)=2.08, p=.09). Both individual GNetIc conditions outperformed the other conditions. Again, this relationship is stronger for the condition *ind-2* (*ind-2*/*no gesture*: t(42)=2.40, p=.02, CI=[0.12;1.38]; *ind-2*/*random gestures*: t(42)= 2.09, p=.04, CI=[0.02;1.27]; *ind-2*/combined: t(41)=2.30, p=.03, CI=[0.09;1.38]). For the other individual GNetIc condition *ind-1*, the mean rating of human-likeness is also higher than in the *combined* GNetIc condition and the two control conditions, but these differences are not significant. No difference was found between the *combined* GNetIc condition and the two control conditions (*random* and *no gesturing*).

7.3 Summary and Discussion

This chapter evaluated the GNetIc generation results in two ways, according to the two-fold research objective pursued in this thesis.

7.3.1 Prediction-based Evaluation

Results of the prediction-based evaluation study can be summarized in four points. First and foremost, the analysis of the prediction accuracy of the GNetIc model as compared to the corpus data yielded very promising results. Both chance node and decision node evaluations provided substantial accuracy values, clearly above the chance level baseline for all production choices in all conditions under investigation. Given that natural gesture use in humans is often sloppy and the fact that in a particular situation a large number of gestures may be appropriate, as can be seen from the obvious inter-individual differences in gesturing behavior, one cannot expect the model to precisely predict every single gesture feature accurately. The achieved prediction values are, however, a positive indication for the appropriateness of the selected modeling methodology. Nevertheless, there is still a potential for improvement making use of the iterative character of the cyclic design methodology applied in this thesis. Ideas for advancing the GNetIc model will be sketched in Section 8. Any extension of the model can easily be evaluated in comparison relative to the results achieved so far.

Second, the comparison of different learning techniques showed that prediction accuracy of constraint-based algorithms was higher than the prediction accuracy of score-based algorithms. Therefore, the application of constraint-based methods is to be preferred since they provide several advantages (Section 5.3.2). In particular, as a local measure and by providing the possibility to vary significance levels, they are of higher value to discover mechanisms and principles of the underlying data, and allow for elucidating iconic gesture production in humans.

Third, the comparison of networks learned from either individual or combined data sets revealed an advantage for individual data. This holds for the results of all four learning methods. That is, the effect of data heterogeneity seems to be stronger than overfitting tendencies (cf. Section 5.4.2). Accordingly, individual data sets are to be preferred when aiming to reproduce gestural behavior.

Finally, accuracy results for both probabilistic and rule-based decision making yielded similar values. That is, neither of the two modeling methods provided by BDNs was found to be superior with regard to prediction accuracy. The application of the hybrid method, therefore, is shown to be reasonable also from the perspective of predication-based accuracy.

7.3.2 Perception-based Evaluation

The perception-based evaluation comprised two major aspects, namely the *communicative* role of gesture use in terms of information uptake and the role of gestures in the *subjective impression* of the interaction in terms of presentation quality and perception of the virtual agent.

Information Uptake

The purpose of the analysis of information uptake was to test whether human addressees pick up semantic information from a virtual agent's communicative behavior, and whether there are any characteristics of speech and gesture use modulating the uptake of gesture information. With regard to information uptake from gestural behavior of virtual agents the present analysis goes beyond the study carried out by Krämer et al. (2003) and Buisine and Martin (2007) in that it employs a micro-analytic feature-based methodology. Results can be summarized in four major points.

First, it is remarkable that the difference across experimental conditions with regard to information uptake in general was only marginal. The *no gesture* condition, actually, received the highest proportion of correctly recalled semantic features. That is, gesture use, in whatever condition, did not increase participants' recall accuracy. For the *random* gesture condition this is not surprising at all, as one could argue that the ununsual way of gesture use, becoming apparent for instance in gesture occurrence with the sentence's theme (see Figure 7.3), led to confusion and, therefore, reduced recall accuracy. The lower recall accuracy in the GNetIc conditions, however, is more difficult to explain, in particular, since most of the gestures were redundant with speech (exceptions are only the window and the door gestures in the GNetIc conditions) and redundant gestures are prevalently seen as supplementing the verbal communication channel. A reason for these findings might be that the gesturing agent or the integration of speech and gestures during perception and understanding needs cognitive resources, resulting in impaired capacity for memorization. Further research addressing this issue is definitelt required.

Second, there was a discrepancy between verbal and graphical recall. As also reported in Buisine and Martin (2007), accuracy scores were higher in graphical recall as compared to written recall. A likely explanation for this finding is that prototypical information about particular objects is implied when drawing the object which was described. For instance, a majority of church doors might be arc-shaped or most church clocks round. So it seems as if participants combined information recalled from the description with their previous knowledge of similar objects which became apparent in the graphical recall situation where participants were 'forced' to give those objects a shape. That is, verbal recall seems to be a more direct or explicit measure with regard to question how much information users really pick up from the agent's communicative behavior.

Third, and related to the previous point, there was a crucial difference between redundant and complementary gestural information in verbal recal, but not in graphical recall: redundancy yielded a slight increase of verbal recall accuracy in contrast to the *no gesture* control condition (65% vs. 61%). By contrast, complementarity information yielded a significant decrease of verbal recall in contrast to the control condition (24% vs. 61%). It is therefore concludable that information encoded redundantly in speech and gestures tended to help the users to memorize and encode the information verbally (i.e., explicitly), while information that was only present in gestures was not recalled very well. With regard to the slightly beneficial role of redundant gestures, it is remarkable that there was relatively much variation across conditions. For instance, in the GNetIc *ind-2* condition, redundancy led to a recall proportion of even 82%. That is, a reasonable hypothesis is that the kind of gesture use is decisive for the degree of recall ability.

Another reason for the low recall accuracy of complementary gestures is that complementary gestures were used towards the end of presentations and, in general, recall accuracy was decreased for semantic information given at the end of the presentations—contrary to the 'recency effect' (when asked to recall a list of items in any order, people tend to begin recall with the end of the list, recalling those items best (Murdock, 1962)). An explanation for the missing recency effect would be that the presentation of content was not independent of order. Rather, utterances built upon each other: the first sentence, for example, introduced the church, while the second sentence elaborated on the appearance of the church by introducing the church tower. The third sentence again elaborated on this by giving information about the tower's roof etc. This order of information was also reflected in participant's verbal recalls, providing evidence for the fact that order was important in recall. Further research is definitely needed at this point to identify the factors that make up successful communicative gestures.

Finally, addressees' information uptake accuracy was further found to be considerably different for the different semantic categories. This is not surprising, since similar differences between SF categories were also reported for information uptake from human gesturing (Beattie and Shovelton, 1999b, 2001). What is remarkable from the present study, however, is that there was a tendency for the SF categories *entity* and *color* to be better recalled in the three GNetIc conditions as compared to the control conditions. Both SF categories are not explicitly conveyed in gestures: for color it is not possible at all, and object identity was always depicted gesturally by specific properties such as shape, position, or size. That is, information uptake accuracy in the GNetIc conditions was improved for information that was not conveyed gesturally. An explanation could be that gestures, in general, helped to visualize and memorize information—not only restricted to the particular SFs explicitly depicted by those gestures.

Quality of Presentation and Perception of the Virtual Agent

Going beyond the purely communicative functions of gesture use, another goal of the perception-based evaluation study was to explore the user acceptance of the GNetIc-generated gestures, as well as to investigate how the virtual agent itself is judged by human users. Results can be summarized in five major points.

First, Max' gesturing behavior was rated positively regarding gesture quantity and quality, and no difference across gesture conditions was found concerning these issues. That is, building generative models of co-verbal gesture use can yield good results with actual users. The fact that gesture quality was rated more or less equally across conditions rules out the possibility that other effects of the experimental conditions were due to varying quality of gesture use and realization in the virtual agent.

Second, both individual GNetIc conditions outperformed the other conditions in that gestures were perceived as more helpful, overall comprehension of the presentation was rated higher, and the agent's mental image was judged as being more vivid. Similarly, the two individual GNetIc conditions outperformed the control conditions regarding agent perception in terms of likeability, competence, and human-likeness.

Third, the *combined* GNetIc condition, notably, was rated worse than the individual GNetIc conditions throughout. This finding underlines the important role of inter-individual differences in communicative behavior and implies that the common approach to inform behavior models from empirical data by averaging over a population of subjects is not the best choice.

Finally, the *no gesture* condition was rated more positively than the *random* condition, in particular for the subjective measures of overall comprehension, the gesture's role for comprehension, and vividness of the agent's mental image. That is, with regard to these aspects it seems even better to make no gestures than to generate random, though still considerably iconic gestural behavior.

Overall, individualized gesturing was strikingly beneficial with regard to how virtual agents and their communicative skills are judged by human users, suggesting that modeling individual speakers with proper abilities for the target behavior results in even better be behavior judged from the perspective of human interaction partners. This may be due to the fact that individual networks ensure a greater coherence of the produced behavior. As a consequence, the agent may appear more coherent and self-consistent which, in turn, may make its behavior more predictable and easier to interpret for the user. This is in line with Nass et al. (2000), who found that people like virtual agents better when they show consistent personality characteristics across modalities. On the contrary, however, Foster and Oberlander (2007) recently argued for more variation in the generation of non-verbal behavior based on evidence from the evaluation of automatically produced head and eyebrow motion. Since the two individual GNetIc conditions were not judged equally well (*ind-2* outperformed *ind-1*), it seems reasonable, in any case, to detect particularly appropriate speakers and to individualize agents according to them. Further research is needed to identify the characteristics of ‘successful’ gesture use in detail. Given the close relation of gestures and speech it is, for instance, supposable that the relation of gestures and speech also plays a role. To make the experimental conditions comparable, the verbal part of the communicative behavior was identical across all conditions. It might be, however, that combination with automatically generated verbal constructions would have been more appropriate, e.g., in the *ind-2*-condition.

CHAPTER 8

Conclusion

The goal of this thesis was to develop a computational simulation model for the production of speech-accompanying iconic gestures to be realized in virtual agents. The rationale behind this objective was to devise and probe a predictive model of gesture use in order to gain insight into human gesture production, and thereby to improve human-agent interaction such that it progresses towards intuitive and human-like communication.

8.1 Summary of Results

As a starting point, evidence that iconic gestures are shaped not only by iconicity, but also by contextual constraints as well as inter-individual differences was collected. Existing models of gesture production—theoretical and computational—were reviewed and discussed with respect to fundamental design choices concerning these factors. The conclusion was that none of those models provides a comprehensive account of how gestures are produced.

Empirical Results As empirical basis for the generation model, the SaGA corpus provided an extensive data collection of communicative behavior for a spatial task. The statistical data analysis revealed the following results:

- Gesture use was found to be influenced by multiple factors: characteristics of the referent, the linguistic and discourse-contextual situation, as well as the speakers' previous gestural behavior.
- The corpus analysis revealed novel insights with respect to the sub-classification of iconic gestures into gestural representation techniques: gestures of these techniques were shown to differ significantly from each other with regard to the employed form features. They were, further, found to be structured by

representation technique-specific constraints, while at the same time also being sensitive to characteristics of the referent as well as inter-individual differences.

- There were obvious inter-individual differences found, both at the surface of gestural behavior and also in how strong particular influencing relations were.

Modeling Results Based on the empirical insights, the *Generation Network for Iconic Gestures* (GNetIc) was developed—a computational simulation model for the production of speech-accompanying iconic gestures:

- The model combines data-driven and rule-based decision making to account for both inter-individual differences in gesture use, as well as patterns of representation technique-specific form-meaning mappings.
- The physical appearance of generated gestures is influenced by multiple factors: characteristic features of the referent accounting for iconicity, as well as contextual factors like the given communicative goal, information state, or previous gesture use.
- Learning gesture networks from individual speaker's data gives an easily interpretable visual image of preferences and strategies in composing gestures and makes them available to generate novel gesture forms in the style of the respective speaker.

GNetIc models were brought to application in an overall architecture for integrated speech and gesture generation. Being equipped with proper knowledge sources, i.e., communicative plans, lexicon, grammar, propositional, and imagistic knowledge, a virtual agent was enabled to autonomously explain buildings of a virtual environment using speech and gestures. By switching between the respective decision networks, the system has the ability to simulate speaker-specific gesture use.

Evaluation Results Accounting for the two-fold rationale followed in this thesis, the GNetIc model was evaluated in two ways:

- In comparison with empirically observed gestural behavior, the model was shown to be able to successfully approximate human use of iconic gestures. Individualized models yielded a slightly better prediction accuracy than 'average' models learned from the combined data of several speakers.
- When brought to application in a virtual agent, individualized GNetIc-generated gestures were found to increase the perceived quality of object descriptions as judged by human recipients. Moreover, the agent was rated more positively in terms of likeability, competence, and human-likeness. Individualized gesture use was, further, shown to be slightly helpful for human users' ability to generally memorize the objects which were described.

8.2 Implications

Looking back at the point from where this thesis started— with its widely open question of why different gestures take the particular physical form they actually do—the results of this work provide first steps towards a more thorough understanding of iconic gesture production in humans and also on how gesture use may improve human-agent interaction. Implications for both research areas will be sketched in the following.

Implications for Research on Communicative Behavior From the point of view of gesture research, the results show that computational modeling with virtual agents is a highly valuable tool to discover mechanisms and principles of gestural behavior. The analysis of modelling results actually revealed novel insights into the production process of iconic gestures: inter-individual differences in gesture use were shown to be present not only in the overt gestures, but also in the production process they originate from. That is, a set of different gesture generation strategies seems to exist from which individuals typically apply a particular subset. The major conclusion to be taken from this is that inter-subjective differences are a key factor with regard to understanding the mechanisms underlying gesture production in humans. This is a fact that has neither been considered in most existing theoretical/psycholinguistic nor in computational models of gesture production. As a result, although existing production models do not provide a comprehensive account of gesture production in human speakers, they may provide increments towards the set of factors and processes involved.

The observation that individual speakers apply a subset of available generation patterns has also been made in research on the generation of referring expressions (Dale and Viethen, 2009; Viethen and Dale, 2010).The insight might, thus, have the potential of having even wider consequences for research on communicative behavior in humans. This view actually offers a new perspective for research on production processes of communicative behavior and implies further, exciting research questions and challenges: what is the range of possible production strategies? How are they combined in individual speakers? Are these combinations idiosyncratic or are there general ‘clusters’ of speakers sharing similar strategy sets? Are there any factors influencing this combination, e.g., personality traits or cognitive skills? Although not providing an answer to these questions, the GNetIc approach nonetheless provides a valuable instrument for their exploration.

Implications for Research on Virtual Agents From the point of view of virtual agent research, the results showed that automatically generated gestural behavior is actually beneficial with regard to the impact of virtual agents on human addressees. Notably, different models were found to result in noticable different behavior, with consistently differing perception and evaluation by human recipients. As a consequence, it seems reasonable to detect particularly appropriate speakers and to individualize a virtual agent's communicative behavior accordingly. This does, of course, raise the question concerning the characteristics of 'successful' gesture style. At this point, the potential of virtual agents comes to the fore as they provide the flexibility to turn on and off aspects of the behavior model to observe how human addressees respond.

Individualization of communicative behavior, however, bears the danger of narrowing acceptance down to a certain population of users, since gesture perception, like production, may be subject to inter-individual differences. For instance, Martin et al. (2007) found the rating of gestural expressivity parameters to be influenced by a human addressee's personality traits. Accordingly, an important lesson to be learned, therefore, concerns the role of evaluation studies as an integral part of the communicative behavior modelling process. While prediction accuracy is highly prized in many evaluations of behavior simulation, the impact of how humans perceive a virtual agent's expressive behavior should always be a major citerion to help producing adequate behavior and increase the acceptance of the agent. Do people, e.g., prefer an agent that simulates their own gesture style?

8.3 Outlook

The GNetIc approach presented in this thesis goes well beyond related work in computational gesture generation, which has either emphasized common patterns in the formation of iconic gestures or concentrated on the individualization of gesture use. Nevertheless, there is of course potential for improvements:

- The model is, so far, restricted to gestures used in object descriptions for simplified VR objects. The description of more realistic entities or other forms of gesture use, like verb-phrase aligned gestures, e.g., pantomime gestures or typical direction-giving gestures as in 'turn right', pose further challenges. With regard to the former, it is likely that further shape properties as well as other referent characteristics, e.g., orientation, have to be considered. The extension regarding verb- and action/motion-related gestures requires an extension of the represenation formalism on which generation is based, as well as the consideration of further gestural representation techniques.
- The GNetIc account, as presented here, focused on decision making with regard to important form aspects of gesture use. Further characteristics of gesture quality, not yet considered, include what Hartmann et al. (2006) called *expressivity parameters*: modulations in spatial or temporal extent, repetitions etc. These parameters were identified as also being subject to inter-individual differences.

Hence, there is still potential to improve the degree of how speaker-specific automatically GNetIc-generated gestures are. Instead of simply applying these parameters to readily planned gestures, the GNetIc account allows integrating them into the overall generation process. This would provide the advantage that those parameters can be placed in relation to other modulating factors. For instance, the spatial extent of a gesture may vary depending on, say, the information state.

- Gesture production was viewed from a speaker-internal perspective in this thesis. One crucial aspect that has not been considered accordingly is the fact that gestures are typically produced in face-to-face *dialogue* situations. There is, however, evidence that gesture use is additionally constrained by dialogue-related factors. One such factor is the visibility of the addressee. There is much evidence that speakers gesture at a higher rate when they can see the person they are talking to (e.g., Cohen and Harrison, 1973; Cohen, 1977; Krauss et al., 1995). Bavelas et al. (2008) additionally found that visibility significantly affects *how* speakers gesture: in face-to-face interaction, participants were more likely to make life-size gestures than in monologue or telephone situations. Kimbara (2006, p. 41) described the phenomenon of *gestural mimicry* in terms of a given speaker's gesture being contingent upon the gesture of an interlocutor that occurred in the previous discourse: "the form-meaning relationship of a given speaker's gesture appears to influence how an interlocutor's gesture is formed when the interlocutor refers to the same topic in subsequent discourse". The gesture generation model, therefore, awaits integration of these interactional and dialogue-related sensitivities of gesture use. The model developed so far lends itself to this purpose, as it is extensible with regard to both further factors as well as the possibility to incorporate functions judging the utility of single decisions or collections of decisions.

To conclude, the fundamental principles of the GNetIc model seem to be suited to deal with these potential enhancements. Accordingly, the work presented here can also serve as a valuable basis to investigate future research questions.

Bibliography

Abellán, J., Gómez-Olmedo, M., and Moral, S. (2006). Some variations on the PC algorithm. In *Proceedings of the 3rd European Workshop on Probabilistic Graphical Models*, pages 1–8.

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

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, 22:261–295.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34:555–596.

Baake, V. (2009). Salienzbasierte Inhaltsplanung für multimodale räumliche Beschreibungen. Master's thesis, Faculty of Technology, Bielefeld University.

Ball, G. and Breese, J. (2000). Emotion and personality in a conversational agent. In Cassell, J., Sullivan, J., Prevost, S., and Churchill, E., editors, *Embodied Conversational Agents*, pages 189–219. MIT Press, Cambridge, MA.

Bavelas, J., Chovil, N., Lawrie, D., and Wade, A. (1992). Interactive gestures. *Discourse Processes*, 15:469–491.

Bavelas, J., Gerwing, J., Sutton, C., and Prevost, D. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, 58:495–520.

Bavelas, J., Kenwood, C., Johnson, T., and Philips, B. (2002). An experimental study of when and how speakers use gestures to communicate. *Gesture*, 2(1):1–17.

Beattie, G. and Shovelton, H. (1999a). Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental investigation. *Semiotica*, 123:1–30.

Beattie, G. and Shovelton, H. (1999b). Mapping the range of information contained in the iconic hand gestures that accompany spontaneous speech. *Journal of Language and Social Psychology*, 18:438–462.

Beattie, G. and Shovelton, H. (2001). An experimental investigation of the role of different types of iconic gesture in communication: A semantic feature approach. *Gesture*, 1:129–149.

Beattie, G. and Shovelton, H. (2002). What properties of talk are associated with the generation of spontaneous iconic hand gestures? *British Journal of Psychology*, 41:403–417.

Bergmann, K., Damm, O., Fröhlich, C., Hahn, F., Kopp, S., Lücking, A., Rieser, H., and Thomas, N. (2008). Annotationsmanual zur Gestenmorphologie.

Bergmann, K., Fröhlich, C., Hahn, F., Kopp, S., Lücking, A., and Rieser, H. (2007). Wegbeschreibungsexperiment: Grobannotationsschema.

Bergmann, K. and Kopp, S. (2006). Verbal or visual: How information is distributed across speech and gesture in spatial dialog. In Schlangen, D. and Fernandez, R., editors, *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue*, pages 90–97.

Bergmann, K. and Kopp, S. (2008). Multimodal content representation for speech and gesture production. In Theune, M., van der Sluis, I., Bachvarova, Y., and André, E., editors, *Proceedings of the 2nd Workshop on Multimodal Output Generation*, pages 61–68.

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

Bergmann, K. and Kopp, S. (2009b). Increasing expressiveness for virtual agents–Autonomous generation of speech and gesture in spatial description tasks. In *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems*, pages 361–368, Budapest, Hungary.

Bergmann, K. and Kopp, S. (2010a). Modelling the production of co-verbal iconic gestures by learning Bayesian Decision Networks. *Applied Artificial Intelligence*, 24:530–551.

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

Bergmann, K., Kopp, S., and Eyssel, F. (2010). Individualized gesturing outperforms average gesturing–evaluating gesture production in virtual humans. In Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., and Safonova, A., editors, *Proceedings of the 10th Conference on Intelligent Virtual Agents*, pages 104–117, Berlin/Heidelberg. Springer.

Bickmore, T. and Cassell, J. (2005). Social dialogue with embodied conversational agents. In van Kuppevelt, J., Dybkjaer, L., and Bernsen, N., editors, *Advances in Natural, Multimodal Dialogue Systems*, New York. Kluwer Academic Publishers.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94:115–147.

Bortz, J. (2005). *Statistik f"ur Human- und Sozialwissenschaftler*. Springer: Berlin, 6. edition.

Bressem, J. (2008). Notating gestures—proposal for a form based notation system of coverbal gestures.

Buisine, S. and Martin, J.-C. (2007). The effects of speech-gesture cooperation in animated agents' behavior in multimedia presentations. *Interacting with Computers*, 19:484–493.

Buschmeier, H. (2008). Alignment-Supported Microplanning in Natural Language Generation. Master's thesis, Faculty of Technology, Bielefeld University.

Butterworth, B. and Beattie, G. (1978). Gesture and silence as indicators of planning in speech. In Campbell, R. and Smith, P., editors, *Recent advances in the psychology of language: Formal and experimental approaches*, pages 347–360. Plenum Press, New York, NY.

Calbris, G. (1990). *The Semiotics of French Gesture*. Indiana University Press, Bloomington.

Carletta, J. (1996). Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics*, 22:249–254.

Cassell, J. (2000). More than just another pretty face: Embodied conversational interface agents. *Communications of the ACM*, 43:70–78.

Cassell, J., McNeill, D., and McCullough, K.-E. (1999). Speech-gesture mismatches: Evidence for one underlying representation of linguistic and non-linguistic information. *Pragmatics and Cognition*, 7:1–33.

Cassell, J., Stone, M., and Yan, H. (2000a). Coordination and context-dependence in the generation of embodied conversation. In *Proceedings of the First International Conference on Natural Language Generation*.

Cassell, J., Sullivan, J., Prevost, S., and Churchill, E., editors (2000b). *Embodied Conversational Agents*. MIT Press, Cambridge.

Cassell, J. and Tartaro, A. (2007). Intersubjectivity in human–agent interaction. *Interaction Studies*, 8:391–410.

Cassell, J. and Thórisson, K. (1999). The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13:519–538.

Cassell, J., Vilhjálmsson, H., and Bickmore, T. (2001). BEAT: The behavior expression animation toolkit. In *Proceedings of SIGGRAPH '01*, pages 477–486, New York, NY.

Chu, M. and Kita, S. (2009). Co-speech gestures do not originate from speech production processes: Evidence from the relationship between co-thought and co-speech gestures. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pages 591–595.

Clark, H. (1996). *Using Language*. Cambridge University Press, Cambridge, UK.

Cohen, A. (1977). The communicative functions of hand illustrators. *Journal of Communication*, 27:54–63.

Cohen, A. and Harrison, R. (1973). Intentionality in the use of hand illustrators in face-to-face communication situations. *Journal of Personality and Social Psychology*, 28:276–279.

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

Collins, A. and Loftus, E. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82:407–428.

Cooper, G. F. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning Journal*, 9:308–347.

Croft, D. and Thagard, P. (2002). Dynamic imagery: A computational model of motion and visual analogy. In Magnani, L. and Nersessian, N., editors, *Model-based reasoning: Science, technology, values*, pages 259–274. Kluwer/Plenum, New York.

Dale, R. and Viethen, J. (2009). Referring expression generation through attribute-based heuristics. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 58–65, Athens, Greece.

de Ruiter, J. (1998). *Gesture and Speech Production*. PhD thesis, University of Nijmegen.

de Ruiter, J. (2000). The production of gesture and speech. In McNeill, D., editor, *Language and gesture*, pages 284–311. Cambridge University Press, Cambridge, UK.

de Ruiter, J. (2007). Some multimodal signals in humans. In *Proceedings of the 1st Workshop on Multimodal Output Generation*, pages 141–148. CTIT.

de Saussure, F. (1916). *Course de linguistique générale*. Payot, Paris.

DeCarolis, B., Pelachaud, C., Poggi, I., and Steedman, M. (2004). APML, a mark-up language for believable behavior generation. In Prendinger, H. and Ishizuka, M., editors, *Life-like Characters. Tools, Affective Functions and Applications*. Springer, Berlin/Heidelberg.

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

Druzdzel, M., van der Gaag, L., Henrion, M., and Jensen, F. (2000). Building probabilistic networks: Where do the numbers come from? *IEEE Transactions on Knowledge and Data Engineering*, 12:481–485.

Duda, R., Hart, P., and Stork, D. (2001). *Pattern Classification*. John Wiley & Sons, New York.

Efron, D. (1941/1970). *Gesture, Race and Culture*. Mouton, The Hague.

Ekman, P. and Friesen, W. (1969). The repertoire of nonverbal behavior: Categories, origins, usage and coding. *Semiotica*, 1:49–98.

Ekman, P. and Friesen, W. (1972). Hand movements. *Journal of Communication*, 22:353–374.

Fast, A. (2009). *Learning the Structure of Bayesian Networks with Constraint Satisfaction*. PhD thesis, University of Massachusetts Amherst.

Feyereisen, P., van de Wiele, M., and Dubois, F. (1988). The meaning of gestures: What can be understood without speech? *Cahiers de Psychologie Cognitive*, 8:3–25.

Firbas, J. (1971). On the concept of communicative dynamism in the theory of functional sentence perspective. *Philologica Pragensia*, 8:135–144.

Fiske, S. T., Cuddy, A. J., and Glick, P. (2006). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Science*, 11:77–83.

Foster, M. and Oberlander, J. (2007). Corpus-based generation of head and eyebrow motion for an embodied conversational agent. *Language Resources and Evaluation*, 41:305–323.

Foster, M. and White, M. (2007). Avoiding repetition in generated text. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*.

Gao, Y. (2002). Automatic extraction of spatial location for gesture generation. Master's thesis, MIT, Cambridge, MA.

Gerwing, J. and Bavelas, J. (2004). Linguistic influences on gesture's form. *Gesture*, 4:157–195.

Gibbon, D., Gut, U., Hell, B., Looks, K., Milde, J.-T., Thies, A., and Trippel, T. (2004). CoGesT: A formal transcription system for conversational gesture. In *Proceedings of LREC 2004*.

Glasgow, J. (1993). The imagery debated revisited: A computational perspective. *Computational Intelligence*, 9:310–333.

Goffman, E. (1981). Replies and responses. In Goffman, E., editor, *Forms of Talk*, pages 5–77. University of Pennsylvania Press, Philadelphia, PA.

Goldin-Meadow, S. (2003). *Hearing Gesture—How Our Hands Help Us Think*. Harvard University Press, Cambridge, MA.

Goldin-Meadow, S. and Butcher, C. (2003). Pointing toward two-word speech in young children. In Kita, S., editor, *Pointing*, pages 85–107. Lawrence Erlbaum Associates, Mahwah, NJ.

Goodman, N. (1976). *Languages of Art: An Approach to a Theory of Symbols*. Hecket Publishing Company.

Gullberg, M. (1998). *Gesture as a communication strategy in second language discourse: A study of learners of French and Swedish*. Lund University Press, Lund.

Gullberg, M. (2010). Language-specific encoding of placement events in gestures. In Bohnemeyer, J. and Pederson, E., editors, *Event Representation in Language and Cognition*. Cambridge University Press, Cambridge, UK.

Gullberg, M. and Holmqvist, K. (2006). What speakers do and what listeners look at. Visual attention to gestures in human interaction live and on video. *Pragmatics and Cognition*, 14:53–82.

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

Habets, B., Kita, S., Shao, Z., Özyürek, A., and Hagoort, P. (2010). The role of synchrony and ambiguity in speech-gesture integration during comprehension. *Journal of Cognitive Neuroscience*.

Hadar, U., Burstein, A., Krauss, R., and Soroker, N. (1998). Ideational gestures and speech in brain-damaged subjects. *Language and Cognitive Processes*, 13:59–76.

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

Hartmann, B., Mancini, M., and Pelachaud, C. (2002). Formational parameters and adaptive prototype instantiation for MPEG-4 compliant gesture synthesis. In *Proceedings of Computer Animation 2002*. IEEE Computer Society Press.

Hartmann, B., Mancini, M., and Pelachaud, C. (2006). Implementing expressive gesture synthesis for embodied conversational agents. In Gibet, S., Courty, N., and Kamp, J.-F., editors, *Gesture in Human-Computer Interaction and Simulation*, pages 45–55. Springer, Berlin/Heidelberg.

Haviland, S. and Clark, H. (1974). What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior*, 13:512–521.

Hayes-Roth, B. (1985). A blackboard architecture for control. *Artificial Intelligence*, 26:251–321.

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

Heylen, D., van Es, I., Nijholt, A., and van Dijk, B. (2002). Experimenting with the gaze of a conversational agent. In *Proceedings International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems*, pages 93–100.

Hiyakumoto, J., Prevost, S., and Cassell, J. (1997). Semantic and discourse information for text-to-speech intonation. In *Proceedings of the ACL Workshop on Concept-to-Speech Generation*, pages 47–56.

Hoffmann, A., Krämer, N., Lam-Chi, A., and Kopp, S. (2009). Media equation revisited. Do users show polite reactions towards an embodied agent? In Ruttkay, Z., Kipp, M., Nijholt, A., and Vilhjálmsson, H., editors, *Proceedings of the 9th Intern. Conf. on Intelligent Virtual Agents*, pages 159–165, Berlin. Springer.

Hofs, D., Theune, M., and op den Akker, R. (2010). Natural interaction with a virtual guide in a virtual environment—a multimodal dialogue system. *Journal of Multimodal User Interfaces*, 3:141–153.

Holler, J. and Beattie, G. (2002). A micro-analytic investigation of how iconic gesture and speech represent core semantic features in talk. *Semiotica*, 142:31–69.

Holler, J. and Beattie, G. (2003). How iconic gestures and speech interact in the representation of meaning: Are both aspects really integral to the process? *Semiotica*, 146/1:81–116.

Holler, J. and Beattie, G. (2004). The interaction of iconic gesture and speech. In Camurri, A. and Volpe, G., editors, *Proceedings of the 5th International Gesture Workshop*, pages 63–69, Berlin/Heidelberg. Springer.

Holler, J., Shovelton, H., and Beattie, G. (2009). Do iconic hand gestures really contribute to the communication of semantic information in a face-to-face context? *Journal of Nonverbal Behavior*, 33:73–88.

Holler, J. and Stevens, R. (2007). An experimental investigation into the effect of common ground on how speakers use gesture and speech to represent size information in referential communication. *Journal of Language and Social Psychology*, 26:4–27.

Holler, J. and Wilkin, K. (2009). Communicating common ground: How mutually shared knowledge influences speech and gesture in a narrative task. *Language and Cognitive Processes*, 24:267–289.

Hostetter, A. and Alibali, M. (2007). Raise your hand if you're spatial—relations between verbal and spatial skills and gesture production. *Gesture*, 7:73–95.

Hostetter, A. and Alibali, M. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin and Review*, 15/3:495–514.

Hostetter, A. and Hopkins, W. (2002). The effect of thought structure on the production of lexical movements. *Brain and Language*, 82:22–29.

Howard, R. and Matheson, J. (1981/2005). Influence diagrams. *Decision Analysis*, 2:127–143.

Iverson, J. and Goldin-Meadow, S. (1998). Why people gesture when they speak. *Nature*, 396:228.

Jacobs, N. and Garnham, A. (2007). The role of conversational hand gestures in a narrative task. *Journal of Memory and Language*, 56:291–303.

Jameson, A. (2003). Adaptive interfaces and agents. In Jacko, J. and Sears, A., editors, *Handbook of Human-Computer Interaction in Interactive Systems*, pages 305–330. Erlbaum, Mahwah, NJ, 1st edition.

Jensen, F. and Nielsen, T. (2007). *Bayesian Networks and Decision Graphs*. Springer, New York, NY.

Kelly, S., Kravitz, C., and Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language*, 89:253–260.

Kendon, A. (1972). Some relationships between body motion and speech: An analysis of an example. In Siegman, A. and Pope, B., editors, *Studies in Dyadic Communication*, pages 177–210. New York: Pergamon.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In Key, M., editor, *The Relationship of Verbal and Nonverbal Communication*, pages 207–227. The Hague.

Kendon, A. (1987). On gesture: Its complementary relationship with speech. In Siegman, A. and Feldstein, S., editors, *Nonverbal Behavior and Communication*, pages 65–97. Lawrence Erlbaum, Hillsdale, NJ.

Kendon, A. (1988). How gestures can become like words. In Poyatos, F., editor, *Cross-Cultural Perspectives in Nonverbal Communication*, pages 131–141. Hogrefe, Toronto.

Kendon, A. (2004). *Gesture—Visible Action as Utterance*. Cambridge University Press.

Kieras, D. (1978). Beyond pictures and words: Alternative information-processing models for imagery effects in verbal memory. *Psychological Bulletin*, 85:532–554.

Kimbara, I. (2006). On gestural mimicry. *Gesture*, 6:39–61.

Kimbara, I. (2008). Gesture form convergence in joint description. *Journal of Nonverbal Behavior*, 32:123–131.

Kimura, D. (1973a). Manual activity during speaking—1. right-handers. *Neuropsychologia*, 11:45–50.

Kimura, D. (1973b). Manual activity during speaking—2. left-handers. *Neuropsychologia*, 11:51–55.

Kipp, M. (2004). *Gesture Generation by Imitation–From Human Behavior to Computer Character Animation*. Dissertation.com, Boca Raton, Florida.

Kita, S. (1993). *Language and Thought Interface: A Study of Spontaneous Gestures and Japanese Mimetics*. PhD thesis, University of Chicago.

Kita, S. (2000). How representational gestures helps speaking. In McNeill, D., editor, *Language and gesture*, pages 162–185. Cambridge University Press, Cambridge, UK.

Kita, S. (2009). Cross-cultural variation of speech-accompanying gesture: A review. *Language and Cognitive Processes*, 24(2):145–167.

Kita, S. and Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, 48:16–32.

Kita, S., Özyürek, A., Allen, S., Brown, A., Furman, R., and Ishizuka, T. (2007). Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production. *Language and Cognitive Processes*, 22:1212–1236.

Kita, S., van Gijn, I., and van der Hulst, H. (1998). Movement phases in signs and co-speech gestures, and their transcription by human coders. In Wachsmuth, I. and Fröhlich, M., editors, *Gesture and Sign Language in Human-Computer Interaction*, pages 23–25. Springer, Berlin/Heidelberg.

Kontkanen, P., Myllymäki, P., Silander, T., and Tirri, H. (1997). Comparing predictive inference methods for discrete domains. In *Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics*, pages 311–318, Ft. Lauderdale, FL.

Kopp, S. (2003). *Synthese und Koordination von Sprache und Gestik für Virtuelle Multimodale Agenten*. Akademische Verlagsgesellschaft Aka, Berlin.

Kopp, S. (2010). Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors. *Speech Communication*, 52:587–597.

Kopp, S., Bergmann, K., and Wachsmuth, I. (2008). Multimodal communication from multimodal thinking—towards an integrated model of speech and gesture production. *Semantic Computing*, 2(1):115–136.

Kopp, S., Krenn, B., Marsella, S., Marshall, A., Pelachaud, C., Pirker, H., Thorisson, K., and Vilhjalmsson, H. (2006). Towards a common framework for multimodal generation: The behavior markup language. In *Proceedings the 6th Conference on Intelligent Virtual Agents*, volume 4133, pages 205–217, Berlin/Heidelberg. Springer.

Kopp, S., Tepper, P., and Cassell, J. (2004). Towards integrated microplanning of language and iconic gesture for multimodal output. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 97–104, New York, NY.

Kopp, S., Tepper, P., Ferriman, K., Striegnitz, K., and Cassell, J. (2007). Trading spaces: How humans and humanoids use speech and gesture to give directions. In Nishida, T., editor, *Conversational Informatics*, pages 133–160. John Wiley, New York.

Kopp, S. and Wachsmuth, I. (2004). Synthesizing multimodal utterances for conversational agents. *Computer Animation and Virtual Worlds*, 15:39–52.

Kosslyn, S. (1987). Seeing and imagining in the celebral hemispheres: A computational approach. *Psychological Review*, 94:148–175.

Krämer, N., Tietz, B., and Bente, G. (2003). Effects of embodied interface agents and their gestural activity. In Rist, T., Aylett, R., Ballin, D., and Rickel, J., editors, *Proceedings of the 4th International Workshop on Intelligent Virtual Agents*, pages 292–300, Berlin/Heidelberg. Springer.

Kranstedt, A., Kopp, S., and Wachsmuth, I. (2002). MURML: A multimodal utterance representation markup language for conversational agents. Technical Report 2002/05, SFB 360, Bielefeld University.

Krauss, J., Morrel-Samuels, P., and Colasante, C. (1991). Do conversational hand gestures communicate? *Journal of Personality and Social Psychology*, 61:743–754.

Krauss, R., Chen, Y., and Chawla, P. (1996). Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? *Advances in Experimental Social Psychology*, 28:389–450.

Krauss, R., Chen, Y., and Gottesman, R. (2000). Lexical gestures and lexical access: A process model. In McNeill, D., editor, *Language and gesture*, pages 261–283. Cambridge University Press, Cambridge, UK.

Krauss, R., Dushay, R., Chen, Y., and Rauscher, F. (1995). The communicative value of conversational hand gestures. *Journal of Experimental Social Psychology*, 31:533–552.

Krippendorff, K. (1980). *Content Analysis*. Sage Publications, Beverly Hills, CA.

Kruijff-Korbayova, I. and Steedman, M. (2003). Discourse and information structure. *Journal of Logic, Language and Information*, 12:249–259.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Lang, E. (1989). The semantics of dimensional designation of spatial objects. In Bierwisch, M. and Lang, E., editors, *Dimensional adjectives: Grammatical structure and conceptual interpretation*, pages 263–417. Springer, Berlin.

Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis*, 19:191–201.

Lausberg, H. and Kita, S. (2003). The content of the message influences the hand choice in co-speech gestures and in gesturing without speaking. *Brain and Language*, 86:57–69.

Lausberg, H. and Slöetjes, H. (2009). Coding gestural behaviour with the NEUROGES–ELAN system. *Behavior Research Methods*, 41(3):841—849.

Lee, J. and Marsella, S. (2006). Nonverbal behavior generator for embodied conversational agents. In Gratch, J., Young, M., Aylett, R., Ballin, D., and Olivier, P., editors, *Proceedings of the 6th International Conference on Intelligent Virtual Agents*, pages 243–255, Berlin/Heidelberg. Springer.

Levelt, W. (1989). *Speaking: From Intention to Articulation.* MIT Press.

Levinson, S. (1996). Frames of reference and molyneux's question: Cross-linguistic evidence. In Bloom, P., Peterson, M., Nadel, L., and Garrett, M., editors, *Space and Language*, pages 109–169. MIT Press.

Levy, E. and McNeill, D. (1992). Speech, gesture, and discourse. *Discourse Processes*, 15:277–301.

Lücking, A., Bergmann, K., Hahn, F., Kopp, S., and Rieser, H. (2010). The Bielefeld speech and gesture alignment corpus (SaGA). In Kipp, M., Martin, J.-P., Paggio, P., and Heylen, D., editors, *LREC 2010 Workshop: Multimodal Corpora—Advances in Capturing, Coding and Analyzing Multimodality.*

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, 63:215–232.

Madsen, A., Jensen, F., Kjærulff, U., and Lang, M. (2005). HUGIN–The tool for Bayesian networks and influence diagrams. *International Journal of Artificial Intelligence Tools*, 14:507–543.

Mancini, M. and Pelachaud, C. (2010). Generating distinctive behavior for embodied conversational agents. *Journal of Multimodal User Interfaces*, 3:249–261.

Marr, D. and Nishihara, H. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. In *Proceedings of the Royal Society of London*, volume 200, pages 269–294.

Marsh, T. and Watt, A. (1998). Shape your imagination: Iconic gestural-based interaction. In *Proceedings of the IEEE Virtual Reality Annual International Symposium*.

Martin, J.-C., Abrilian, S., and Devillers, L. (2007). Individual differences in the perception of spontaneous gesture expressivity. In *Integrating Gestures*, page 71.

McClave, E. (1994). Gestural beats: The rhythm hypothesis. *Journal of Psycholinguistic Research*, 23:45–66.

McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review*, 92:271–295.

McNeill, D. (1992). *Hand and Mind—What Gestures Reveal about Thought*. University of Chicago Press, Chicago.

McNeill, D. (2000). Introduction. In *Language and Gesture*, pages 1–10. Cambridge University Press, Cambridge, UK.

McNeill, D. (2005). *Gesture and Thought*. University of Chicago Press, Chicago, IL.

McNeill, D. and Duncan, S. (2000). Growth points in thinking-for-speaking. In McNeill, D., editor, *Language and gesture*, pages 141–161. Cambridge University Press, Cambridge, UK.

McNeill, D. and Levy, E. (1982). Conceptual representations in language activity and gesture. In Jarvella, R. and Klein, W., editors, *Speech, place and action*, pages 271–295. John Wiley & Sons, Chichester.

Menke, P. and Mehler, A. (2010). The Ariadne system: A flexible and extensible framework for the modeling and storage of experimental data in the humanities. In Kipp, M., Martin, J.-P., Paggio, P., and Heylen, D., editors, *LREC 2010 Workshop: Multimodal Corpora—Advances in Capturing, Coding and Analyzing Multimodality*.

Müller, C. (1998). *Redebegleitende Gesten: Kulturgeschichte–Theorie–Sprachvergleich*. Berlin Verlag, Berlin.

Murdock, B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, 64:482–488.

Murphy, K. (2001). *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California.

Nass, C., Isbister, K., and Lee, E.-J. (2000). Truth is beauty: Researching embodied conversational agents. In Cassell, J., Sullivan, J., Prevost, S., and Churchill, E., editors, *Embodied Conversational Agents*, pages 374–402, Cambridge. MIT Press.

Neff, M., Kipp, M., Albrecht, I., and Seidel, H.-P. (2008). Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics*, 27(1):1–24.

Özyürek, A. (2002). Speech-gesture relationship across languages and in second language learners: Implications for spatial thinking and speaking. In *Proceedings of the 26th annual Boston University Conference on Language Development*, pages 500–509.

Özyürek, A., Kita, S., Allen, S., Furman, R., and Brown, A. (2005). How does linguistic framing influence co-speech gestures? Insights from crosslinguistic differences and similarities. *Gesture*, 5:216–241.

Paivio, A. (1986). *Mental Representations*. Oxford University Press.

Parrill, F. (2010). The hands are part of the package: Gesture, common ground and information packaging. In Rice, S. and Newman, J., editors, *Empirical and Experimental Methods in Cognitive/Functional Research*, pages 285–302. CSLI Publications, Stanford, CA.

Pearl, J. (1985). Bayesian networks: A model of self-activated memory for evidential reasoning". In *Proceedings of the 7th Conference of the Cognitive Science Society*, pages 329—334.

Peirce, C. (1965). *Collected Papers of Charles Sanders Peirce*. The Belknap Press of Harvard University Press, Cambridge, MA.

Pelachaud, C. and Poggi, I. (2002). Subtleties of facial expressions in embodied agents. *Journal of Visualization and Computer Animation*, 13:301–312.

Poggi, I., Pelachaud, C., de Rosis, F., Carofiglio, V., and De Carolis, B. (2005). GRETA. A believable embodied conversational agent. In Stock, O. and Zancarano, M., editors, *Multimodal Intelligent Information Presentation*. Kluwer.

Prillwitz, S., Leven, R., Zienert, H., Hanke, T., and Henning, J. (1989). *HamNoSys Version 2.0: Hamburg Notation System for Sign Languages: An Introduction*. Signum Press, Hamburg.

Prince, E. F. (1981). Towards a taxonomy of given-new information. In Cole, P., editor, *Radical Pragmatics*, pages 223–255. Academic Press, New York, NY.

Quek, F. and McNeill, D. (2000). Gesture and speech multimodal conversational interaction in monocular video. In *Proceedings of the 3rd International Conference on Methods and Techniques in Behavioral Research, Measuring Behavior*, page 215.

Quinlan, J. (1986). Induction of decision trees. *Machine Learning*, 1(1):81–106.

Rauscher, F., Krauss, R., and Chen, Y. (1996). Gesture, speech, and lexical access: The role of lexical movements in speech production. *Psychological Science*, 7:226–231.

Rehm, M. and André, E. (2007). Informing the design of agents by corpus analysis. In Nishida, T. and Nakano, Y., editors, *Conversational Informatics*. John Wiley & Sons, Chichester, UK.

Rehm, M., Nakano, Y., Nishida, E. A. T., Bee, N., Endrass, B., Wissner, M., Lipi, A., and Huang, H.-H. (2008). From observation to simulation: Generating culture-specific behavior for interactive systems. *AI & Society*, 24:267–280.

Reiter, E. and Dale, R. (2000). *Building Natural Language Generation Systems*. Cambridge University Press, Cambridge, UK.

Rimé, B. and Schiaratura, L. (1991). Gesture and speech. In Feldman, R. and Rimé, R., editors, *Fundamentals of nonverbal behavior*, pages 239–281. Press Syndicate of the University of Cambridge, New York, NY.

Ritz, J., Dipper, S., and Götze, M. (2008). Annotation of information structure: An evaluation across different types of texts. In *Proceedings of the 6th LREC conference*, pages 2137–2142.

Rogers, W. (1978). The contribution of kinesic illustrators toward the comprehension of verbal behavior within utterances. *Human Communication Research*, 5:54–62.

Ruttkay, Z. (2007). Presenting in style by virtual humans. In Esposito, A., editor, *Verbal and Nonverbal Communication Behaviours*, pages 23–36. Springer, Berlin/Heidelberg.

Schmidt, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Sowa, T. (2006). *Understanding Coverbal Iconic Gestures in Shape Descriptions*. Akademische Verlagsgesellschaft Aka, Berlin.

Sowa, T. and Wachsmuth, I. (2009). A computational model for the representation and processing of shape in coverbal iconic gestures. In Coventry, K., Tenbrink, T., and Bateman, J., editors, *Spatial Language and Dialogue*, pages 132–146. Oxford University Press, Oxford, UK.

Spirtes, P. and Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social Science Computing Review*, 9:62–72.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT Press, 2nd edition.

Stalnaker, R. (1978). Assertion. In Cole, P., editor, *Syntax and semantics: Pragmatics*, volume 9, pages 315–332. Academic Press, New York, NY.

Steck, H. and Tresp, V. (1999). Bayesian belief networks for data mining. In *Proceedings of the 2nd Workshop on Data Mining and Data Warehousing*.

Stegmann, J. and Lücking, A. (2005). Assessing reliability on annotations (1): Theoretical considerations. Technical Report 2, SFB 360, Bielefeld University.

Stokoe, W. C. (1972). *Semiotics and Human Sign Languages*. Mouton, The Hague.

Stone, M. (2002). Lexicalized grammar 101. In *Proceedings on the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Provcessing and Computational Linguistics*, pages 77–84, Philadelphia, PA.

Stone, M., DeCarlo, D., Oh, I., Rodriguez, C., Stere, A., Lees, A., and Bregler, C. (2004). Speaking with hands: Creating animated conversational characters from recordings of human performance. In *Proceedings of SIGGRAPH '04*, pages 506–513.

Stone, M., Doran, C., Webber, B., Bleam, T., and Palmer, M. (2003). Microplanning with Communicative Intentions: The SPUD System. *Computational Intelligence*, 19:311–381.

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

Striegnitz, K., Tepper, P., Lovett, A., and Cassell, J. (2007). Knowledge representation for generating locating gestures in route directions. In Coventry, K., Tenbrink, T., and Bateman, J., editors, *Spatial Language and Dialogue (Explorations in Language and Space)*. Oxford University Press, Oxford.

Teyssier, M. and Koller, D. (2005). Ordering-based search: A simple and effective algorithm for learning bayesian networks. In *Proceedings of the 21st Conference on Uncertainty in AI*, pages 584–590.

Thomas, N. (2010). Mental imagery. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy (Fall 2010 Edition)*.

Thompson, L. and Massaro, D. (1986). Evaluation and integration of speech and pointing gestures during referential understanding. *Journal of Experimental Child Psychology*, 42:144–168.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65:31–78.

Tuite, K. (1993). The production of gesture. *Semiotica*, 93:83–105.

Viethen, J. and Dale, R. (2010). Speaker-dependent variation in content selection for referring expression generation. In *Proceedings of the 8th Australasian Language Technology Workshop*, pages 81–89.

Vilhjalmsson, H., Cantelmo, N., Cassell, J., Chafai, N., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A., Pelachaud, C., Ruttkay, Z., Thorisson, K., van Welbergen, H., and van der Werf, R. (2007). The Behavior Markup Language: Recent developments and challenges. In Pelachaud, C., Martin, J.-C., Andre, E., Chollet, G., Karpouzis, K., and Pelé, D., editors, *Proceedings of the 7th International Conference on Intelligent Virtual Agents*, pages 99–111, Berlin/Heidelberg. Springer.

Webb, R. (1996). *Linguistic features of metaphoric gestures*. PhD thesis, University of Rochester, Rochester, NY.

Witte, B. and Weisemann, P. (2008). Entwicklung einer multimodalen Wissensrepräsentation zur Sprach- und Gestengenerierung. Master's thesis, Faculty of Technology, Bielefeld University.

Wittig, F. (2002). *Maschinelles Lernen Bayes'scher Netze für benutzeradaptive Systeme*. PhD thesis, Saarland University.

Wundt, W. (1900/1973). *The Language of Gestures*. Mouton, The Hague.

Yu, Q. and Terzopoulos, D. (2007). A decision network framework for the behavioral animation of virtual humans. In *Proceedings of SIGGRAPH '07*, pages 119–128.