From Annotated Multimodal Corpora to Simulated Human-Like Behaviors

Rehm, Matthias; André, Elisabeth

Published in:
Modeling Communication with Robots and Virtual Humans

Publication date:
2008

Document Version
Publisher's PDF, also known as Version of record

From Annotated Multimodal Corpora to Simulated Human-Like Behaviors

Matthias Rehm and Elisabeth André

Augsburg University, Institute of Computer Science
86159 Augsburg, Germany
{rehm,andre}@informatik.uni-augsburg.de
http://mm-werkstatt.informatik.uni-augsburg.de

Abstract. Multimodal corpora prove useful at different stages of the development process of embodied conversational agents. Insights into human-human communicative behaviors can be drawn from such corpora. Rules for planning and generating such behavior in agents can be derived from this information. And even the evaluation of human-agent interactions can rely on corpus data from human-human communication. In this paper, we exemplify how corpora can be exploited at the different development steps, starting with the question of how corpora are annotated and on what level of granularity. The corpus data can be used either directly for imitating the human behavior recorded in the corpus or rules can be derived from the data which govern the behavior planning process. Corpora can even play a vital role in the evaluation of agent systems. Several studies are presented that make use of corpora for the evaluation task.

Keywords: Multimodal interaction, embodied conversational agent, behavior modelling, multimodal corpora.

1 Introduction

A number of approaches to modeling the behaviors of embodied conversational agents (ECA's) are based on a direct simulation of human behaviors. Consequently, it comes as no surprise that the use of data-driven approaches which allow us to validate design choices empirically has become increasingly popular in the ECA field. To get insight into human-human conversation, researchers rely on a large variety of resources including recordings of users in "natural" or staged situations, TV interviews, Wizard of Oz studies, and motion capturing data. Various annotation schemes have been designed to extract relevant information for multimodal behaviors, such as facial expressions, gestures, postures and gaze. In addition, there has been increasing interest in the design of annotation schemes to capture emotional behaviors in human-human conversation. Progress in the field has been boosted by the availability of new tools that facilitate the acquisition and annotation of corpora.

The use of data-driven approaches provides a promising approach to the modeling of ECA behaviors since it allows us to validate design choices empirically.

I. Wachsmuth and G. Knoblich (Eds.): Modeling Communication, LNAI 4930, pp. 1–17, 2008.

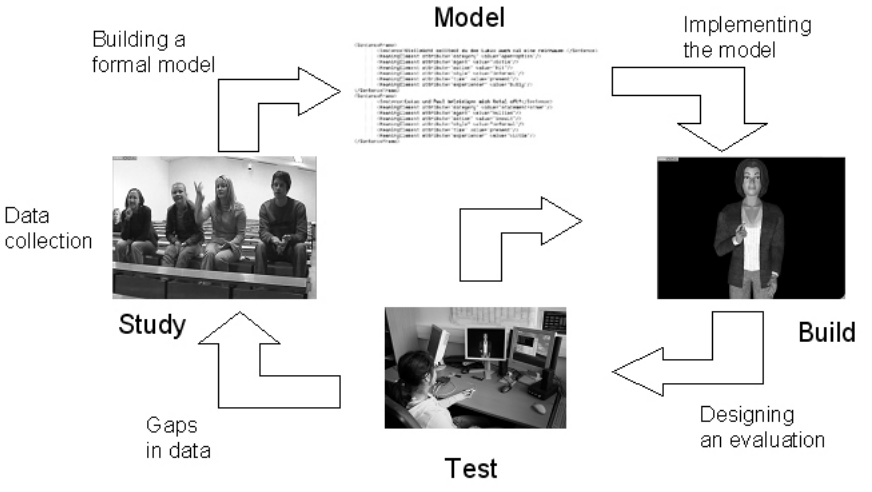

Fig. 1. Development cycle for embodied conversational agents

Nevertheless, the creation of implementable models still leaves many research issues open. One difficulty lies in the fact that an enormous amount of data is needed to derive regularities from concrete instantiations of human-human behavior. In rare cases, we are interested in the replication of behaviors shown by individuals. Rather, we aim at the extraction of behavior profiles that are characteristic of a group of people, for example, introverts versus extroverts. Furthermore, the resulting ECA behaviors only emulate a limited amount of phenomena of human-human behaviors. In particular, the dynamics of multimodal behaviors has been largely neglected so far. Last but not least, there is the danger that humans expect a different behavior from an ECA than from a human conversational partner which might limit the potential benefits of a simulation-based approach.

The methodological approach for modeling communicative behavior for embodied conversational agents is well exemplified by Cassel's Study-Model-Build-Test development cycle [8]. Figure 1 gives an overview of the different steps in this development cycle. To build a formal model for generating realistic agent behaviors, data of humans that are engaged in a dialogue with other humans are collected. In most cases, formal models are not built from scratch. Rather, the data analysis serves to refine existing models found in the literature. The resulting models of human-human conversational behavior then serve as a basis for the implementation of ECAs that replicate the behaviors addressed by the models. To evaluate the resulting system, experiments are set up in which humans are confronted with ECAs following the model. Depending on the outcome of such experiments, the developers might decide to acquire new data from humans so that the existing models may be refined.

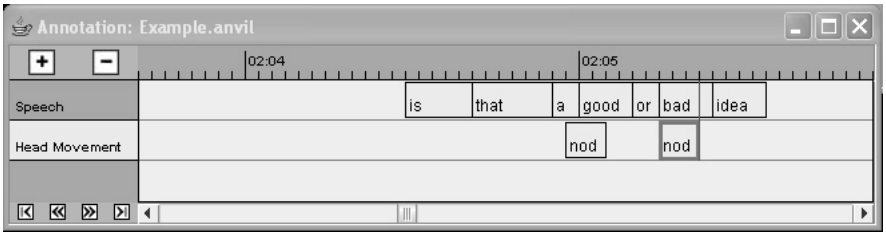

Fig. 2. Information of different modalities is annotated in parallel on a temporal score

In the rest of this article, we exemplify how the single steps of the development process have been realized by ECA researchers. First we will provide an overview of existing corpus-based work that has been conducted in order to get insight on multimodal human-human dialogue with the aim to replicate such behaviors in an embodied conversational agent. We will present several approaches to bridge the gap from corpus analysis to behavior generation including copy-synthesis, generate-and-filter as well as first attempts to realize trainable generation approaches. Finally, we will discuss several empirical studies that have been conducted with the aim to validate models derived from a corpus.

2 Multimodal Corpora for Studying Human Behavior

It is undeniable that a rich literature on human communicative behavior exists covering such diverse areas as dialogue management (e.g. [2]) gesture use (e.g., [26]; [20]) or gaze behavior (e.g., [4]; [19]). But for the explicit task of emulating human communicative behavior by an embodied conversational agent, this literature is often deficient in one way or another. This is due to the fact that the proposed theories and models were of course not created with the generation task in mind and thus often consider only one modality or lack crucial information, e.g. about the synchronization of different modalities, making it sometimes necessary to collect a completely new corpus for deriving this information.

Corpus work has a long tradition in the social sciences where it is employed as a descriptive tool to gain insights into human communicative behavior. Due to the increased multimedia abilities of computers, a number of tools for video analysis have been developed over the last decade allowing for standardized annotation of multimodal data.

2.1 Annotating Multimodal Information

A corpus is a collection of video recordings of human (or human-agent) communicative behavior that is annotated or coded with different types of information. A multimodal corpus analyses more than one modality in a single annotation, e.g. speech and gesture, and ideally explicates the links and crossmodal relations between the different modalities. Which kind of information is coded in a given

corpus is defined in an annotation scheme that specifies the coding attributes and values. Figure 2 gives an impression of a very simple annotation board that codes utterances and head movements. Each modality has its own layer in a data format that resembles a musical score i.e. the information on different modalities is coded in parallel on a temporal axis. The more widely used tools for annotating multimodal corpora like Anvil[1] or the Nite Workbench[2] support tailoring the annotations scheme to a given task often using an XML format that is proprietary to the given tool. The annotated data is also generally given in XML format allowing for accessing the information by simply parsing the XML tree.

There is no limit on the number of annotation layers one can define to describe communicative behavior. Speech, gestures, gaze or facial expressions are good candidates for which information can be annotated on different levels, such as the spatio-temporal course of actions, semantic content, or communicative function. Depending on the research area, a great number of annotation schemes are available focusing on these different levels of annotation. Basically, we can distinguish between *how* the information in this channel is realized at the signal level and *what* kind of information is realized. How the information is realized often includes the temporal course of an action in a given channel, for gestures e.g. the distinction between preparation, stroke, and retraction phases (see below). What kind of information is realized often includes categorizing the elements of a specific channel and emphasizing the communicative function of the modality by coding e.g. what kind of feedback is given by a specific gaze behavior. Whereas the how level of annotation supports the extraction of concrete animation parameters for an agent and information about the temporal synchronization of modalities, the what level allows us to derive information concerning crossmodal functional relations that are of relevance to the behavior planning process.

Elaborate annotation schemes exist (for an overview see e.g. [23]) ranging from gesture analysis (e.g. [15]; [26]), over general movement analysis (see e.g. [24]) to the coding of facial expressions (e.g. [12]). If they are suitable for the task of modeling the communicative behavior of an agent — be it at the concrete level of controlling the animation or at the more abstract level of planning appropriate multimodal behavior — has to be decided from case to case based on a given agent system.

The MUMIN corpus focuses on the analysis of gestures and facial displays which accompany communicative phenomena, such as feedback, turn-taking or sequencing ([3]), and is a good example of a multimodal corpus that was not created with the generation of multimodal output in mind. While not focusing on the generation task, the corpus shows how different modalities converge in their contribution to communicative functions. Feedback, turn management, and sequencing constitute communicative functions that are not bound to a single modality, such as speech, but are inherently multimodal in face to face communication. Gestures and facial displays convey such information and are only annotated if they play a relevant role for feedback, turn management or

[1] http://www.anvil-software.de/
[2] http://nite.nis.sdu.dk/

sequencing. Turn management for example consists of the three attributes turn-gain, turn-end, and turn-hold. Turn-gain can have the values turn-accept and turn-take distinguishing between a situation where the turn is freely offered by the other participant and accepted and a situation where the speaker takes the turn although the other participant was not finished yet. For each modality, these functional attributes are annotated allowing for a multimodal analysis of communicative phenomena, i.e. an analysis of the correlations between the different modalities in performing the functions. The shape and dynamics of the facial displays and gestures are only roughly annotated with the aim to characterise and distinguish the non-verbal expressions. Gestures e.g. are annotated only by handedness (single vs. double) and trajectory (a number of simple trajectories was defined).

The standard annotation scheme for the coding of facial expressions is the facial action coding system (FACS, e.g. [12]). The basic parameter of FACS is an action unit which corresponds to a facial muscle. A facial expression can thus be described as a vector of activated action units. This scheme is very successful for describing human facial expressions, but suitable only to a limited extent for the generation of multimodal behavior because in general the animation of facial expressions for agents does not correspond directly to facial muscles. A different approach takes into account one of the current animation standards (MPEG-4). MPEG-4 defines a number of facial animation parameters that correspond to reference points in the face, such as the middle of the right eyebrow. Karpouzis et al. [18] describe how these reference points and their movement can be recognized automatically allowing for automatically anntotating facial expressions in a format that is directly suitable for the animation of a virtual character [34].

A special case is Laban movement analysis [24] which is a detailed description scheme first introduced to describe dance movements. This scheme was successfully utilized in the EMOTE model [11] to control the gestural behavior of a virtual character (see Sec. 3). Attempts to exploit this scheme also for the annotation of multimodal corpora were only a limited success due to the many dimensions which make the annotation far too tedious to be reliable [17].

2.2 Multimodal Corpora for ECA Design

So far we have described approaches that annotate information on different levels, such as the signal or the functional level, and that use corpora to achieve quite diverse goals. In the following, we will concentrate on corpora that have been collected with the goal of generating appropriate communicative behavior in virtual agents and exemplify the still diverse annotation approaches with the annotation of gesture use in human communication.

Basically, we can distinguish between a direct use of corpus data e.g. to generate animations that directly correspond to the behavior found in the data, and an indirect use of corpus data, for example to extract abstract rules that govern the planning and generation process. The direct use calls for annotations that can be mapped onto instructions for a generation component. An example includes the annotation of facial expressions using the MPEG-4 standard from

which facial expressions with an MPEG-4 compliant agent system are generated. Rule derivation for controlling the behavior planning process on the other hand requires annotations that refer to a more abstract functional level. An example is the annotation of categories of facial expressions, such as smiles or frowns, and their communicative function ideally linked to other modalities, such as speech.

Kipp et al. [22] suggest an annotation scheme for gestures that draws on the distinction between the temporal course of a gesture and its type and relies on a gesture typology introduced by McNeill [26]. The temporal course of the gesture is described by a *phase layer*. Gesture phases are preparation, hold, stroke, and retraction. Generally, the hands are brought from a resting position into the gesture space during preparation. The stroke is the phase of the gesture that carries/visualizes its meaning. Afterwards, the hands are brought back to a resting position during the retraction phase. Because gestures are often co-expressive with the speech channel, sometimes a hold is necessary. A hold is a break in the gesture execution if e.g. the utterance has not yet proceeded to the word which should be accompanied by the gesture. What kind of gesture is realized has to be annoted in a second layer, the *phrase layer*. Following McNeill, Kipp et al. distinguish between adaptors, beats, emblems, deictic, iconic, and metaphoric gestures. Adaptors comprise every hand movement to other parts of the body, such as scratching one's nose. Beats are rhythmic gestures that may emphasize certain propositions made verbally or that link different parts of an utterance. Emblems are gestures that are meaningful in themselves, i.e., without any utterance. An example is the American "OK"-emblem, where the thumb and first finger are in contact at the tips while the other fingers are extended. Deictic gestures identify referents in the gesture space. The referents can be concrete, for example, when somebody is pointing to the addressee, or they can be abstract, for example, when somebody is pointing to the left and the right while uttering the words "the good and the bad". Iconic gestures depict spatial or shape-oriented aspects of a referent, e.g., by using two fingers to indicate someone walking while uttering "he went down the street". Metaphoric gestures at last visualize abstract concepts by the use of metaphors, e.g. by employing a box gesture to visualize "a story". This is an example of the conduit metaphor that makes use of the idea of a container — in this case a container holding information. The goal of Kipp et al. is the imitation of gestural behavior by a virtual agent. To achieve this goal, information on the spatial layout of the gesture is also indispensable and coded in terms of attributes, such as handedness, straightness of trajectory, start and end positions for the stroke, three-dimensional hand position and elbow inclination.

The proposed scheme has the advantage of an economic balance between coding effort and generation effect. Using the different phase categories for annotating gestures with movement information means that in the ideal case a gesture is coded by three categories (preparation, stroke, retraction). The information of the spatial layout is annotated in a way that corresponds to traditional keyframes of animation. Thus, the data derived from the corpus can more or less

be directly used to control the gestural behavior of a virtual character resulting in an imitation of the recorded human behavior.

Abrilian et al. [1] as well as Chafai et al. [10] annotate instead the expressivity dimensions of gestural activity focusing on how a gesture is accomplished and not on what kind of gesture is used. They employ six parameters to rate the movement quality of the gestures (and of head and torso movements) in the investigated clips: activation, repetition, spatial extent, speed, strength, and fluidity. All parameters are annotated continuously between two values. Activation e.g. ranges from passive to active, speed from slow to fast, and fluidity from jerky to fluid. The annotation revealed correlations between the different parameters. For example, highly active gestural movements are often observed together with repetitive and strong movements. Chafai et al. link their expressivity parameters — fluidity, power, spatial expansion, repetition — to the above mentioned gesture phases that describe the different movement phases of a gesture. They analyse the temporal course of these parameters allowing to pinpoint irregularities and discontinuities that are interpreted as pragmatic functions in the ongoing interaction. Irregular and discontinuous movements are interpreted as attentional clues for the addressee that provide information about relevant parts of an utterance.

To sum up, the information on a different level than the actual gesture can serve useful for the generation task. The found regularities about the temporal course of the parameters and the correlations between them allow to derive rules for the generation of an agent's behavior. Moreover, expressivity parameters are not bound to a single modality, and the consistent use of a parameter, such as fluidity, over the different modalities, such as gesture, head and body movement, supports the coherent generation of believable behavior.

Rehm and André [31] describe an annotation scheme that analyzes gestures also on a more abstract functional level. The SEMMEL corpus was created to capture the relation between linguistic and nonverbal strategies of politeness. When humans interact with each other, they risk continuously threatening the face of their conversational partners, for example by showing disapproval or by putting the other person under pressure. To mitigate such face threats, humans usually rely on various politeness strategies. The seminal work by Brown and Levinson [5] contains a rich repertoire of linguistic means of politeness, but ignores multimodal aspects. Therefore, Rehm and André decided to collect their own corpus. To code politeness strategies, they follow Walker et al.'s [35] categorization into direct, approval-oriented, autonomy-oriented, and off-record strategies. In direct strategies, no redress is used, the speaker just expresses his concerns. Approval-oriented strategies are related to the positive face needs of the addressee, using means to approve of her self-image. Autonomy-oriented strategies on the other hand are related to the negative face wants of the addressee, trying to take care of her want to act autonomously. Off record strategies at last are the most vague and indirect form to address someone, demanding an active inference on the side of the addressee to understand the speaker. The coding of strategies uses a simplified version of Brown and Levinson's hierarchy

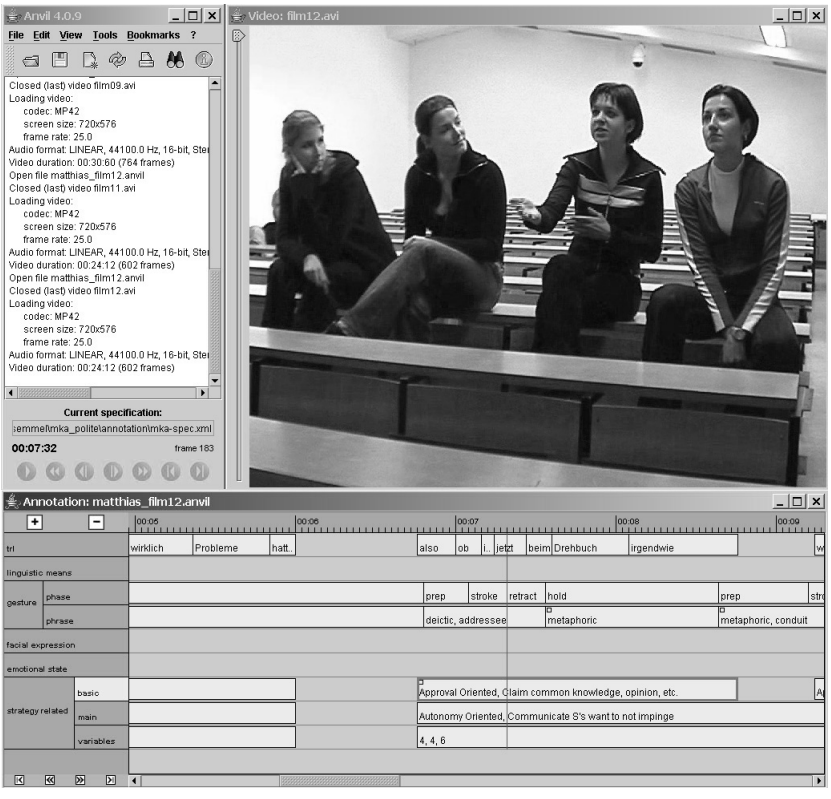

Fig. 3. Snapshot from the SEMMEL corpus. Above the video is displayed, below the annotation board.

distinguishing between seven different approval-oriented, five different autonomy-oriented, and four different off-record strategies. The coding of gestures follows Kipp's approach (see above). Accordingly, two gesture layers are distinguished: the gesture phase layer and the gesture phrase layer.

The aim of this annotation scheme was to derive information about the functional co-occurence of linguistic politeness strategies and gesture use to inform the behavior planning process of an embodied conversational agent (see Sec. 3). A statistical analysis revealed a correlation between gesture types and linguistic politeness strategies. The more indirect the strategy, the more abstract gestures (metaphoric) were used. This correlation is utilized in an overgenerate-and-filter approach to agent behavior selection (see [31] and Sec. 3).

Poggi et al. ([29]; [7]) realize a multimodal score that integrates both signal and functional level information in a single annotation scheme. Apart from the signal type and signal description which specifies the surface features of a movement, the meaning type and meaning description as well as the function of a

signal are annotated. The meaning description of a gesture is an interpretation of what can be seen e.g. raising the right hand is interpreted as *just wait, be careful*. To classify the meaning type, a semantic typology is established that distinguishes between content information, information on the speaker's mind, and self presentation (information on the speaker's identity). The function at last represents the information contribution of a signal relative to other modalities. The function of a gesture is given in comparison to the co-expressive speech signal. Five different functions are annotated. Repetition denotes that speech and gesture bear the same meaning, addition is used if additional information is given by the gesture, substitution, if a word is omitted, but its information is given by the gesture, contradiction describes the fact that contradicting information is revealed on the speech and on the gesture channel, and independence indicates that speech and gesture co-occur, but relate to different parts of the communicative plan.

Thus, Poggi et al.'s scheme explicitly codes the relations between different modalities (here exemplified for speech and gesture) on a functional level. This information is employed to model how different modalities are coordinated during the behavior planning process.

The presented list of corpora is necessarily incomplete and was selected to highlight the advantages and challenges of using multimodal corpora. These challenges include on the one hand the question of how to utilize corpus data for the control of non-verbal communicative behavior in virtual agents (or robots). This is dwelled upon in the next section. A second challenge is the question of how to discover what kinds of links exist between modalities (temporal, spatial, functional, semantic/conceptual) and how they are represented. Kipp et al. [22] give an example that this is no trivial problem. They discovered in their data that the often claimed temporal synchronicity between words and co-expressive gesture (e.g. [26]) is not as strict as they thought. Thus, other mechanisms than purely temporal relations seem to be necessary to synchronize these two modalities that are correlated on a conceptual level.

3 Multimodal Corpora for Modeling Human Behavior

To derive implementable models from empirical data, ECA researchers have analyzed various aspects of multimodal human behavior in an annotated corpus, such as the frequency of specific behaviors, the transitions between them, their co-occurrence with other behaviors as well as expressivity parameters, such as fluidity. The approaches may be distinguished by the level of the annotations (signal level versus functional level) from which models are built, the extent to which the context of a multimodal behaviour is taken into account and the employed generation mechanism which may involve a direct or indirect use of the corpus (see Sec. 2).

Some researchers generate ECA behaviors directly from motion capturing data. For instance, Stone and colleagues [33] recorded a human actor that was given a script capturing multimodal behaviors that were anticipated as relevant

to the target domain. Multimodal ECA behaviors were then generated by recombining the speech and motion samples from the recorded data. The technique produces more naturalistic behaviors than techniques that synthesize behaviors from scratch. However, the approach requires a mechanism to sequence behaviours in a coherent manner. Furthermore, the question arises of how to cope with situations for which appropriate motion capturing data and speech samples are missing. To allow for variations in the performance of an ECA, data have to be collected for different kinds of situation, personality, emotion etc. The problems may be compared to problems occurring when using a unit selection approach to synthesize speech.

Another approach is to control an agent by high-level expressivity parameters. For instance, the EMOTE system by Chi et al. [11] is based on dance annotation as described by Laban (see Sec. 2). The system is able to modify the execution of a given behavior by changing movement qualities in particular the Laban principles of effort and shape. Pelachaud and colleagues made use of six dimensions of expressivity that were derived from perceptual studies [28] (see Sec. 2.2). The advantage of both methods is that they enable the modulation of action performance at a high level of abstraction. Furthermore, they rely on a small set of parameters that may affect different parts of the body at the same time. The hypothesis behind the approaches is that behaviors that manifest themselves in various channels with consistent expressivity parameter will lead to a more believable agent behavior. Pelachaud and colleagues extract the setting of the expressivity parameters from the corresponding annotations in the corpora. They realized a so-called copy-synthesis approach which replays the annotations in the corpus using an ECA and corresponds to a direct use of a corpus.

Others perform a statistical analysis of human data to derive rules that guide the generation process. For instance, Foster and Oberlander [14] conducted experiments with a majority-choice and a weighted-choice model for the generation of facial displays. In the first case, the facial display that occurred the largest number of times in a given context is chosen. In the latter case, a random choice is made where the choice is weighted according to the relative frequency of facial displays. Context was either defined as non-existing making use of frequencies calculated over the whole corpus, as simple e.g. by considering the words in the sentence or the semantic classes of the words, or extended by taking into account also specific contextual clues like pitch-accent specifications.

Statistical models may be easily combined with an over-generate-and-filter approach as proposed in the BEAT system (e.g. [9]). The basic idea is to annotate text with plausible gestures based on rules that are derived from studies of human-human dialogue. Since it may happen that the initially proposed multimodal behaviors cannot co-occur physically, modifiable filters are then applied to trim the gestures down to a set appropriate for a particular character.

An example of an over-generate-and-filter approach includes the work by Kipp [21] who allow for different degrees of automation in behavior generation. The human author has the possibility to completely pre-author scripts that are annotated with instructions for a gesture generator. In addition, the human

author may devise rules that may be used to automatically generate annotated scripts. Finally, machine learning methods are employed to derive further rules. At runtime, all rules that fire are applied to an utterance. After that process, an utterance may contain a lot of non-verbal actions which may not occur simultaneously. The system then applies a filtering approach where manual annotations are preferred over automated ones.

Rehm and André [31] make use of an over-generate-and-filtering approach to enhance natural language utterances with suitable gestures making use of a gesticon and rules derived from the statistical analysis described in Sec. 2.2. In the first step, a probabilistic process selects a gesture type (iconic, metaphoric, etc.) based on the statistical results of the corpus study. For instance, deictic gestures may be given a higher priority than iconic gestures when suggesting non-verbal behaviors for approval-oriented strategies. The enriched natural language utterance is passed on to the animation engine. Since non-verbal behaviors are generated independently of each other, the system may end up with a set of incompatible gestures. The set of proposed gestures is therefore reduced to those gestures that are actually realized by the animation module. The findings of corpus studies may not only inform the generation, but also the filtering of gestures. For instance, iconic gestures may be filtered out with a higher probability than metaphoric gestures when realizing off record strategies.

Another question is to what extent the context in which specific multimodal behaviors occur should be taken into account when generating multimodal behaviors. One extreme would be to simply determine the frequency of multimodal behaviors, such as certain kinds of gesture. In this case, the context would not be considered at all. Instead non-verbal behaviors would be chosen based on the frequency with which they occur in the corpus. A more context-sensitive approach would be to consider the context provided by the words, by the semantic class of words or by the communicative strategy used. Kipp [21] introduces rules based on keyword spotting to annotate utterances with gestures. Rehm and André define rules that are based on the relative frequency of gestures in combination with certain strategies of politeness. Foster [13] discusses a complete representation of context which is not just defined by linguistic features, but that captures all aspects of a multimodal utterance including intonation, facial displays and gestures. Most approaches neglect the temporal context in which multimodal behaviors occur. An exception includes Kipp who proposed an approach relying on bigram estimations in order to derive typical sequences of two-handed and one-handed gestures.

Usually, rules for selecting multimodal behaviors are manually extracted from a corpus. Kipp [21] discusses the use of machine learning techniques to derive rules automatically. Unfortunately, such an approach requires a large amount of data - especially if the context in which a rule may be applied is captured as well. Therefore, Kipp [21] does not rely on recordings of humans, but on manually authored presentations. Unlike most previous work, he does not emulate multimodal human-human communicative behavior, but tries to derive design guidelines of human animators.

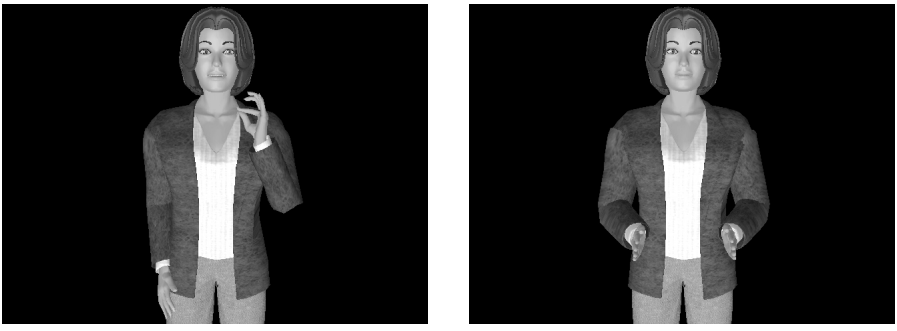

Fig. 4. Greta realizing an iconic vs. a metaphoric gesture

4 Evaluation of Corpus-Based Approaches to Generation

The question arises of how to evaluate ECAs whose behaviour is driven by empirical data. One possibility is to investigate to what extent the derived models enable a prediction of human-like behaviours. Such an approach has been used to evaluate the performance of an approach to the data-driven generation of empathetic facial displays by Foster and Oberlander [14]. To compare the performance of the models against the corpus, they employed 10-fold-cross-validation. For each fold, 90% of the data were used to derive a behavior model and 10% of the data were used to validate the models. For each sentence, they measured precision and recall by comparing the predicted facial displays with the actual displays in the corpus and then averaged the scores across the sentence. Their evaluation revealed that majority-choice models resulted into higher precision and recall than weighted-choice models. A similuar evaluation methods was proposed by Kipp who partioned his corpus into a training (60%) and a test test(40%) and measured precision and recall for manual annotations of non-verbal behaviors. Instead of comparing predicted behaviors with actual behaviors, Buisine and colleagues [6] conduct a perceptive evaluation where humans judge to what extent the replayed behaviors by the agent ressemble the original behaviors. In particular, they investigated whether humans are able to detect blends of emotions in an embodied agent.

Of course, a great similarity to human-like behaviors or to pre-authored behaviors does not necessarily mean that the resulting agent is positively perceived by a human observer. To shed light on this question, perception studies are performed which compare how human observers respond to ECAs whose behaviors are informed by an empirical model in comparison to ECAs with randomized multimodal behaviors. Garau and colleagues [16] as well as Lee and colleagues [25] investigate the effect of informed eye gaze models on the perceived quality of communication. Both research teams observed a superiority of informed eye gaze behaviors over randomized eye gaze behaviors. Rehm and André [31] 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). They presented subjects with two variants of utterances including criticism: one in which the criticism was accompanied by a gesture of the concrete, and the other one in which the criticism was accompanied by an abstract gesture (see Fig. 4). The subjects then had to rate the perceived tone of politeness. 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 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. That is well designed gestures may strengthen, but badly designed gestures weaken pragmatic effects. The studies by Foster and Oberlander [14] enable a direct comparison of prediction-based evaluation methods and perception-based evaluation methods. Foster and Oberlander investigated how a talking head that was driven by different variants of a generation algorithm was perceived by human observers. They observed that humans seem to prefer behaviors that follow a weighted-choice model over behaviors that follow a majority-choice model. They conclude that humans prefer non-verbal behaviors that reflect more of the variations in the corpus even if the non-verbal behaviors that accompany specific sentences did not correspond to the non-verbal behaviors in the corpus. The results of their studies show that a perception-based evaluation method may indeed lead to different results than a prediction-based evaluation method. Furthermore, they noticed that the users' opions regarding the acceptability of facial displays may vary systematically. In particular, they observed interesting gender-specific differences. All preferences for the weighted-choice models were expressed by the female subjects while the male subjects did not have any preference at all or seem to slightly prefer the majority-choice models.

Besides asking users directly for their impression of the agent, researchers investigated whether an agent that is based on an empirically driven model changes the nature of the interaction. Garau and colleagues [16] found that model-based eye gaze improved the quality of communication when a realistic avatar was used. For cartoonish avatars, no such effect was observed. A study by Nakano and colleagues [27] revealed that an ECA with a grounding mechanism seems to encourage more non-verbal feedback from the user than a system without any grounding mechanism. Sidner and colleagues [32] showed that users are sensitive to a robot's conversational gestures and establish mutual gaze with it even if the set of communicative gesture of the robot is strongly limited.

In contrast to the work above, Rehm and André [30] focus on a direct comparison of human-agent and human-human interaction. The objective of their work was to investigate whether humans behave differently when interacting with an agent as opposed to interacting with another human. As a first step, they focused on gaze behaviors as an important predictor of conversational attention. To this end, they recorded users interacting with a human and a synthetic game partner in a game of dice called Mexicali (see Fig. 5). The scenario allowed

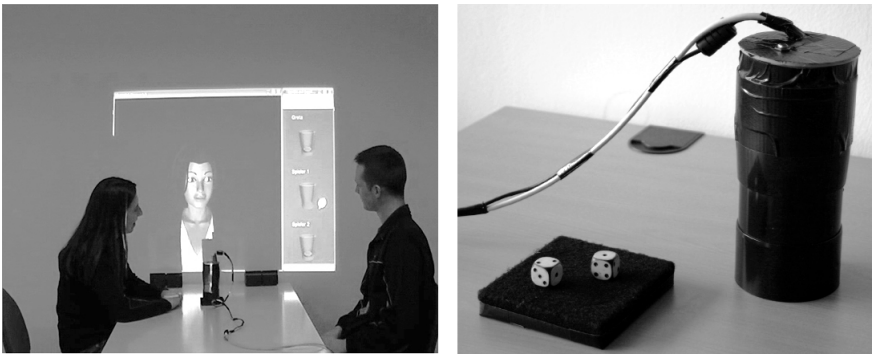

Fig. 5. The Gamble system and the CamCup

them to directly compare gaze behaviors in human-human with gaze behaviors in human-agent interaction. On the one hand, they were able to confirm a number of findings about attentive behaviors in human-human conversation. For instance, their subjects spent more time looking at an individual when listening to it than when talking to it - no matter whether the individual was a human or a synthetic agent. Furthermore, the addressee type (human vs. synthetic) did not have any impact on the duration of the speaker's gaze behaviors towards the addressee. Even though the game was in principle playable without paying any notice to the agent's nonverbal behaviors, the users considered it as worthy of being attended to. While the users' behaviors in the user-as-speaker condition were consistent with findings for human-human conversation, we noticed differences for the user-as addressee condition. People spent more time looking at an agent that is addressing them than at a human speaker. Maintaining gaze for an extended period of time is usually considered as rude and impolite. The fact that humans do not conform to social norms of politeness when addressing an agent seems to indicate that they do not regard the agent as an equal conversational partner, but rather as a (somewhat astonishing) artefact that is able to communicate. This attitude towards the agent was also confirmed by the way the users addressed the agent verbally.

5 Conclusion

Annotated multimodal corpora serve as useful tools for developing embodied conversational agents with a rich repertoire of multimodal communicative behaviors. We have seen how corpora are employed in the study of human behavior with the aim of simulating human communication. Different approaches were presented on how the information derived from such corpora is utilized to control the behavior generation process for an agent. And finally we have exemplified that corpora can even play a role in evaluating human-agent interactions.

Although the use of corpora in the development process of embodied conversational agents has increased significantly, a number of open research issues remain. Standardized schemes are not easy to establish due to the different levels of granularity possible in the annotation process. Despite of new annotation tools, the collection and annotation of corpora is still cumbersome and time-consuming. A great challenge for the future is therefore de-contextualization of multimodal data and their automated adaptation to a new context.

References

1. Abrilian, S., Martin, J.-C., Buisine, S., Devillers, L.: Perception of movement expressivity in emotional tv interviews. In: HUMAINE Summerschool (2006)
2. Allwood, J., Nivre, J., Ahlsén, E.: On the semantics and pragmatics of linguistic feedback. Journal of Semantics 9, 1–26 (1992)
3. Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., Paggio, P.: The mumin annotation scheme for feedback, turn management and sequencing. In: Gothenburg Papers in Theoretical Linguistics 92: Proceedings from The Second Nordic Conference on Multimodal Communication, pp. 91–109 (2005)
4. Argyle, M., Cook, M.: Gaze and mutual gaze. Cambridge University Press, Cambridge (1976)
5. Brown, P., Levinson, S.C.: Politeness — Some universals in language usage. Cambridge University Press, Cambridge (1987)
6. Pelachaud, C., Martin, J.-C., Niewiadomski, R., Abrilian, S., Devillers, L., Buisine, S.: Perception of Blended Emotions: From Video Corpus to Expressive Agent. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 93–106. Springer, Heidelberg (2006)
7. Caldognetto, E.M., Poggi, I., Cosi, P., Cavicchio, F., Merola, G.: Multimodal Score: an ANVIL Based Annotation Scheme for Multimodal Audio-Video Analysis. In: Proceedings of the LREC-Workshop on Multimodal Corpora, pp. 29–33 (2004)
8. Cassell, J.: Body Language: Lessons from the Near-Human. In: Riskin, J. (ed.) The Sistine Gap: History and Philosophy of Artificial Life, University of Chicago Press, Chicago (in press)
9. Cassell, J., Vilhjalmsson, H., Bickmore, T.: BEAT: The Behavior Expression Animation Toolkit. In: Proceedings of SIGGRAPH 2001, Los Angeles, CA, pp. 477–486 (2001)
10. Chafai, N.E., Pelachaud, C., Pelé, D.: Analysis of gesture expressivity modulations from cartoon animations. In: Proceedings of the LREC-Workshop on Multimodal Corpora (2006)
11. Chi, D., Costa, M., Zhao, L., Badler, N.: The EMOTE Model for Effort and Shape. In: Proceedings of SIGGRAPH, pp. 173–182 (2000)
12. Ekman, P., Rosenberg, E. (eds.): What the Face Reveals: Basic & Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS). Oxford University Press, Oxford (1998)
13. Foster, M.E.: Issues for corpus-based multimodal generation. In: Proceedings of the Workshop on Multimodal Output Generation (MOG 2007), pp. 51–58 (2007)
14. Foster, M.E., Oberlander, J.: Data-driven generation of emphatic facial displays. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL), pp. 353–360 (2006)

15. Frey, S., Hirschbrunner, H.P., Florin, A., Daw, W., Crawford, R.A.: A unified approach to the investigation of nonverbal and verbal behavior in communication research. In: Doise, W., Moscovici, S. (eds.) Current Issues in European Social Psychology, pp. 143–199. Cambridge University Press, Cambridge (1983)
16. Garau, M., Slater, M., Vinayagamoorthy, V., Brogn, A., Steed, A., Sasse, M.A.: The Impact of Avatar Realism and Eye Gaze Control on Perceived Quality of Communication in a Shared Immersive Virtual Environment. In: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 529–536 (2003)
17. Höysniemi, J., Hämäläinen, P.: Describing children's intuitive movements in a perceptive adventure game. In: Martin, J.-C., Os, E.D., Kühnlein, P., Boves, L., Paggio, P., Catizone, R. (eds.) Multimodal Corpora: Models Of Human Behaviour For The Specification And Evaluation Of Multimodal Input And Output Interfaces, pp. 21–24 (2004)
18. Karpouzis, K., Raouzaiou, A., Drosopoulos, A., Ioannou, S., Balomenos, T., Tsapatsoulis, N., Kollias, S.: Facial expression and gesture analysis for emotionally-rich man-machine interaction. In: Sarris, N., Strintzis, M. (eds.) 3D Modeling and Animation: Synthesis and Analysis Techniques, pp. 175–200. Idea Group, USA (2004)
19. Kendon, A.: Some functions of gaze direction in social interaction. Acta Psychologica 32, 1–25 (1967)
20. Kendon, A.: Gesture — Visible Action as Utterance. Cambridge University Press, Cambridge (2004)
21. Kipp, M.: Creativity meets Automation: Combining Nonverbal Action Authoring with Rules and Machine Learning. In: Gratch, J., et al. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 230–242. Springer, Heidelberg (2006)
22. Kipp, M., Neff, M., Albrecht, I.: An annotation scheme for conversational gestures: How to economically capture timing and form. In: Proceedings of the LREC-Workshop on Multimodal Corpora, pp. 24–27 (2006)
23. Knudsen, M.W., Martin, J.-C., Dybkjr, L., Ayuso, M.J.M., Bernsen, N.O., Carletta, J., Heid, U., Kita, S., Llisterri, J., Pelachaud, C., Poggi, I., Reithinger, N., van Elswijk, G., Wittenburg, P.: ISLE Natural Interactivity and Multimodality Working Group Deliverable D9.1: Survey of Multimodal Coding Schemes and Best Practice (2002), URL(07.02.07): `http://isle.nis.sdu.dk/reports/wp9/D9.1-7.3.2002-F.pdf`
24. Lamb, W., Watson, E.: Body Code: The Meaning in Movement. Routledge & Kegan Paul, London (1979)
25. Lee, S.P., Badler, J.B., Badler, N.I.: Eyes alive. In: Proceedings of the 29th annual conference on Computer graphics and interactive techniques, pp. 637–644 (2002)
26. McNeill, D.: Hand and Mind — What Gestures Reveal about Thought. The University of Chicago Press, Chicago, London (1992)
27. Nakano, Y.I., Reinstein, G., Stocky, T., Cassell, J.: Towards a Model of Face-to-face Grounding. In: Proceedings of the Association for Computational Linguistics, Sapporo, Japan, July 1–12, pp. 553–561 (2003)
28. Pelachaud, C.: Multimodal expressive embodied conversational agents. In: Proceedings of ACM Multimedia, pp. 683–689 (2005)
29. Poggi, I., Pelachaud, C., Magno Caldognetto, E.: Gestural Mind Markers in ECAs. In: Gesture-Based Communication in Human-Computer Interaction, pp. 338–349. Springer, Heidelberg (2004)
30. André, E., Rehm, M.: Where Do They Look? Gaze Behaviors of Multiple Users Interacting with an Embodied Conversational Agent. In: Panayiotopoulos, T., Gratch, J., Aylett, R.S., Ballin, D., Olivier, P., Rist, T. (eds.) IVA 2005. LNCS (LNAI), vol. 3661, pp. 241–252. Springer, Heidelberg (2005)

31. Rehm, M., André, E.: Informing the design of agents by corpus analysis. In: Nishida, T., Nakano, Y. (eds.) Conversational Informatics, John Wiley & Sons, Chichester (2007)
32. Sidner, C.L., Lee, C., Kidd, C., Lesh, N., Rich, C.: Explorations in engagement for humans and robots. Artificial Intelligence 166(1–2), 140–164 (2005)
33. Stone, M., DeCarlo, D., Oh, I., Rodriguez, C., Stere, A., Lees, A., Bregler, C.: Speaking with Hands: Creating Animated Conversational Characters from Recordings of Human Performance. ACM Transactions on Graphics 23(3), 506–513 (2004)
34. Tsapatsoulis, N., Raouzaiou, A., Kollias, S., Cowie, R., Douglas-Cowie, E.: Emotion Recognition and Synthesis based on MPEG-4 FAPs. In: Pandzic, I., Forchheimer, R. (eds.) MPEG-4 Facial Animation, pp. 141–167. John Wiley & Sons, Chichester (2002)
35. Walker, M.A., Cahn, J.E., Whittaker, S.J.: Improvising Linguistic Style: Social and Affective Bases for Agent Personality. In: Proceedings of AAMAS 1997, pp. 96–105 (1997)

Appendix: Where to Find Multimodal Corpora

Some of the corpora mentioned in this article can be accessed by interested researchers. The specifics concerning data protection and access regularities vary from corpus to corpus. A good starting point to search for available multimodal corpora is the website of the Humaine Association (former Humaine Network of Excellence): http://emotion-research.net/wiki/Databases. Mostly linguistic corpora are available from the European Language Resources Association (ELRA, http://catalog.elra.info/) or from the Lingustic Data Consortium (LDC, http://www.ldc.upenn.edu/).

At last, we would like to mention explicitely three exemplary corpora. The AMI corpus contains around 100 hours of multiparty meeting interactions and is freely accessible (http://corpus.amiproject.org/). The Smartkom corpus is a German corpus of a Wizard of Oz experiment on human-computer interactions in an information kiosk scenario. There is a service charge for accessing this corpus (http://www.bas.uni-muenchen.de/Bas/BasSmartKomHomeeng.html). The CUBE-G corpus contains around 20 hours of culture-specific interactions from Germany and Japan in three standardized scenarios (first meeting, negotiation, status difference). Information on this corpus can be found under http://mmwerkstatt.informatik.uni-augsburg.de/projects/cube-g/.