In A. Esposito et al. (Eds.): Multimodal Signals: Cognitive and Algorithmic Issues
Lecture Notes in Artificial Intelligence 5398, pp. 47-59, 2009
Springer-Verlag Berlin Heidelberg 2009

Representing Communicative Function and Behavior in Multimodal Communication

Hannes Högni Vilhjálmsson

Center for Analysis and Design of Intelligent Agents (CADIA)
School of Computer Science, Reykjavík University
Kringlan 1, 103 Reykjavík, Iceland
hannes@ru.is

Abstract. In this paper I discuss how communicative behavior can be represented at two levels of abstraction, namely the higher level of communicative intent or function, which does not make any claims about the surface form of the behavior, and the lower level of physical behavior description, which in essence instantiates intent as a particular multimodal realization. I briefly outline the proposed SAIBA framework for multimodal generation of communicative behavior, which is an international research platform that fosters the exchange of components between different systems. The SAIBA framework currently contains a first draft of the lower level Behavior Markup Language (BML) and is starting work on the higher level Function Markup Language (FML). I also briefly explain the usefulness of this distinction by using examples of several implemented systems that each draws different strengths from it. These systems range from autonomous conversational agents to computer mediated communication.

Keywords: Communicative behavior, communicative function, multimodal communication, embodied conversational agents, mediated communication

1. Introduction

There is an international community of researchers working on creating what has been called embodied conversational agents. These are autonomous agents that have a human-like embodiment, either graphical or physical, and possess the skill to engage people in face-to-face conversation [4]. One of the main challenges of this research is the proper coordination of verbal and nonverbal behavior, since face-to-face communication relies on the careful production of a wide range of multimodal cues that serve many important communicative functions. For example we often user our body to indicate that we would like the chance to say something and then while we speak, we often emphasize parts of our sentences with synchronized hand movement.

Research groups that work on embodied conversational agents at a high level, for example dialogue planning, have typically had to build the agent's body and implement an animation system from scratch that is capable of producing the right variety of communicative behavior that dialogue requires. Similarly, those research

groups looking into sophisticated animation techniques for virtual humans or robots, have had a hard time finding high level "brains" that properly fit on top of their expressive bodies. Therefore the best looking embodiments often end up being animated through scripted behavior.

Building a fully functional and beautifully realized embodied conversational agent that is completely autonomous, is in fact a lot more work than a typical research group can handle alone. It may take individual research groups more than a couple of years to put together all the components of a basic system, where many of the components have to be built from scratch without being part of the core research effort.

Recognizing that collaboration and sharing of work between research groups would get full conversational systems up and running much quicker and reduce the reinvention of the wheel, an international group of researchers started laying the lines for a framework that would help make this happen. In particular, the emphasis of the group was on defining common interfaces in the multimodal behavior generation process for embodied conversational agents. While the seeds for this work were planted at the 2002 AAMAS workshop "Embodied conversational agents – let's specify and evaluate them!", the first official working meeting of the steering group took place at Reykjavik University in 2005 under the title "Representations for Multimodal Generation".

At this meeting the members of the group pooled their knowledge of various full agent systems and looked into what processes seemed common to all of them, identifying possible areas of reuse and employment of standard interfaces. The group proposed the so-called SAIBA framework as a result [12], [15]. This framework divides the overall behavior generation process into three sub-processes, each bringing the level of communicative intent closer to actual realization through the agent's embodiment (see Figure 1).

In this framework lies an opportunity to define one interface at the high level, between intent planning and behavior planning, and another interface at the lower level, between behavior planning and behavior realization. The group then set out to start defining these interfaces, called Function Markup Language [10] and Behavior Markup Language [12], [14], respectively.

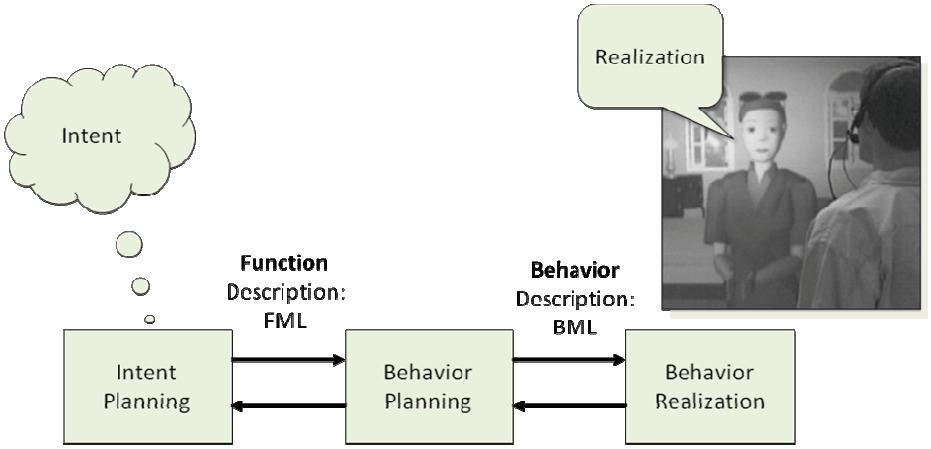

Figure 1: The SAIBA framework for multimodal behavior generation, showing how the overall process consists of sub-processes at different levels of abstraction, starting with communicative intent and ending in actual realization in the agent's embodiment.

2. Communicative Function vs. Behavior

To appreciate and better understand the difference between communicative function, which is specified with FML, and communicative behavior, specified with BML, let's look at an example. This example is relatively extreme and is constructed solely for illustration, but not to show an actual system implementation.

If you want to tell someone a story about something that happened, you would first shape your communicative intent regarding the story, for example picking a recipient and then organizing major topics and key points in your mind based on who the recipient is. If the recipient is not present, you may then decide to realize this communicative intent in writing. The intent is then transformed into concrete form, governed by rules of written discourse (see Figure 2, top). However, if the recipient suddenly shows up in person, you may decide to discard the written realization and instead engage in oral realization of that very same communicative intent. Now the delivery is governed by different rules, essentially creating a different concrete

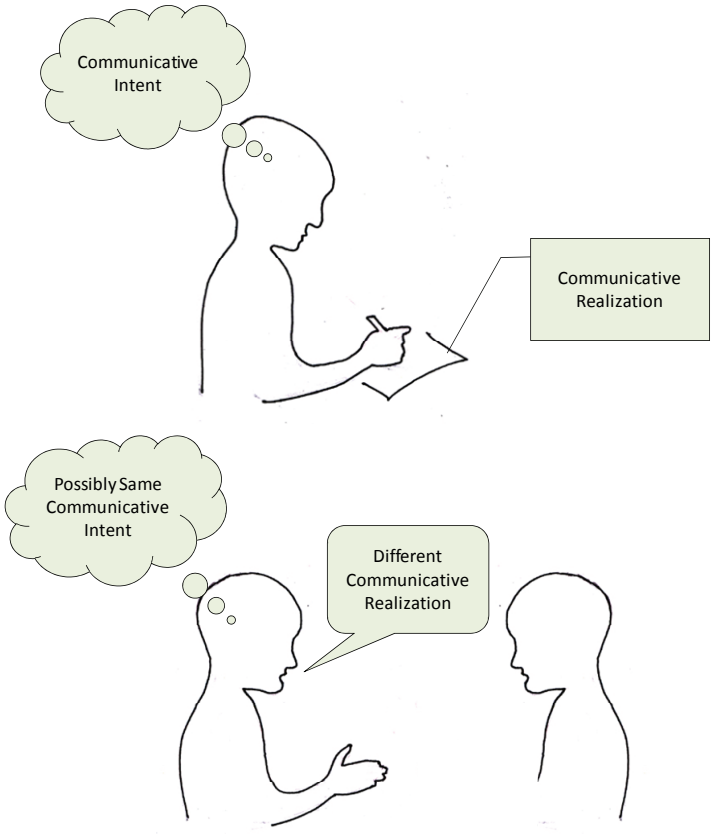

Fig. 2: The same communicative intent could get transformed into written form (top) or oral form (bottom)

rendering of the original intent (see Figure 2, bottom). The more abstract form of the story stays intact, while the concrete form changes. This seems to indicate that it is useful to distinguish between these two levels of representation.

To further illustrate this distinction, let's imagine that you wish to tell your friend Alex that you love dark chocolate but hate white chocolate, and want to emphasize this difference. Representing this at the level of communicative intent could produce something like the (made up) formal notation shown on the left side in Figure 3. This representation indicates the communicative functions that you wish to perform, starting with a function intended to introduce the new topic of chocolates. You then

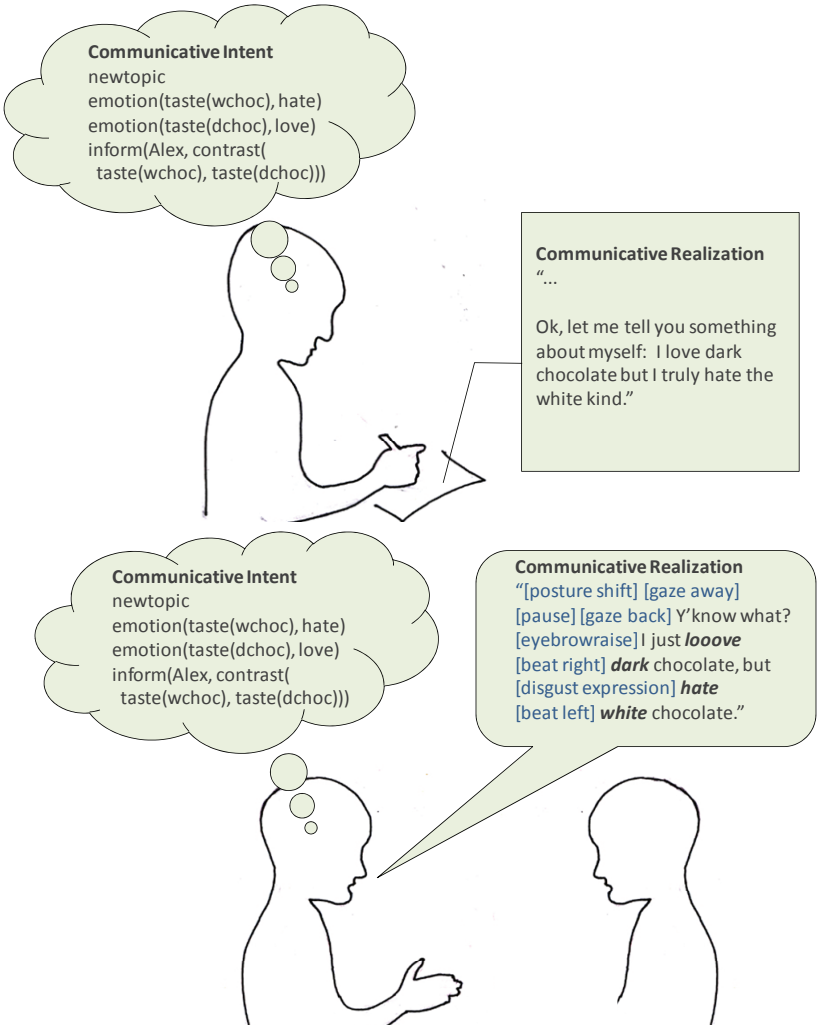

Fig. 3: A story about contrasting tastes gets realized in a letter (top) and face-to-face (bottom). Linguistic devices realize functions in the former, but nonverbal devices also serve functions in the latter.

have a representation of your likes and dislikes, and finally the communicative function of informing Alex of the contrast between your tastes. When you turn this into writing, each function gets realized through some written device, some according to the conventions of letter writing, others through more general rules of grammar and coherent discourse. For example, the introduction of a new topic may manifest itself as the start of a new paragraph and the word "Ok" at the beginning of the first sentence (see top-right side in Figure 3).

If Alex shows up in person and oral delivery is used instead, each of the communicative functions now get realized through a new set of multimodal devices that become available. Of course some of them still involve the production of coherent words, but both the availability of new forms of multimodal expression and the new kind of discourse, call for a surprisingly different realization. For example, the switch to a new topic may now be accompanied by a visible shift in posture, and instead of relying completely on the words to carry the contrast between the likes and the dislikes, hand gestures play an important role (see right side in Figure 3).

This example shows what gets produced is the realization that best serves the original intent in a particular communication environment, given its unique constraints and conventions. In embodied conversational agents, it is helpful to have a way to represent communicative intent before the exact form is chosen, leaving behavior planning to a dedicated process, based on the particular context of delivery.

It is worth noting that dividing the production of human-like behavior into two distinct levels of representation has been attempted across various fields of study, and these levels have received different names. In Figure 4 we see on the left hand side various terms that have been used to describe communicative function at the higher level of abstraction, while on the right we have terms that describe the more concrete realization. The words in the two lists roughly correspond to each other, so that for example the words meaning/sign often occur together. It is important to realize that these different terms exist and that it is quite unlikely that a single pair will become the standard usage. There are always going to be some slight differences in the interpretations of these words, but their purpose is the same: To create meaningful and useful levels of abstraction to better understand or produce discernable acts.

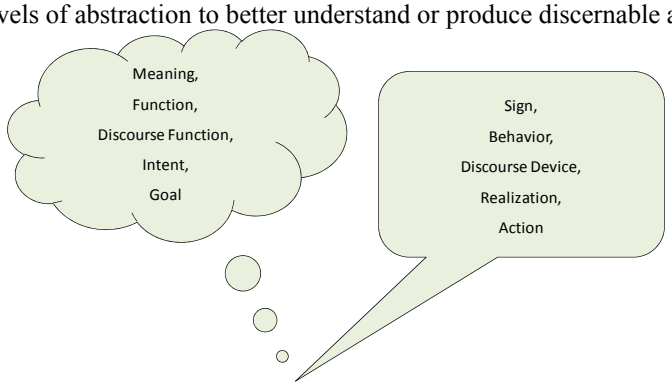

Fig. 4: Terms used to describe communicative function at higher level of abstraction (left) and a more concrete level of realization (right)

3. The Relationship

There is a large body of work that attempts to describe human behavior in terms of its functions in communication [2], [8], [9], [11]. By thinking about the functions, we start to see the bigger picture and construct whole systems of behavior that somehow seem connected or serve similar purpose. When looking at face-to-face conversation in particular, communicative functions seem to generally fall into one of three broad categories illustrated in Figure 5 [10]. The first category (A) has to do with the establishing, maintaining and closing the communication channel between participants. A communication channel is merely a metaphor for the social contract that binds people together in the common purpose of conversing. This category of functions has received many names, of which *interactional*, *envelope* and *management* functions have been popular [6], [11], [13]. Some examples of functions that fall into this category are given in Table 1A. The second category (B) covers the actual content that gets exchanged across a live communication channel. Given that the functions in category A are doing their job, those in B are made possible. Typically this is the deliberate exchange of information, which gets organized and packaged in chunks that facilitate uptake. The various functions in this category can be divided across the different organizational levels, from the largest organizational structure of discourse down to the specification of each proposition. This is shown in Table 1B. If the second category covers only deliberate exchange of information, then another category is needed that takes care of the various functions contributing to visible behavior giving off information, without deliberate intent. The third category (C) is perhaps a bit of a catch-all category, but generally deals with functions describing mental states and attitudes, which in turn may affect the manner in which other functions get realized or give rise to their own independent behavior. Table 1C lists some examples of these.

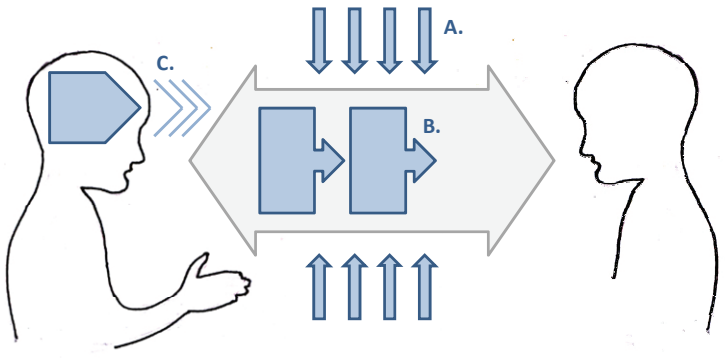

Fig. 5: General categories of communicative functions: Interaction Functions (A), Content Functions (B) and Mental States and Attitudes (C)

Table 1A: INTERACTION FUNCTIONS	
Function Category	**Example Functions**
Initiation / Closing	*react, recognize, initiate, salute-distant, salute-close, break-away, …*
Turn-taking	*take-turn, want-turn, yield-turn, give-turn, keep-turn, assign-turn, ratify-turn, …*
Speech-Act	*inform, ask, request, …*
Grounding	*request-ack, ack, repair, cancel, …*

Table 1B: CONTENT FUNCTIONS	
Function Category	**Example Functions**
Discourse Structure	*topics and segments*
Rhetorical Structure	*elaborate, summarize, clarify, contrast, …*
Information Structure	*rheme, theme, given, new, …*
Propositions	*any formal notation (e.g. "own(A,B)")*

Table 1C: MENTAL STATE AND ATTITUDE FUNCTIONS	
Function Category	**Example Functions**
Emotion	*anger, disgust, fear, joy, sadness, surprise, …*
Interpersonal Relation	*framing, stance, …*
Cognitive Processes	*difficulty to plan or remember*

Now that we have seen that each conversation is governed by a system of communicative functions carried out by participants, we must ask next how they manifest themselves as discernable behavior. What are the rules that map functions to supporting behaviors?

In order to answer this question, scientists generally need to engage in a four step process for each of the communicative functions they wish to map. The first step is the literature search where it is established whether this function has been studied before and existing empirical results provide enough evidence for certain behavior mapping rules. Keep in mind that prior studies may have been performed in situations that are different from those currently being modeled, and therefore it may be necessary to make a few assumptions about how well the data generalizes, or repeat the study in a different setting. If a study is to be repeated or a new study is called for, the second step involves gathering new data. Data is gathered with a good video and audio recording of a communicative event that strikes a balance between being completely naturally occurring and somewhat engineered to ensure a context that matches the one being modeled. But even if the situation is engineered, the subjects being recorded should not be aware of what functions and behaviors are being studied to avoid biased behavior. Once the data has been gathered, the third

step is the coding of that data. This is done in at least two distinct passes, preferably with two separate coders. One pass through the data only annotates what is going on at the level of communicative intent or function. For example, one could annotate all places where it is apparent from the transcript that the subjects are changing the topic. On a second separate pass, a particular behavior is annotated, using only the modality of that behavior. For example, looking through the video with the audio turned off and without seeing any previous annotation or transcript, one can annotate the occurrences of head nods. The last step is then to look at these annotations together and understand how well, if at all, the annotated function seems to predict the occurrence of the annotated behavior. Table 4 shows a few examples of communicative functions and how they have been correlated with visible behavior.

Table 2: EXAMPLE MAPPING RULES

rheme/new	Gestures are more likely to occur with new material than given material [5]
emphasis	Emphasis commonly involves raising or lowering of the eyebrows [2]
new-topic	People often change posture when they change the topic of conversation [3]
give-turn	Speaker usually selects next speaker with gaze near the end of their own turn (Kendon, 1990)
take-turn	Speakers generally break eye-contact at turn beginning [1]
request-ack/ack	Speaker looks at listeners and raise their eyebrows when they want feedback and listeners raise eyebrows in response [8]

Strictly speaking, it is not correct to talk about the mapping from function to behavior as absolute rules. In fact, they are merely regularities that have been discovered empirically, but there are always going to be some exceptions. We tend to talk about them as rules because that is often how they are implemented in embodied conversational agents. When we apply these rules inside the agents, we are not only assuming they occur without exception, but we are also assuming a certain context that makes the rule applicable, for example a particular culture or social situation (see Figure 6).

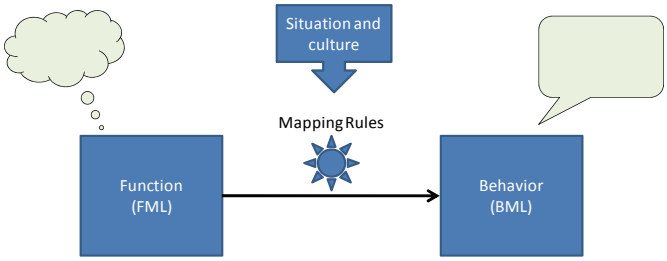

Fig. 6: Rules that map functions to behavior assume a certain context like the social situation and culture

4. Applications

To provide an idea of how a SAIBA-like framework with a clear division between function and behavior can contribute to working systems, we'll take a quick look at three different applications where FML and BML play an important role.

The first one is a classic embodied conversational agent where we have a human user interacting with a graphical representation of the agent on a large wall-size display. The agent is capable of receiving multimodal input and produce multimodal output in near real-time, creating the sense of full face-to-face conversation with the human user. In this application, a pipeline architecture is possible where the multimodal input from the user is first described using something like BML (see Figure 7). A special Understanding Module interprets the behavior in the current context (including situation and culture) and produces an abstract description of the user's communicative intent in FML. The agent's decisions about how to respond are then made purely at the abstract level inside a central Decision Module. Decisions are similarly described in FML, and finally a Generation Module maps the agent's intended functions into behavior, visible to the human user, using the current context. Rea, the real-estate agent, was developed with this architecture [6]. BML and FML did not exist at the time, but corresponding representations were used based on the KQML messaging standard.

One benefit of creating an agent in this fashion, is that the abstract decision making module, which can be quite complex, is isolated from the surface form of behavior, both on the input side and the output side. It may therefore be easier to adapt this agent to interacting in different cultures or even use different means for communication, like calling the user on the phone instead of interacting face-to-face.

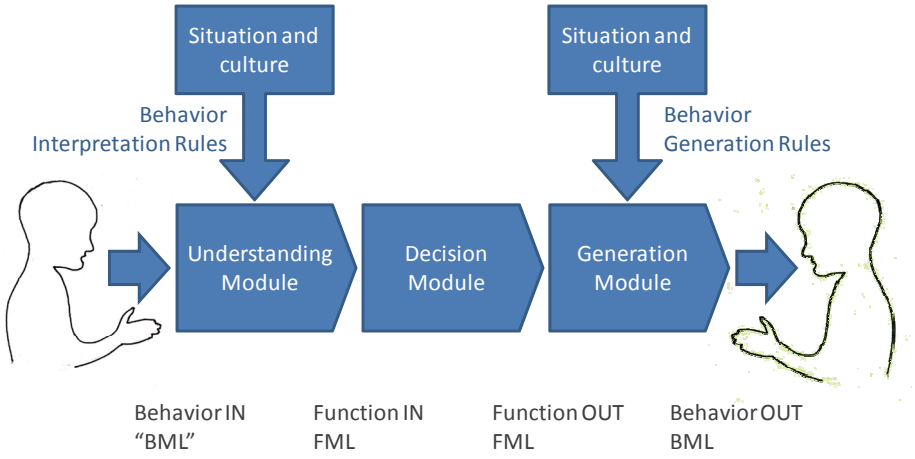

Fig. 7: An embodied conversational agent architecture where the central Decision Module only deals with an abstract representation of intent

Another application is a tool for animators that wish to make a character speak lines from a script and produce appropriate and synchronized nonverbal behavior. The system can then analyze the text to be spoken and look for various linguistic devices (i.e. behavior) that have been shown to be associated with particular communicative functions (see Figure 8). These functions are then annotated in the text using FML. Finally a Generation Module receives the annotated text and applies mapping rules to produce BML that carries out those functions, just like the example with the conversational agent above. Since the functions line up with certain parts of the spoken text, the behaviors that support those functions will coincide with the same parts. The BML can be used to drive animation in real-time or schedule behavior clips in an off-line 3D rendering tool like Maya. It can also be used to simply annotate the script with behavior annotations that a human animator can read and use as suggestions. The BEAT toolkit used this architecture, and did in fact use an early version of BML and FML for annotation purposes [7].

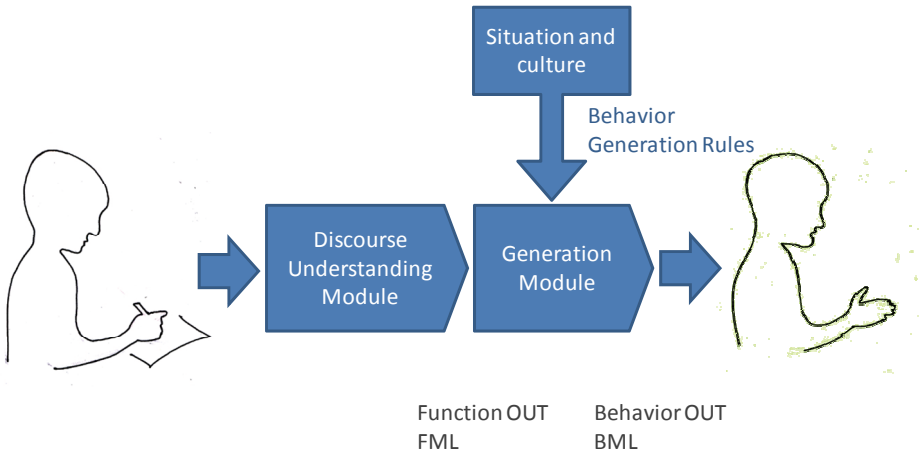

Fig. 8: A script is annotated with communicative functions that a generation module uses to produce appropriate animation

The final application mention here is real-time computer-mediated-communication. In this situation we have people communicating with other people over some kind of a communication channel that is narrower than a typical face-to-face conversation. For example, a person may be sending written messages to a friend over instant messaging. By using the same technique as the animation toolkit described above, the mediating system could analyze the sender's message and annotate communicative functions (see Figure 9). Once the message arrives at the recipient's computer, a Generation Module can look at the message along with the function annotation, and generate a realization of the message that best supports the intended communication. For example, if the sender has an animated avatar, the Generation Module could deliver the message as if it was being spoken by the avatar and produce all the supporting nonverbal behavior according to the FML to BML mapping rules.

Interestingly, these mapping rules could be different on different recipient computers, so the same message sent to two different continents could result in two different avatar performances on recipient computers, reflecting the local culture and setting. The Spark system, for group communication and collaboration, is implemented in this manner and uses an early version of BML and FML [16].

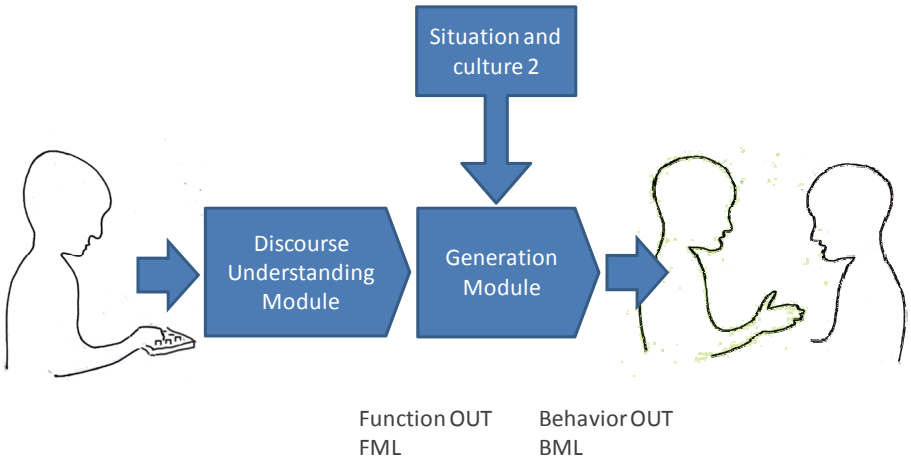

Fig. 9: Communicative functions annotated in a real-time chat message help produce an animated avatar that augments the delivery

5. Conclusions

There is a growing group of researchers that conspires to build a horde of virtual humans capable of mingling socially with humanity. Like in all good conspiracy plots, a plan to make this possible is already underway, namely the construction of a common framework for multimodal behavior generation that allows the researchers to pool their efforts and speed up construction of whole multimodal interaction systems. The overall framework is called SAIBA, and within it, two important interfaces are being defined: FML and BML. FML describes communicative function at the level of intent, regardless of surface behavior, whereas BML describes the surface form which then is realized by an animation engine. This division into two levels of abstraction has helped scientists to make sense of human social and linguistic behavior in the past, but it is also very useful in modern applications that either participate in or support multimodal communication. We are still at an early stage of defining SAIBA and its interfaces, but the growing interest and some promising early tools, are indication that we may be onto something important. Only time will tell.

The purpose of this paper is to clarify the distinction between function and behavior, as well as outlining some related methodologies and system architectures. While providing some answers, it is also meant to provoke questions and get people interested in joining the ongoing discussion within the SAIBA community. A good starting point are the online wiki pages and forums[1].

6. Acknowledgements

The talk that this paper is based on was given at the "Multimodal Signals: Cognitive and Algorithmic Issues" international school supported by the EU COST Action number 2102. Special thanks go to the COST 2102 management chair, Professor Anna Esposito. The ideas presented in this paper are drawn from collaborations with numerous researchers, especially at RU/CADIA, USC/ISI and MIT, and of course in the SAIBA community. My work is supported by the Icelandic Centre for Research grant number 080027011.

7. References

1. Argyle, M., & Cook, M.: Gaze and mutual gaze. Cambridge University Press, Cambridge (1976)
2. Argyle, M., Ingham, R., Alkema, F. et al.: The Different Functions of Gaze. Semiotica, (1973)
3. Cassell, J., Nakano, Y., Bickmore, T. et al.: Non-Verbal Cues for Discourse Structure. Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics, (2001) 106-115
4. Cassell, J., Sullivan, J., Prevost, S. et al. (eds.): Embodied conversational agents. MIT Press, Cambridge, MA (2000)
5. Cassell, J., Steedman, M., Badler, N. et al.: Modeling the Interaction between Speech and Gesture. (1994)
6. Cassell, J., Bickmore, T., Campbell, L. et al.: More than just a Pretty Face: Conversational Protocols and the Affordances of Embodiment. Knowledge-Based Systems, Vol. 14. Elsevier (2001) 55-64
7. Cassell, J., Vilhjalmsson, H., Bickmore, T.: BEAT: The Behavior Expression Animation Toolkit. SIGGRAPH, ACM Press (2001) 477-486
8. Chovil, N.: Discourse-Oriented Facial Displays in Conversation. Research on Language and Social Interaction, Vol. 25. (1991) 163-194
9. Fehr, B. J., & Exline, R. V.: Social visual interaction: A conceptual and literature review. In: A. W. Siegman, & S. Feldstein (eds.): Nonverbal Behavior and Communication. Lawrence Erlbaum Associates, Inc., Hillsdale (1987) 225-326

[1] http://wiki.mindmakers.org/projects:saiba:main

10. Heylen, D., Kopp, S., Marsella, S. et al.: The Next Step Towards a Functional Markup Language. Proceedings of Intelligent Virtual Agents 2008, Springer (2008)
11. Kendon, A.: Conducting interaction: Patterns of behavior in focused encounters. Cambridge University Press, New York (1990)
12. Kopp, S., Krenn, B., Marsella, S. et al.: Towards a Common Framework for Multimodal Generation in ECAs: The Behavior Markup Language. Lecture Notes in Artificial Intelligence, Vol. 4133. Springer (2006)
13. Thorisson, K.: Gandalf: An Embodied Humanoid Capable of Real-Time Multimodal Dialogue with People. 1st International Conference on Autonomous Agents, ACM (1997) 536-537
14. Vilhjalmsson, H., Cantelmo, N., Cassell, J. et al.: The Behavior Markup Language: Recent Developments and Challenges. Proceedings of Intelligent Virtual Agents 2007, Vol. LNAI 4722. Springer (2007) 99-111
15. Vilhjalmsson, H., & Stacy, M.: Social Performance Framework. Workshop on Modular Construction of Human-Like Intelligence at the 20th National AAAI Conference on Artificial Intelligence, AAAI (2005)
16. Vilhjalmsson, H.: Augmenting Online Conversation through Automatice Discourse Tagging. HICSS 6th Annual Minitrack on Persistent Conversation, IEEE (2005)