AAMAS 2009 Workshop

Towards a Standard Markup Language for Embodied Dialogue Acts

PROCEEDINGS

IN CONJUNCTION WITH AAMAS 2009

Budapest, May 12, 2009

Dirk Heylen, Catherine Pelachaud, Roberta Catizone, David Traum (eds.)

Contents

An extension of FML for a lexicon of engaging communicative strategies

Carole Adam
RMIT University, School of CS& IT
Melbourne, VIC 3000, Australia
carole.adam.rmit@gmail.com

Lawrence Cavedon
RMIT University, School of CS& IT
Melbourne, VIC 3000, Australia
lawrence.cavedon@rmit.edu.au

ABSTRACT

We propose an extension of the FML-APML specification to match our needs for specifying the communicative strategies of an engaging interactive toy. This intelligent toy uses various strategies defined in previous work to try to engage the child in a long-term relationship. We have identified several useful tags that are not yet incorporated in FML, in particular, tags to represent actions, emotions of other agents, and mental attitudes. This paper motivates such tags and describes a specification of our toy's communicative strategies using them.

1. INTRODUCTION

SAIBA is a multimodal behavior generation framework for Embodied Conversational Agents aiming at allowing researchers to share modules of their agents architectures. It is composed of the Functional Markup Language (FML [5]), used by an Intent Planner to describe the agent's intentions, goals and plans; and the Behaviour Markup Language (BML), used by a Behaviour Planner to specify the behavioral realization of these intentions.

We are designing an intelligent interactive toy that can engage a child in a long-term relationship through emotional strategies, such as expressing empathy or curiosity, or using coping strategies on behalf of the child [2]. We believe such strategies to be important for building a relationship with the child and therefore crucial to engagement. This intelligent toy is endowed with a logical model of emotions [1] built on Ortony, Clore and Collins' psychological theory of emotions (below referred to as the OCC theory, [9]), Our research focus (and expertise) is about modelling the influence of the child's emotions on the agent's goals. Concerning the emotional expressive capabilities of the Toy (via voice, facial expression, etc.) we hope to reuse existing work. We thus find the SAIBA initiative very interesting since it will allow us to use Behaviour Planners designed by other researchers, who are experts of the field of emotion expression and agent animation. Therefore we need to produce the toy's intention in a compliant format, namely in FML.

Following [6, p.6], our Intent Planner will ground on a lexicon of strategies, similarly to the Gesticon, a lexicon of gestures that can be used by the Behaviour Planner. Concretely, we intend to write a lexicon of templates of strategies, where some variables will have to be instantiated in real-time during the interaction by the Strategy Planner (our realization of SAIBA Intent Planner). The intelligent toy can then choose a strategy in this lexicon and send the intention to perform this strategy to the Behaviour Planner. The strategies available to the toy have been formalised in a BDI logic framework in previous work [2].

We describe here extensions to the current FML specification that we think to be required for representing these communicative engaging strategies. Since FML is still under development, we use the FML-APML specification proposed in [8]. While specifying our strategies in this language, we have identified a number of missing tags that we introduce and discuss here. This proposal of new tags will be illustrated by examples of intentions using them, from our lexicon of strategies.

We start by introducing the engaging interactive toy on which we are working (Section 2), in particular its architecture and the communicative strategies it uses to engage the child. We then discuss the current FML-APML specification and the tags we propose to add to it in order to be able to specify the toy's engaging strategies (Section 3). We then provide our lexicon of templates of strategies along with a discussion of our choices of formalisation (Section 4). Finally we conclude the paper by discussing future works to be done for FML and for our project (Section 5).

2. AN INTERACTIVE ENGAGING TOY

At the previous AAMAS FML workshop [4], four scenarios involving the description of different kinds of functions have been envisaged: task-oriented collaboration and negotiation with a human for the construction of a physical object; long monolog from a museum presentation agent perceiving no feedback from the human audience; multiparty dynamic social conversation in a restaurant; long term companion agent in the health domain. Our intelligent toy project is in line with the fourth scenario: it is a long-term companion agent, even if in our current focus it is not aimed at the health domain but at children.

2.1 Architecture

Our architectural framework is based on the BDI model of agent implementation: incoming utterances are interpreted as communicative intentions which are processed within a dialogue manager based on the *information-state update* ap-

Cite as: EDAML Workshop @ AAMAS'09, Dirk Heylen, *Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, Decker, Sichman, Sierra and Castelfranchi (eds.), May, 10–15, 2009, Budapest, Hungary, pp. XXX-XXX.

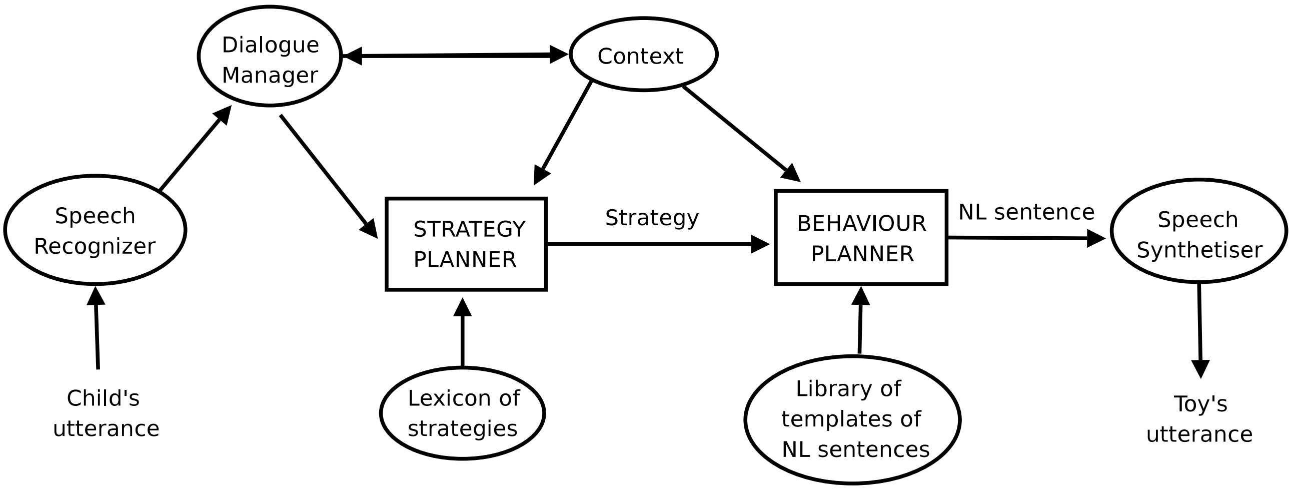

Figure 1: SAIBA-compliant architecture of the Toy

proach [7]. For the purpose of this paper, we focus on the generation part of the architecture.

The SAIBA design requires agents to be made up of an Intent Planner using FML to describe the computed intentions, a Behaviour Planner matching these intentions with a BML description of the behaviour needed to achieve them, and a Behaviour Realizer able to understand the BML description and perform the requested behaviours. We have adapted the architecture of the Toy to comply with this design. The toy is thus made up of:

- a Strategy Planner (matching SAIBA Intent Planner) that users a Strategy Lexicon to choose a strategy template, then uses the context to instantiate this template, and finally sends the instantiated strategy (more precisely the toy's intention to perform this strategy) to the Behaviour Planner;
- a Behaviour Planner: this module is very simple for now, picking up and filling in an appropriate template in a library of precomputed natural language sentence templates[1], and sending the instantiated chosen sentence to the Behaviour Realizer;
- a Behaviour Realizer, concretely for now only a speech synthetiser. In future work this speech synthetiser should be able to interpret BML markup on the received text, and the Toy's architecture will be enriched with more complex Behaviour Realizers (facial expressions, gestures, etc.).

2.2 Mechanism of the strategy planner

In the Toy architecture, the Strategy Planner module continuously analyses the context and generates in real time the adopted strategies. In particular, information about the context allows it to choose the appropriate template of strategy in the lexicon, and then to fill in its slots with the correct values. For example the strategy "express own emotion" must have the field "emotion" filled in with the current emotion of the toy. Another example is the conversation partner's emotion, that is part of the context, and can thus be used to further specify a strategy template. This allows the toy's reaction to be adapted to the current context and to the user.

The generated file can contain several applicable strategies (intents) with their relative priorities. The priority is the value of the *importance* tag, and is determined at runtime by the Strategy Planner from the context[2], to ensure that the best strategy is applied in each situation rather than a random strategy. Then the Behaviour Planner will also use the context information to determine which strategies are realizable, and which behaviour to use to achieve them: for example if the received strategy is to perform a given action, the Behaviour Planner has to determine if this action is possible in the current (physical) context, allowed in the current (normative) context, etc; otherwise it has to try another strategy. However, we are not interested in the Behaviour Planner in this paper.

2.3 Strategies

The following strategies are available to our engaging toy (formalised in previous work [2]):

- **Empathy** consists in expressing the appropriate goodwill *fortunes-of-others* emotion from the OCC theory [9]: *happy-for* when the child feels a positive emotion, and *sorry-for* when he/she feels a negative one;
- **Expression** consists in expressing the toy's own emotion if it is feeling one in relation to the current subject;
- **Curiosity** consists in asking the child about his/her attitude towards the subject at hand, a way to provide the toy with useful information to then compute the child's emotion, and to make it seem interested in the child;
- **Confirmation** consists in asking the child confirmation about the computed emotion, making the toy look aware of the child's emotion;

[1] In later work, we hope to take advantage of the expertise of other researchers by using an existing Behaviour Planner.

[2] The link between the context and the priority of the various strategies is for now empiric. Ideally it should be assessed after interaction experiments with children, but we do not discuss such details here.

- **Positive reinterpretation** is a coping strategy consisting in finding positive aspects in the negative subject being discussed. Here the toy helps the child to cope by communicating these positive aspects to him/her;

- **Active coping** is another coping strategy, consisting in taking active steps against a stressful situation. Here the toy can either correct the child's incorrect beliefs that lead to a negative emotion, perform an action to reduce this emotion, or encourage the child to perform such an action;

- **Mental disengagement** is also a coping strategy, consisting in shifting the child's mind off the problem by engaging in another activity. Here the toy can propose such a relaxing activity to divert the child from the stressor.

3. WHICH FML SPECIFICATION FOR THIS STRATEGY LEXICON?

FML is still under development as a standard, so we ground our strategies in the FML-APML proposal of specification that was the more complete in our view [8]. In particular emotions are represented in EARL [10]. The following subsections list the tags we need to add to the current FML-APML specification to make it expressive enough to specify the communicative strategies of our toy, as listed above (Section 2.3).

3.1 Expressive performatives and emotion tag

We need a specific performative for expressive communicative intentions (intentions to communicate the agent's emotion) corresponding to the expressive category in Searle's typology of speech acts [11]. This would replace the current *emotion* tag used in FML-APML to represent a communicative goal to express an emotion. Instead, we would still have an *emotion* tag, but use it to represent an emotion as an object and not as a communicative goal (see Section 3.2). Thus this emotion-object should for example be encapsulated in a performative tag as its theme or rheme. This expressive performative accounts for the verbal expression of an emotion (for example to say "I am happy").

But we also need to be able to specify emotions that are non-verbally expressed along with a speech act. For this we propose to have an *accompanying-emotion* tag inside each performative, encapsulating an emotion-object describing which emotion should be expressed by the agent while performing the speech act (of course, the communicative act itself may express a different emotion to that being communicated non verbally). This allows us to specify the behaviour of agents who would assert that they feel a certain emotion (for instance because social norms force them to do so) while their gestures or expression show their real (felt) emotion.

3.2 A tag for emotions as objects

Just as the agent must be able to refer to events or objects in the world, it may wish to refer to an emotion, be it its own emotion or another agent's one. We thus need to have an *emotion* tag to represent emotions as objects (and not as intentions). This tag should have the following attributes to fully specify the emotion:

- standard FML-APML attributes: *id*, *start*, *end*;
- EARL attributes: category, intensity, regulation, *etc*;
- agent: the agent feeling this emotion;
- target: the agent at whom this emotion is directed, if any (for example for *anger* or *sorry-for*);
- object: the object of the emotion (for instance *content* about the sunny weather, or *pride* of a good mark at school).

Please note that our new version of the *emotion* tag represents an object and not a communicative intention anymore, so it does not need an *importance* attribute. Actually, the *importance* attribute of FML-APML should be restricted to the performative tag; the *world* tag for example does not require an importance.

This *emotion* tag is useful for example in the *confirmation* strategy, when the toy wants to talk about the child's emotion, or in the *expression* or *empathy* strategies, when it refers to its own emotion.

3.3 Referring to mental attitudes

Agents may not only refer to emotions but also to their and others' mental attitudes. For example in some strategies the toy can query the child about his attitude towards something (desires, likings, etc.), or talk about its own mental attitudes (for instance its beliefs about the situation). Thus we need to be able to refer to agents' mental attitudes with a specific tag. FML-APML had introduced the *world* tag to allow the possibility of refering to the external world. We propose a *mental-attitude* tag which, along with the *emotion* tag proposed in Section 3.2, aims at making it possible to also refer to the internal world.

The *mental-attitude* tag would have the following attributes:

- agent: the agent owning this mental attitude (the one who believes, or desires, *etc.*)
- type: among a list of possible mental attitudes (non exhaustive): belief, desire, intention, choice, goal, ability, ideal, likes, *etc.*;
- object: the object of the mental attitude (the proposition that is believed, or desired, *etc.*).

For example: the child (agent) likes (type) tigers (object).

3.4 Describing actions

The *world* tag introduced in FML-APML has for now the following subtypes: events, places and objects. Now if we consider the OCC theory of emotions [9], emotions can be triggered by three types of stimuli: objects, events and actions. This is an argument in favour of adding an *action* type to the *world* tag. This type of *world* element should have a property *author* to specify the name of the agent performing the action.

Further, some of our strategies consist in talking about actions and their effects, for instance informing that performing such action would make the stressor disappear, and thus suggest to the child that he should perform this action. There are two things that need to be described here:

- *actions* as elements of the world, that would be covered by this new type of *world* tag;
- *relations* between actions and situations in the world (conditions and effects of actions); these are discussed in Section 3.5.

3.5 Relations between elements of the world

FML-APML and various other specifications of FML have a tag to represent meta-discourse level relations between parts of the discourse (tag called *belief-relation* in FML-APML for example). Similarly we wish to represent relations between elements of the world (and not of the discourse). For example:

- causal relations between events (useful in the positive reinterpretation strategy);
- causal relations between actions and situations in the world (useful for the active coping strategy);
- size relations between objects (bigger than...), temporal relations between events (before, after...), or distance relations between places (far, close...), etc;
- finally if we had another type of *world* tag to refer to the participants of the interaction, then this new *relation* tag would also allow us to represent the social relationships between two agents, a notion whose importance has been highlighted in [3].

We thus propose a *world-element-relation* tag with the following attributes:

- type of relation (size, distance, causality, effect, before, after...);
- first element;
- second element.

For example this tag will allow us to represent and converse about the effect of the action of closing the door (that the door is closed), and about less trivial actions and effects (see Section 4.6).

```
<world-element-relation type=”action-effect-law”>
    <element1>
        <world type=”action” name=”close-door”>
    </element1>
    <element2>
        <world type=”event” name=”door-closed”>
    </element2>
</world-element-relation>
```

In the following section we give a specification of our strategies in this extended FML-APML, using in particular the new tags introduced above.

4. FML LEXICON OF STRATEGIES

We currently represent our strategies using rules: for each strategy there are conditions for generation depending on the content of the toy's knowledge base, and an FML template of intent with slots to be filled in depending on the triggering context. The Strategy Planner determines in real time, depending on the context, the values of these slots as well as the priority of each strategy. It then sends the instantiated strategies to the Behaviour Planner.

In this section we thus provide templates of strategies whose slots are marked by upper-case letters, representing the variables to be instantiated at runtime. As a simplification, in all the following strategy templates we do not specify the *start*, *end* and *importance* tags. We also illustrate each strategy using a natural language sample sentence that the Behaviour Planner could use to achieve it.

4.1 Empathy

4.1.1 *Condition*

The *empathy* strategy is applicable whenever the context indicates that the user is feeling an emotion with enough intensity. Positive empathy (*happy-for*) is triggered by positive child's emotions while negative empathy (*sorry-for*) is triggered by negative ones.

4.1.2 *FML specification*

```
<FML>
    <performative id=”p1” type=”expressive”>
        <theme>
            <emotion-object id=”e1” type=T
                                regulation=”felt”/>
        </theme>
        <accompanying-emotion>
            <emotion-object id=”e1”/>
        </accompanying-emotion>
    </performative>
</FML>
```

4.1.3 *Example*

If the child is sad because he has just broken his bicycle, the toy could say “I am sorry about your broken bicycle”.

4.1.4 *Discussion*

The *emotion* tag that is encapsulated in the expressive *performative* tag as its theme represents an object emotion as discussed above, as opposed to a communicative intention. Its type T will be instantiated by either ”happy-for” or ”sorry-for” depending on the context, *i.e.* concretely on the (positive or negative) valence of the child's emotion. This will inform the Behaviour Planner that the agent intends to verbally express this emotion. Further, the *accompanying-emotion* tag specifies which emotion should be expressed non-verbally along with the resulting speech act (in this case the same emotion that is verbally expressed)[3].

4.2 Expression

4.2.1 *Condition*

If the toy feels an emotion (due to its own personality and profile) relevant to the subject of the dialogue, then it can express it. This may be less prioritary than expressing empathy to the kid, depending on the relative intensities of the toy's and the kid's emotions; however we are not interested in the management of priorities here.

[3] Since this same emotion appears twice, we would probably need a mechanism to refer to a full tag already defined in the file. Here we just give the second one the same identifier and leave the other attributes empty to mean that it is actually a copy of the first one already defined.

4.2.2 FML specification

```
<FML>
    <performative id=”p1” type=”expressive”>
        <theme>
            <emotion id=”e1” type=E regulation=”felt”/>
        </theme>
        <accompanying-emotion>
            <emotion id=”e1”/>
        </accompanying-emotion>
    </performative>
</FML>
```

4.2.3 Example

If the child is talking about his visit to the zoo and says that he saw koalas, the toy (assuming that its personality is in particular that it loves animals) may answer: “I love koalas”.

4.2.4 Discussion

This specification makes the toy verbally (with an expressive *performative*) and non-verbally (with the *accompanying-emotion* tag) express its emotion. The variable E in this strategy template would be replaced at runtime with the toy’s current emotion type.

4.3 Curiosity

4.3.1 Condition

When the toy cannot compute the child’s emotion about something because it is missing information about the child’s attitude towards this situation, it can ask the child for this information.

4.3.2 FML specification

```
<FML>
    <performative id=”p1” type=”ask”>
        <theme>
            <mental-attitude agent=”child” type=T
                                    object=S />
        </theme>
        <accompanying-emotion>
            <emotion category=”interest” regulation=”felt”/>
        <accompanying-emotion>
    </performative>
</FML>
```

4.3.3 Example

The child is still talking about his visit to the zoo, and this time informs the toy that he saw tigers. Now the toy does not know the child’s preferences about tigers so it cannot deduce the resulting child’s emotion. So it may ask: “Do you like tigers?”.

4.3.4 Discussion

Since we must not put any text in the FML specification, we cannot write the theme as follows:

$$< theme > Do\ you\ like\ S\ ? < /theme >$$

Thus the theme is rather a mental attitude of the child, described by the *mental-attitude* tag we propose to add. The type T and object S of this mental attitude have to be instantiated at run-time: the variable S will be replaced by the subject at hand; the variable T will be replaced either by *desire*, *ideal* or *like* depending on the particular type of S (event, action or object respectively). Indeed according to the OCC theory events are appraised wrt one’s desires, actions are appraised wrt one’s standards (what we call ideals), and objects are appraised wrt one’s likings. This will result in the agent asking questions such as “Do you think it is ideal to have the best score in mathematics?”, “Do you want to go to the zoo?”. It could also express the question in a negative way, for instance “Do you hate tigers?”, but this will only be decided by the Behaviour Planner (both the positive and negative questions achieve the same intention to ask the child about his attitude towards tigers).

4.4 Confirmation

4.4.1 Condition

Whenever the toy is able to compute the child’s emotion about the subject being discussed, it can choose to ask the child to confirm it. This strategy may have higher priority when the toy is not sure about its deduction.

4.4.2 FML specification

```
<FML>
    <performative id=”p1” type=”ask”>
        <theme>
            <emotion agent=”child” category=C
                                    object=S target=T />
        </theme>
        <accompanying-emotion>
            <emotion category=”interest” regulation=”felt”/>
        <accompanying-emotion>
    </performative>
</FML>
```

4.4.3 Example

The child now announces that he saw a lion at the zoo, and this lion was roaring at him. Thanks to its profile of the child, the toy can deduce that the child may have been afraid that the lion was going to attack him (believing that roaring was a signal of that). The toy may choose to ask a confirmation of this deduction: “Were you afraid that the lion would attack you?”. The child’s answer will allow it to update the profile.

4.4.4 Discussion

We see with this strategy the interest of having a tag to describe emotions (including emotions of other agents) as objects with all their characteristics (in particular here their object). The variables representing these characteristics (namely C, S and T) will be instantiated in real time depending on the context.

We chose to specify an accompanying emotion of interest, since it would contribute to making the toy seem interested in the child, and should thus help in engaging the child in the interaction.

4.5 Positive reinterpretation

4.5.1 Condition

This strategy is applicable if the child is feeling a negative emotion about a situation that the toy believes to have positive aspects.

4.5.2 FML specification

```
<FML>
    <performative id=”p1” type=”inform”>
        <theme>
            <world-element-relation type=”causality”
                                element1=E1 element2=E2/>
        </theme>
        <accompanying-emotion>
            <emotion category=”relief” regulation=”felt”/>
        <accompanying-emotion>
    </performative>
</FML>
```

4.5.3 Example

The child is at home and cannot go to the zoo because it is snowing. Seeing that the child is sad about this, and knowing that he likes making snowmen, the toy can say: “With all this snow we can make a snowman in the garden”.

4.5.4 Discussion

We see here the usefulness of the tag for expressing causal relations between events in the world. The variable $E1$ will be instantiated with the object of the child’s negative emotion, and $E2$ with the positive consequence found by the toy. The accompanying emotion should be a positive emotion, such as relief or content.

4.6 Active coping: child’s action

4.6.1 Condition

This strategy is applicable when the child is feeling a negative emotion and can perform an action against the stressor (the object of his emotion).

4.6.2 FML specification

```
<FML>
    <performative id=”p1” type=”incite”>
        <theme>
            <world ref-type=”action” ref-id=A
                    prop-type=”author” prop-value=”child” />
        </theme>
        <accompanying-emotion>
            <emotion category=”trust” regulation=”felt”/>
        <accompanying-emotion>
    </performative>
    <performative id=”p2” type=”inform”>
        <theme>
            <world-element-relation type=”action-effect-law”
                    element1=A element2=C />
        </theme>
        <accompanying-emotion>
            <emotion category=”content” regulation=”felt”/>
        <accompanying-emotion>
    </performative>
</FML>
```

4.6.3 Example

If the child is anxious because he has a mathematics exam the next day and is not very good at mathematics, the toy can suggest to him that he should review his lesson: “You should review your math lesson again”, in the hope that being reviewing the lesson would increase the child’s chance to get a good mark. Besides, being well prepared should at least reduce the child’s anxiety about the exam.

4.6.4 Discussion

This strategy contains two performatives, one to announce that an action has a positive effect on the situation, the other to incite the child to perform this action. The agent can actually perform both communicative acts, or one or the other. The accompanying emotion for the *incite* performative has been specified among the EARL emotion categories as *trust*.

This strategy shows the utility of the *action* type in the *world* tag to describe actions, along with the *world-element-relation* tag to describe their effects. The Strategy Planner will instantiate variable A by an action that the child can perform that will eliminate the stressor; it will instantiate the variable C with the positive consequence of this action.

4.7 Active coping: toy’s action

4.7.1 Condition

This strategy is applicable if the child is feeling a negative emotion and the toy can perform an action against its cause.

4.7.2 FML specification

```
<FML>
    <performative id=”p1” type=”propose”>
        <theme>
            <world ref-type=”action” ref-id=A
                    prop-type=”author” prop-value=”toy” />
        </theme>
        <accompanying-emotion>
            <emotion category=”courage” regulation=”felt”/>
        <accompanying-emotion>
    </performative>

    <performative id=”p2” type=”inform”>
        <theme>
            <world-element-relation type=”action-effect-law”
                    element1=A element2=C />
        </theme>
        <accompanying-emotion>
            <emotion category=”content” regulation=”felt”/>
        <accompanying-emotion>
    </performative>
</FML>
```

4.7.3 Example

If the child is disappointed because his mother has no time to tell him a bedtime story, the toy can propose to tell him one: “I can tell you a story if you want”.

4.7.4 Discussion

This strategy is similar to the previous one except that the action is to be performed by the toy, so the performative $p1$ is of type *propose* instead of *incite*. The accompanying emotion could be *courage* or *hope* to show that the toy is confident in the outcome of its action.

4.8 Mental disengagement

4.8.1 Condition

This strategy is applicable when the child is feeling a negative emotion and no active coping strategies apply, but there exists some diverting activity that the participants (the toy and the child) can engage in.

4.8.2 FML specification

```
<FML>
    <performative id=”p1” type=”propose”>
        <theme>
            <world ref-type=”action” ref-id=A
                        prop-type=”author” prop-value=T />
        </theme>
        <accompanying-emotion>
            <emotion category=”excitement” regulation=”felt”/>
        <accompanying-emotion>
    </performative>
</FML>
```

4.8.3 Example

If the child is angry at his brother who has borrowed his bike while he wanted to go for a ride, the toy can try to engage him in another activity to make him forget about his anger: “Let’s play chess together!”.

4.8.4 Discussion

In this case the action may be an action by either participant, or a joint action by both participants (for example, play a game together). The author T and the action A will be specified at runtime. The accompanying emotion of *excitement* has been chosen to motivate the child to engage in the diverting activity.

5. CONCLUSION

In this paper we have provided an FML specification of some communicative strategies used by an interactive toy to engage the child in a long-term relationship. While specifying these strategies we have identified some tags that were required for expressing the involved intentions, but that were currently missing. We have thus proposed an extension of the current FML-APML specification with some new tags.

The interest of SAIBA-compliance is to allow researchers to re-use modules developped by other researchers, specialised in different areas (emotion psychology for the reasoning of the agent, expression of emotions for the translation from communicative intents to BML units, animation for the final rendering of emotions, etc.). We thus hope that our Strategy Planner may be useful in other agents, and we are looking forward to the further developments of SAIBA-compliant reusable modules that could be integrated in the architecture of our interactive toy.

6. ACKNOWLEDGEMENTS

This work is partially supported by the Australian Research Grant under grant number LP0882013. We also wish to thank Real Thing Pty Ltd. for their support.

7. REFERENCES

[1] C. Adam. *Emotions: from psychological theories to logical formalization and implementation in a BDI agent.* PhD thesis, INP Toulouse, France, July 2007.

[2] C. Adam and P. Ye. Reasoning about emotions in an engaging interactive toy. In *Empathic Agents Workshop @ AAMAS’09*, 2009.

[3] T. Bickmore. Framing and interpersonal stance in relational agents. In *Why conversational agents do what they do. Functional representations for generating conversational agent behaviour, workshop at AAMAS’08*, 2008.

[4] FML. Fml: scenarios for next steps. `http://wiki.mindmakers.org/_media/projects:fml:aamas08-scenarios.pdf?id=projects\%3Afml\%3Amain&cache=cache`, 2008.

[5] D. Heylen, S. Kopp, S. Marsella, C. Pelachaud, and H. Vilhjalmsson. The next step towards a functional markup language. In *AAMAS’08*, 2008.

[6] S. Kopp, B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. R. Thorisson, and H. Vilhjamsson. Towards a common framework for multimodal generation: the behaviour markup language. In *IVA’06*, 2006.

[7] S. Larsson and D. Traum. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering*, 6:323–340, 2000.

[8] M. Mancini and C. Pelachaud. The FML-APML language. In *Why conversational agents do what they do. Functional representations for generating conversational agent behaviour, workshop at AAMAS’08*, 2008.

[9] A. Ortony, G. Clore, and A. Collins. *The cognitive structure of emotions.* Cambridge University Press, Cambridge, MA, 1988.

[10] M. Schroder, H. Pirker, and M. Lamolle. first suggestions for an emotion annotation and representation language. In *LREC’06*, 2006.

[11] J. R. Searle. *Speech acts: An essay in the philosophy of language.* Cambridge University Press, New York, 1969.

Reactive behaviors in SAIBA architecture

Elisabetta Bevacqua
Telecom ParisTech
46 rue Barrault, 75013
Paris, France
bevacqua@telecom-paristech.fr

Ken Prepin
Telecom ParisTech
46 rue Barrault, 75013
Paris, France
prepin@telecom-paristech.fr

Etienne de Sevin
Telecom ParisTech
46 rue Barrault, 75013
Paris, France
de-sevin@telecom-paristech.fr

Radosław Niewiadomski
TELECOM ParisTech
46 rue Barrault, 75013
Paris, France
niewiado@telecom-paristech.fr

Catherine Pelachaud
Telecom ParisTech
46 rue Barrault, 75013
Paris, France
pelachaud@telecom-paristech.fr

ABSTRACT

In this paper we propose an extension of the current SAIBA architecture. The new parts of the architecture should manage the generation of Embodied Conversational Agents' reactive behaviors during an interaction with users both while speaking and listening.

General Terms

1. INTRODUCTION

SAIBA [13] is an international research initiative whose main aim is to define a standard framework for the generation of virtual agent behavior. It defines a number of levels of abstraction (see Figure 1), from the computation of the agent's communicative intention, to behavior planning and realization.

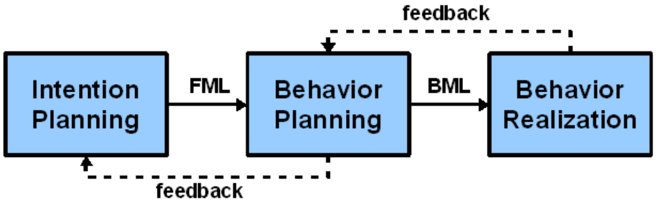

Figure 1: Saiba

The Intent Planning module decides the agent's current goals, emotional state and beliefs, and encodes them into the Function Markup Language (FML) [3] (this language is still being defined). To convey the agent's communicative intentions, the Behavior Planning module schedules a number of communicative signals (e.g., speech, facial expressions, gestures) which are encoded with the Behavior Markup Language (BML). It specifies the verbal and nonverbal behaviors of ECAs [13]. Each BML top-level tag corresponds to a behavior the agent is to produce on a given modality: head, torso, face, gaze, body, legs, gesture, speech, lips.

In a previous work we proposed a first approach to the FML: the *FML-APML* language [7]. FML-APML is an XML-based markup language for representing the agent's communicative intention and the text to be uttered by the agent. The communicative intentions of the agent correspond to what the agent aims to communicate to the user: its emotional states, beliefs and goals. It originates from the APML language [1] which uses Isabella Poggi's theory of communicative acts. It has a flat structure, and allows defining explicit duration for each communicative intention. Each tag represents one communicative intention; different communicative intentions can overlap in time.

However, we believe that FML alone cannot encompass all the behaviors that people perform during an interaction. Some of them do not derive uniquely from a communicative intention, they appear rapidly as a dynamic reaction to external or internal events. For example, a person engaged with friends in conversation will respond to their laugth or she could react to an unexpected shift of the other party's gaze and look (unconsciously) in the same direction.

We think that, to perform these behaviors type, the ECAs must be able, when a new event occurs (expected or not), to compute immediate reaction (Reactive Behavior module), to select between this reaction and the previously planned behavior (action-selection module), and if necessary, to replan behavior dynamically (FML chunked representation). In the next Section, we propose an extension of the current SAIBA architecture that should manage these tasks. Then we will explain how this architecture allows us to generate both speaker's and listener's behaviors. In Section 2 we present some scenarios/applications that can be realized with the exptended architecture. In Section 4.1 we present how adaptation (viewed here as reaction) between interactants is possible in the new SAIBA architecture. In Section 4.2 we will argue the importance of an *Action Selection* module that selects the appropriate behavior the agent should display. We also suggest in Section 4.3 that, for real-time purpose, FML input files should not be sent as a whole but in chunks. Finally, we will describe some examples that make

Cite as: Reactive behaviors in SAIBA architecture, E. Bevacqua, K. Prepin, E. de Sevin, R. Niewiadomski and C. Pelachaud, *Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, Decker, Sichman, Sierra and Castelfranchi (eds.), May, 10–15, 2009, Budapest, Hungary, pp. XXX-XXX.

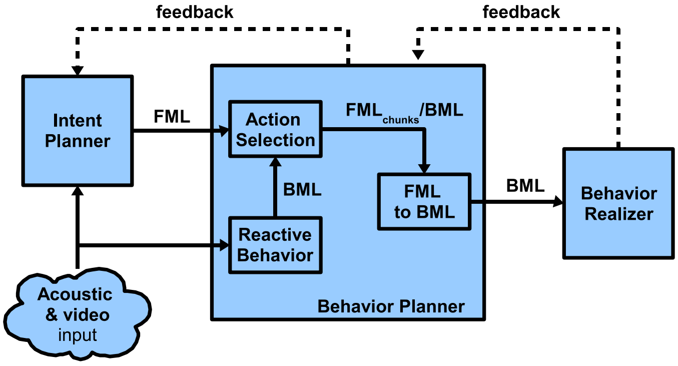

Figure 2: Proposed extension of the current SAIBA architecture. Three elements are added to the Behavior Planner module: Reactive Behavior (see Section 4.1), Action Selection (see Section 4.2)and the FMLchunk (linking 'Action Selection' to 'FML to BML' elements, see Section 4.3).

use of this architecture.

2. REALTIME INTERACTIONS

The SAIBA architecture is primarily dedicated to verbal-driven interactions. In these interactions, the speech is used to transmit sense and meaning to a partner. FML represents these meanings by mean of communicative intentions. However speech alone is not enough to enable realtime verbal communication to take place between speaker and listener. The speaker needs feedbacks from the listener, and the listener needs that the speaker adapts its speech depending on these feedbacks. There is a need of realtime adaptation of the agents to both, the context and the reaction of their partner.

To be involved in natural verbal interactions with humans, we believe that the Behavior Planner module needs to be modified. This module should be able to receive visual and acoustic input signals (described with BML tags) and to influence agent's actions in a very reactive way. We have added three elements to the Behavior Planner module (see Figure 2): one element to compute reactive response (Reactive Behavior), another to select between this 'on the fly' reaction and the preplanned behavior (Action Selection), and finally one element offering the capacity to replan the behavior whenever necessary (thanks to a Chunked FML representation).

This new version of the Behavior Planner module will be both influenced by the higher level communicative intentions conveyed by FML and be reactive to physical events.

3. REALTIME APPLICATIONS

The proposed architecture in Figure 2 can be easily applied to generate the agent's behavior both while speaking and listening. In both roles the agent can perform behaviors derived from its communicative intentions and reactive responses triggered by external and internal events. We suppose that, while speaking, the system will go mainly through the Intent Planner module to execute all the cognitive processes needed for dialogue generation. However, even while speaking, the agent could perform some reactive behaviors, like smiling back to the listener's smile. On the other hand, while in the role of the listener, the agent's behavior could be mainly reactive, since previous research has shown that the listener's behaviour is often triggered by the verbal and nonverbal signals performed by the speaker [6, 14]. However, even while listening, the agent can intentionally display some signals to show the other party what it thinks about the speech, for example that it agrees or not, believes or not and so on. In conclusion, in both interactive roles, the ECA system must be able to generate cognitive and reactive behaviors.

In particular, when going through the cognitive process, some information in the FML can help the system to generate the right behavior according to the current role of the agent. In fact, that during a human-human communication, participants know exactly where they stand into the interaction. They know when they are speaking or listening, if they aim to give the turn to elicit an answer from the other party. They recognize when they can take the turn or when they have to insist to obtain it. Such a knowledge drives the interlocutors' behavior. For example, if a participant wants to communicate his agreement towards the content of the speech, he will just nod the head if he is listening otherwise he will express his agreement with a full sentence if he is speaking. To fit well in an interaction with users, a conversational agent should know which is its role at any moment of the communication in order to show the right behavior. That is why the FML should contains tags for the turn management. This type of tag would not only influence the choice of the appropriate behavior to convey a certain communicative intention, like in the example described above, but also determine the generation of particular behavioral signals. For example, if the agent wants to take the turn, it can open its mouth and emit short sounds to make the user let him the floor.

3.1 Mimicry

Several researches have shown that in human-human interactions people tend to imitate each other. This copying behaviour, called *mimicry*, has been proven to play an important role during conversations. For example, when fully engaged in an interaction, mimicry of behaviors between interactants may happen [5]. Mimicry behavior can be performed consciously or unconsciously. During the interaction a person could decide to imitate the other party's smile in order to show that he shares his appreciation. To generate this type of behavior, the architecture proposed in Figure 2 would generate a FML containing the communicative intention of mimicry; afterwards the Behavior Planner would translate it in behavioral signals according to the behavior performed by the other party, in this example, the chosen signal would be a smile.

On the other hand one could be completely unaware of mimicking the person he is interacting with. Such a behavior, called by Lakin "chameleon effect" [4], helps to create affiliation and rapport. To generate this type of reactive and unconscious behavior, we propose that the Behavior Planner should include a sub module, the Reactive Behavior in Figure 2. Such a module, triggered by the user's acoustic and non verbal signals, generates the mimicry behavior in BML format. No need for FML in this situation since the agent's behavior is unintentional and since, being a reactive behavior, its generation should be as faster as possible.

3.2 Empathy

Empathy is commonly defined as the capacity to "put your-self in someone else's shoes to understand her emotions" [11]. To be empathic assumes one is able to evaluate the emotional dimension of a situation from the point of view of another person.

Magalie Ochs et al. [10] have proposed a model of empathic emotions elicitation during a dialog. From the subjective evaluation of the interlocutor's speech, the Intent Planner generates the FML representing the empathic responses to be displayed by the agent. These empathic responses can be simple as well as complex expressions (e.g. superposition of empathic and egocentric emotions) [9]. This FML is sent to the Behavior Planner which translates it in behavioral signals.

The empathic expressions should be distinguished from the mimicry of emotional expressions [2, 12]. While the first may result in various emotional responses, the second consists in unconscious imitation of the facial expressions of the interlocutor. According to Dimberg et al. [2] these facial expressions are difficult to inhibit voluntary. This type of emotional expressions can not be generated by the Intent Planner. They ought to be specified more reactively. We believe these mimicry of emotional expressions have to be computed directly by the Reactive Behavior process.

4. MODIFICATION

In the next subsections we present the modifications we have brought to the SAIBA platform.

4.1 Reactive Behavior

The mutual adaptation necessary to enable verbal interaction between an ECA and a human is, in some way, highly cognitive: the speaker can have to re-plan its speech, the emotions of the agents can change throughout the dialogue. However this mutual adaptation is also, in some other way, mostly reactive, just as a dynamical coupling with the partner: the listener will give backchannels, the partners may imitate each other, they may synchronise, or slow down or speed up their rhythms of production.

This dynamical aspect of the interaction is much closer to the low-level of the agent system than to the high-level of the communicative intentions described by FML: this dynamical coupling needs reactivity (realtime perception) and sensitivity (realtime adapted actions). For this reason, the ReactiveBehavior module has a certain autonomy from the rest of the architecture. It will short-cut the Intent Planner, getting directly input signals, i.e. the BML coming from the human (see Figure 2), as well as the currently planned actions, i.e. the BML produced at the output of the BehaviorPlanner.

With these two sources of information, the Reactive Behavior module will propose to the Action Selection module (see Section 4.2) two different types of data. It can propose adaptation of the current behavior. By comparing its own actions to the actions of the speaker at a very low level, among other thing tempo or rhythm of signal production; for example it can propose to slow down or speed up behaviors. This type of propositions may enable synchronisation, or similarity of tempo with the user. The second type of data proposed by the Reactive Behavior are full actions. By extracting from the user's behavior salient events, it will propose actions such as performing a backchannel, imitating the user or following its gaze.

Finally the Reactive Behavior will be able to propose realtime reactions or adaptations to the user's behavior thanks to its partial autonomy. It will act more as an adaptator of the ongoing interaction than as a planner. It is a complementary part of the Intent Planner, much more reactive and also working at a much lower level. The ECA must be able to select or to merge the information coming from both this Reactive Behavior and the Intent Planner, using for instance an Action Selection module.

4.2 Action Selection

The Action Selection receives propositions of actions from the intention planner in FML and the Reactive Behavior module in BML (see Figure 2) and send the chosen action (in FML or BML) to the FMLtoBML module. The Action Selection allows the agent to adapt interactively to the user's behaviors by choosing between actions coming from the Reactive Behavior module and from the intention planner. That is the Action Section module chooses between a more cognitive-driven or a more reactive-driven behavior.

More precisely, the intent planner module and the Reactive Behavior module can propose conflicting actions. The action selection module has to decide which action is the most appropriate. This selection is made by considering the user's interest level as well as the intentions and emotional states of the ECA. To enable the Action Selection module to make a choice, actions are associated to priorities. These priorities are computed depending on the importance the ECA gives to communicate a given intent. Importance of a communicative intent is represented by the importance tag of APML-FML [8].

4.3 FML chunk

To interact with users, the ECA system must generate the agent's behavior in real-time. Computing the agent's animation from a large FML input file, ie that contains several communicative intentions, could create an unacceptable delay that would slow down the agent's response, making the whole interaction unnatural. That is why we think that FMLs should be cut in smaller chunks when needed.

Therefore, we suggest that the FML language should contain additional information to specify if a FML command belongs to a larger FML, which is its order in the subset and how long is the original FML. Knowing the duration of the original FML would help the process of behavior planning. For example, a non verbal signal, bound to a minimum duration time, could start in a FML chunk if the original FML is long enough to allow its whole animation.

The decomposition of FML in a subset of chunks asks for the implementation of a feedback system between the modules of the SAIBA architecture. In order to plan or re-plan the agent's intentions, the Intention Planner module needs to be informed about the current state of the FML that it has generated. Possible states of a FML are: "playing", "completely played", "discarded", "interrupted".

5. CONCLUSIONS

In this paper we discussed how some aspects of interactions can be managed within SAIBA. In our opinion reactive behaviors during an interaction cannot be managed properly in the current architecture. Thus we proposed its extension as well as some examples of scenarios/applications of it. The new nodules of the architecture allows Embodied Conversational Agents for reactive behaviors during an interaction with users both while speaking and listening.

6. ACKNOWLEDGMENTS

This work has been funded by the STREP SEMAINE project IST-211486 (http://www.semaine-project.eu), the IP-CALLAS project IST-034800 (http://www.callas-newmedia.eu) and the project MyBlog-3D, contract ANR/06 TLOG 023

7. ADDITIONAL AUTHORS

8. REFERENCES

[1] B. D. Carolis, C. Pelachaud, I. Poggi, and M. Steedman. APML, a mark-up language for believable behavior generation. In H. Prendinger and M. Ishizuka, editors, *Lifelike Characters. Tools, Affective Functions and Applications*. Springer, 2004.

[2] U. Dimberg, M. Thunberg, and S. Grunedal. Facial reactions to emotional stimuli: Automatically controlled emotional responses. *Cognition & Emotion*, 16(4):449–471, 2002.

[3] D. Heylen, S. Kopp, S. Marsella, C. Pelachaud, and H. Vilhjalmsson. Why conversational agents do what they do? Functional representations for generating conversational agent behavior. the first Functional Markup Language workshop, 2008. The Seventh International Conference on Autonomous Agents and Multiagent Systems Estoril, Portugal.

[4] J. L. Lakin and T. L. Chartrand. Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychological Science*, 14:334–339(6), July 2003.

[5] J. L. Lakin, V. A. Jefferis, C. M. Cheng, and T. L. Chartrand. Chameleon effect as social glue: Evidence for the evolutionary significance of nonconsious mimicry. *Nonverbal Behavior*, 27(3):145–162, 2003.

[6] R. M. Maatman, J. Gratch, and S. Marsella. Natural behavior of a listening agent. In T. Panayiotopoulos, J. Gratch, R. Aylett, D. Ballin, P. Olivier, and T. Rist, editors, *Proceedings of 5th International Working Conference on Intelligent Virtual Agents*, volume 3661 of *Lecture Notes in Computer Science*, pages 25–36, Kos, Greece, 2005. Springer.

[7] M. Mancini and C. Pelachaud. Distinctiveness in multimodal behaviors. In L. Padgham, D. C. Parkes, J. Müller, and S. Parsons, editors, *Proceedings of Conference on Autonomous Agents and Multi-Agent Systems (AAMAS08)*, 2008.

[8] M. Mancini and C. Pelachaud. Distinctiveness in multimodal behaviors. In L. Padgham, D. C. Parkes, J. Müller, and S. Parsons, editors, *Proceedings of Conference on Autonomous Agents and Multi-Agent Systems (AAMAS08)*, 2008.

[9] R. Niewiadomski, M. Ochs, and C. Pelachaud. Expressions of empathy in ecas. In H. Prendinger, J. C. Lester, and M. Ishizuka, editors, *in Proceedings of 8th International Conference on Intelligent Virtual Agents, IVA 2008*, volume 5208 of *Lecture Notes in Computer Science*, pages 37–44, Tokyo, Japan, 2008. Springer.

[10] M. Ochs, C. Pelachaud, and D. Sadek. An empathic virtual dialog agent to improve human-machine interaction. In *Autonomous Agent and Multi-Agent Systems (AAMAS)*, 2008.

[11] E. Pacherie. *L'empathie*, chapter L'empathie et ses degrés, pages 149–181. Odile Jacob, 2004.

[12] W. Sato and S. Yoshikawa. Spontaneous facial mimicry in response to dynamic facial expressions. In *Proceedings of the 4th International Conference on Development and Learning*, pages 13–18, 2005.

[13] H. H. Vilhjálmsson, N. Cantelmo, J. Cassell, N. E. Chafai, M. Kipp, S. Kopp, M. Mancini, S. Marsella, A. N. Marshall, C. Pelachaud, Z. Ruttkay, K. R. Thórisson, H. van Welbergen, and R. J. van der Werf. The Behavior Markup Language: Recent developments and challenges. In C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, and D. Pelé, editors, *Proceedings of 7th International Conference on Intelligent Virtual Agents*, volume 4722 of *Lecture Notes in Computer Science*, pages 99–111, Paris, France, 2007. Springer.

[14] N. Ward and W. Tsukahara. Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, 23:1177–1207, 2000.

The DIT++ taxonomy for functional dialogue markup

Harry Bunt
TiCC, Tilburg Center for Creative Computing
Tilburg University, Tilburg, the Netherlands
harry.bunt@uvt.nl

ABSTRACT

This paper presents the DIT^{++} taxonomy of communicative functions, with some of its background and theoretical motivations. Its applications in dialogue annotation, dialogue generation, dialogue management, and theoretical dialogue research are summarized, and its role is indicated in a recently started effort of the ISO organization to develop an international standard for functional dialogue markup.

1. DIALOGUE ACTS

Dialogue acts are widely used in studies of dialogue phenomena, in dialogue annotation efforts, and in the design of dialogue systems. The very idea of describing dialogue in terms of communicative actions, such as questions, promises, requests, and greetings, goes back to speech act theory (Austin, 1962; Searle, 1969), which has been an important source of inspiration for modern dialogue act theory. Where speech act theory is primarily a theoretical orientation in the philosophy of language, however, dialogue act theory is a data-driven approach to the computational modeling of interactive language use. As a way to describe meaning in communicative behaviour, dialogue acts are semantic concepts that can be defined by the way a dialogue act is intended to affect the information state of an addressee when (s)he understands the behaviour. For instance, when an addressee understands the utterance *Do you know what time it is?* as a question about the time, then the addressee's information state is updated to contain (among other things) the information that the speaker does not know what time it is and would like to know that. If, by contrast, an addressee understands that the speaker used the utterance to reproach the addressee for being late, then the addressee's information state is updated to include the information that the speaker does know what time it is. Distinctions such as that between a question and a reproach refer to the *communicative function* of a dialogue act; the entities, their properties and relations that are referred to, constitute its *semantic content*. The communicative function of a dialogue act expresses what the speaker is trying to achieve, and the semantic content describes the information that is being addressed. Another way of characterizing this distinction is that the communicative function of a dialogue act specifies how the semantic content is to be used to update an information state.

The term 'dialogue act' is often used in the rather loose sense of 'speech act used in dialogue', but such a characterization hardly does justice to the semantic status of dialogue acts. A more accurate characterization could run as follows. *A dialogue act is a unit in the semantic description of communicative behaviour in dialogue, specifying how the behaviour is intended to change the information state of a dialogue participant who understands the behaviour correctly* (i.e. as intended by the speaker). The semantic content of a dialogue act is the information with which the information state is to be updated; the communicative function specifies the way in which that information is to be used in updating the information state. Formally, a dialogue act is an information-state update operator construed by applying a communicative function to a semantic content.

The assignment of meaning to stretches of communicative behaviour in dialogue presupposes a way to identify stretches that are meaningful. The identification of such stretches is called the segmentation of the dialogue. Dialogue segmentation has many intricacies, the discussion of which is beyond the scope of the present paper; see e.g. Larsson (1998); Geertzen et al. (2007). In this paper we will use the theory-neutral term 'markable' to indicate stretches of communicative behaviour that express one or more dialogue acts, and that are the object of dialogue act markup.

Dynamic Interpretation Theory (DIT) is a computational approach to the analysis of the meaning of dialogue utterances in natural human dialogue or in human-computer interaction, with a focus on the functional aspect of utterance meaning.[1] Like Speech Act Theory, Communicative Activity Theory (Allwood, 2000), and Grice's theory of cooperative action, DIT approaches the use of language as action, but different from these and other theories, DIT considers utterances as expressing multiple update actions on an addressee's as well as on the speaker's information state.DIT does not consider purely linguistic utterances only, but also nonverbal communicative behaviour, such as head gestures and facial expressions, or graphical acts like showing an hour glass on a computer screen, and 'multimodal utterances' where language is combined with nonlinguistic sounds, where prosodic aspects of speech are taken into account, and where linguistic and nonverbal elements are used in synchrony in order to perform one or more dialogue acts.

One of the outcomes of the development of DIT has been a taxonomy of communicative functions which in recent years has been extended and modified, taking into account a range of other taxonomies that have been proposed, and resulting in a comprehensive general-purpose taxonomy called the DIT^{++} taxonomy. This taxonomy has been applied in human annotation, in machine anno-

Cite as: Title, Author(s), *Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, Decker, Sichman, Sierra and Castelfranchi (eds.), May, 10–15, 2009, Budapest, Hungary, pp. XXX-XXX.

[1] See Bunt, 1989; 1990; 1994; 1995; 2000; 2004; 2006; 2008; Bunt et al., 2007; Bunt & Girard, 2005; Geertzen, 2009; Geertzen & Bunt, 2006; Keizer & Bunt, 2006; 2007; Morante, 2007; Petukohva & Bunt, 2009

tation, and in the design and implementation of modules for dialogue management and generation in multimodal dialogue (the PARADIME system, see Keizer & Bunt, 2006; 2007). In the European project LIRICS, a slightly simplified version of this taxonomy has been defined and tested for its usability to the annotation of dialogues in several European languages.

This paper is organized as follows. In the next section we discuss the definition of communicative functions and the multifunctionality of dialogue utterances. Section 3 discusses the structural organization of dialogue act taxonomies, in particular the clustering of communicative functions into dimensions. A notion of dimension is introduced which has a conceptual significance, and goes beyond that of a cluster of mutually exclusive tags. Section 4 describes the structure of the DIT^{++} taxonomy, with its general-purpose and dimension-specific functions. Section 5 briefly discusses some of the applications of the DIT^{++} taxonomy, including its role in a recently started ISO project that aims at establishing an international standard for functional dialogue markup.

2. COMMUNICATIVE FUNCTIONS

2.1 Defining Communicative Functions

Existing markup schemes for communicative functions use either one or both of the following two approaches to defining communicative functions: (1) in terms of the intended effects on addressees; (2) in terms of properties of the signals that are used. For example, questions, invitations, confirmations, and promises are nearly always defined in terms of speaker intentions, while repetitions, hesitations, and dialogue openings and closing are typically defined by their form. Defining a communicative function by its linguistic form has the advantage that their recognition is relatively straightforward, but faces the fundamental problem that the same linguistic form can often by used to express different communicative functions. For example, the utterance *Why don't you start?* has the form of a question, and can be intended as such, but can also be used to invite someone to start. Form-based approaches of dialogue acts are also in danger of confusing purely descriptive concepts with semantic ones, since descriptions like 'repetition' and 'hesitation' say something about the form of the behaviour, but don't say anything about the meaning of that behaviour.

DIT takes a strictly 'deep', semantic approach to dialogue acts in terms of the effects on addressees that a speaker intends to occur as the reflection of understanding the speaker's behaviour. Two caveats, though. First, speakers are not conscious of all their intentions when they perform dialogue acts. Definitions in terms of intended effects should therefore not be taken to imply that dialogue participants are necessarily *aware* of the intentions that are ascribed to them by a dialogue act analysis. Second, while we do not take linguistic form to be part of the definition of a communicative function, we do insist that every communicative function which occurs in a taxonomy must be empirically justified, in the sense that there are ways in which a speaker can indicate that his behaviour should be understood as having that particular function through the form of his behaviour. This requirement puts communicative functions on an empirical basis.

The distinction between deep and more shallow approaches is relevant in connection with the different requirements of human and automatic annotation. Human annotators are better at understanding and annotating dialogue utterances in a detailed manner, because they have more knowledge of intentional behaviour and they have richer context models. Since a general dialogue annotation schema should support human annotation, it should contain concepts with a depth and granularity that matches human understanding of the functions of dialogue utterances. In order to support automatic annotation, on the other hand, the schema should also contain concepts that are suitable for a more shallow form of annotation, that relies less more on surface features and less on deep semantic knowledge. These two requirements can be met by defining hierarchies of communicative functions, where functions deeper down in a hierarchy correspond to more specific intentions or assumptions on the part of the speaker than functions higher in the hierarchy.

2.2 Multifunctionality

Studies of human dialogue behaviour indicate that natural dialogue utterances are very often multifunctional. This is due to the fact that participation in a dialogue involves several activities beyond those strictly related to performing the task or activity for which the dialogue is instrumental (such as obtaining certain information, instructing another participant, negotiating an agreement, etc.). In natural conversation, among other things, dialogue participants constantly "evaluate whether and how they can (and/or wish to) continue, perceive, understand and react to each other's intentions" (Allwood, 1997). They share information about the processing of each other's messages, elicit feedback, and manage the use of time and turn allocation, of contact and attention, and of various other aspects. Communication is thus a complex, multi-faceted activity, and for this reasondialogue utterances are often multifunctional. A qualitative and quantitative analysis of this phenomenon (Bunt, 2007; 2009) shows that the multifunctionality of minimal functional segments in spoken dialogue on average amounts to 4-5 functions. For multimodal dialogues, where significant parts of the interaction is performed through nonverbal behaviour, the multifunctionality is even greater.

One of the requirements on a dialogue act taxonomy is that it should support the multifunctional analysis and specification of dialogue, and preferably do this in a way that explains the multifunctionality that is observed in natural dialogue.

3. DIALOGUE ACT TAXONOMIES

3.1 Taxonomy structures

Existing dialogue act taxonomies differ not only in their precise sets of tags, but more importantly with respect to (1) the underlying approach to dialogue modeling; (2) the definitions of the basic concepts; (3) whether the tags are mutually exclusive; (4) the coverage of aspects of interaction; and (5) the level of granularity of the defined tag set. Generally, dialogue act taxonomies can be divided into one- and multidimensional ones.

One-dimensional schemes have a set of mutually exclusive tags, and support coding dialogue utterances with a single tag. Their tag sets are often quite small (such as the LINLIN schema (Ahrenberg et al., 1995) and the HCRC schema (Carletta et al., 1996)), and have the form of a simple flat list. The simplicity of these tag sets is usually considered to make them more reliable and to take less effort to apply consistently by annotators. It has been noted, however, that one-dimensional annotation schemas also have serious disadvantages (see e.g. Klein et al., 1998; Larsson, 1998; Popescu-Belis, 2005), in particular in view of the pervasive multifunctionality of natural dialogue behaviour.

Multidimensional taxonomies support dialogue utterances to be coded with multiple tags and have a relatively large tag set; see e.g. Allen & Core (1997), Larsson (1998); Popescu-Belis (2005), Bunt (2006); Bunt & Schiffrin (2006). Such a large tag set would benefit in several respects from having some internal structure.

First, clustering semantically related tags improves the transparency of the tag set, as the clusters indicate the kind of semantic information that is considered. The introduction of clusters of tags also makes the coverage of the tag set clearer, since the clusters will often corresponds to certain classes of dialogue phenomena.

Second, a taxonomical structure which is based on semantic clustering may support the decision-making process of human annotators: an initial step in such a process can be the decision to consider a particular cluster, and subsequently more fine-grained distinctions may be tested in order to decide on a specific tag within the cluster. Also, the tags within a cluster are typically either mutually exclusive (such as 'signal understanding' and 'signal non-understanding'), or are related by an entailment relation (such as a 'confirmation' also being an 'answer'); in both cases, an annotator should choose only one tag from the cluster. In this way the organisation of the tag set supports annotators in avoiding the consideration of inconsistent or irrelevant combinations of tags.

Third, a hierarchical organisation in the tag set may also be advantageous for automatic annotation and for achieving annotations which are compatible though not identical with those of human annotators. The choice of a particular cluster of tags can typically be made on the basis of less information than that of a particular tag within the cluster. The same is true for choosing a more general tag from the cluster versus a more specific tag (e.g. 'answer' versus 'disconfirmation'). Human (expert) annotation is often more detailed than automatic annotation because of the difference in semantic information that is effectively available. Automatic and human annotation are therefore often not identical, but still may be highly compatible. This can be expressed and measured precisely by taking the semantic relations within a cluster into account for computing annotator-agreement scores (Geertzen & Bunt, 2007). A structured tag set can be searched more systematically (and more 'semantically') than an unstructured one, and this can clearly have advantages for dialogue annotation, interpretation, and generation.

3.2 Clustering and Dimensions

The clusters of communicative functions that can be found in existing annotation schemes are typically characterized by notions of intuitive conceptual similarity, such as the clusters of questions and statements called 'info-request' and 'statement' in the DAMSL taxonomy.

DAMSL (Dialogue Act Markup using Several Layers) is the first and most frequently used annotation scheme that implements the multidimensional approach (Allen & Core, 1997), allowing multiple labels to be assigned to utterances in four layers: Communicative Status, Information Level, Forward-Looking Function (FLF) and Backward-Looking Function (BLF). The FLF layer is subdivided into seven classes, including (roughly) the classes of commissive and directive functions, well known from speech act theory; the BLF layer has four classes, as shown in Table 3.2.

These classes are also referred to as 'dimensions' (Core and Allen, 1997). While the DAMSL documentation does not discuss the notions of 'layer' and 'dimension' as such, the various ways of clustering the tag set are clearly useful for introducing some structure in the tag set and for providing annotation guidelines that can benefit from this structure. Clusters or 'dimensions' like those in DAMSL are usually defined as a set of mutually exclusive functions, related to the same type of information, such as the set {opening, closing} that constitutes the dimension called 'conventional'. Bunt (2006) has shown that this approach to clustering does not always lead to a notion of dimension that has any conceptual and theoretical significance, and that provides a consistent account of the observed multifunctionality of dialogue utterances.

Layer	*Dimension*
Forward-Looking Functions	statement
	info-request
	influencing-addressee-future-action
	committing-speaker-future-action
	conventional
	explicit-performative
	exclamation
	other-forward-function
Backward-Looking Functions	agreement
	understanding
	answer
	information-relation

Table 1: Layers and dimensions in DAMSL.

Popescu-Belis (2005) argues that dialogue act tagsets should seek a multidimensional theoretical grounding, and defines the following aspects of utterance function that could be relevant for choosing dimensions: (1) the traditional clustering of illocutionary forces in speech act theory into five classes: Representatives, Commissives, Directives, Expressives and Declarations; (2) turn management; (3) adjacency pairs; (4) topical organization in dialogue; (5) politeness functions; and (6) rhetorical roles.

Bunt (2004; 2006) suggests that a theoretically grounded multidimensional schema should be based on a theoretically grounded notion of dimension, and proposes to define a *set of dimensions* as follows.

(1) Each member of a set of dimensions is a cluster of communicative functions which all address a certain aspect of participating in dialogue, such that:

1. dialogue participants can address this aspect through linguistic and/or nonverbal behaviour which has this purpose;
2. this aspect of participating in a dialogue can be addressed independently of the other aspects corresponding to elements in the set of dimensions, i.e., an utterance can have a communicative function in one dimension, independent of its functions in other dimensions.

The first of these conditions means that only aspects of communication are considered that can be distinguished according to empirically observable behaviour in dialogue. The second condition requires dimensions to be independent, 'orthogonal'. A set of dimensions that satisfies these requirements can be a good foundation for a multidimensional annotation scheme, especially if the set of functions within each dimension is defined in such a way that any two functions are either mutually exclusive or have an entailment relation, since it would follow that a markable can be annotated with (maximally) as many tags as there are dimensions, one function (at most) for each dimension.

Note that the use of an information-state update semantics for dialogue acts, as underlying DIT, is helpful when designing a multidimensional taxonomy, because it supports the formulation of precise definitions of communicative functions, clarifying the relations between them and providing a formal basis for what may intuitively seem to be sets of related functions, and thus for identifying potential dimensions.

Petukhova (forthc.) provides an up to date survey of the use of communicative functions related to various dimensions in 18 existing annotation schemas. She presents test results, based on co-occurrence frequencies, phi-statistics, and vectorial distance measures to empirically determine to what extent proposed dimensions are well-founded. One of the conclusions from this study is that the dimensions of the DIT^{++} taxonomy, described below, do indeed form a well-defined set of dimensions.

3.3 General-purpose and dimension-specific functions

When we view a dimension in dialogue analysis or specification as a particular aspect of interacting, like topic management, turn management, or trying to perform a certain task, then we see that dialogue acts like questions and answers do not belong to any dimension. One can ask a question about something in the task, or a about agreeing to close a topic, or about whose turn it is to say something, or about any other aspect of interacting, so questions can be said to belong to *all* these dimensions. Similarly for answers, statements, requests, offers, agreements, (dis-)confirmation, and so on. Clusters of general such dialogue acts, which belong to what in speech act theory are sometimes called *core speech acts*, therefore should not be considered as forming certain dimensions, but as *general-purpose functions* that can be used in any dimension. This in contrast with communicative functions that are specific for a particular dimension, such as Turn Keep, Turn Release, Introduce Topic, Change Topic, Apology and Thanking. On this view, which has been developed in DIT, a taxonomy of communicative functions consists of two parts:

1. a set of clusters of *general-purpose functions*;
2. a set of clusters of *dimension-specific functions*.

For the DIT^{++} taxonomy, Table 4.1 shows the structure of the first part (the general-purpose functions) with the main functions in the various clusters; the complete set of functions is shown in Annex A. Table 4.1 lists examples of dimension-specific communicative functions in each of the DIT^{++} dimensions; the complete taxonomy of dimension-functions is shown in Annex A.

General-purpose functions can be used to build a dialogue act in any dimension, depending on the type of semantic content that such a function is combined with. Therefore an adequate annotation or specification of a markable should in general have two components: the communicative function and the dimension that is addressed, as in the following examples. (If the communicative function is a dimension-specific function, then the specification of the dimension is redundant, if the names of these functions have been chosen to be unique.)

(2) a. Please repeat that.
<Feedback, Request>

b. Jim, your turn.
<Turn Management, Instruct>

c. I am very grateful for your help
<Social Obligations Management, Inform>

d. You got that?
<Allo-Feedback, CheckQuestion>

4. DIT++ DIMENSIONS AND FUNCTIONS

4.1 General-purpose functions

The general-purpose communicative functions in the DIT^{++} taxonomy fall into two broad categories:

- the *Information Transfer* functions, which aim at seeking or providing information, and are subdivided accordingly in *information seeking* (or ‘questioning’) and *information providing* functions;

- *Information Transfer Functions*
  - *Information-Seeking Functions*
    - *Direct Questions*
      - propositional question, set question, alternatives question, check question, etc.
    - *Indirect Questions*
      - indirect propositional question, set question, alternatives question, check question, etc.
  - *Information-Providing Functions:*
    - *Informing Functions:*
      - inform, agreement, disagreement, correction;
      - *Informs with Rhetorical or Attitudinal Functions, such as* elaboration, justification, exemplification.. and warning, threat,..
    - *Answer Functions:*
      - propositional answer, set answer, confirmation, disconfirmation
- *Action Discussion Functions*
  - *Commissives*
    - offer, promise, address request
    - *other commissives, expressable by means of performative verbs*
  - *Directives:*
    - instruction, address request, indirect request, (direct) request, suggestion
    - *other directives, such as advice, proposal, permission, encouragement, urge,..., expressable by means of performative verbs*

Table 2: Structure of the DIT^{++} taxonomy of general-purpose communicative functions.

- the *Action Discussion* functions, which aim at bringing certain actions into the discussion that may or should be performed by the speaker, by the addressee, or jointly, and which are subdivided into *commissives*, where the speaker is (conditionally) committing himself to a certain action, and *directives*, where the speaker is putting pressure on the addressee to (conditionally) perform or participate in a certain action.

Table 4.1 shows the structure of the taxonomy of general-purpose functions and a number of functions that inhabit this structure. The complete set of functions can be found in Annex A.

Within the subcategory of questions various question types are distinguished, such as questions which enquire after the truth of a proposition (‘propositional questions’, often also called yes-no questions), questions which aim at identifying those elements of a given set which have a certain property (‘set questions’, often also called ‘WH-questions’ after the question words that are mostly used for expressing this type of questions in English), and questions which aim at discriminating between two ore more alternatives (‘alternatives question’, also known as ‘multiple-choice question’). These three types of question occur in two variants in the taxonomy: a direct and an indirect version. The difference is that in the direct case the speaker expresses an assumption that the addressee knows the answer; in the indirect version no such assumption is expressed. This is one possible way of making a semantic distinction (which one may or may not want to make) between questions like *What time is it?* and *Where is Harry’s office* on the hand, and questions like *Do you know what time it is?* and Ĉan you tell me where Harry’s office is? on the other.

The subcategory of information-providing function falls apart in those functions where the speaker is providing information in

Dimension	*Dimension-specific ommunicative functions*	*Typical expressions*
Task/Activity	OpenMeeting, CloseMeeting; Appoint, Hire, Fire	domain-specific fixed expressions
Auto-Feedback	PerceptionNegative	*Huh?*
	EvaluationPositive	*True.*
	OverallPositive	*OK.*
Allo-Feedback	InterpretationNegative	*THIS Thursday.*
	EvaluationElicitation	*OK?*
Turn Management	TurnKeeping	final intonational rise
	TurnGrabbing	hold gesture with hand
	TurhGiving	*Yes.*
Time Management	Stalling	slowing down speech; fillers
Contact Management	ContactChecking	*Hello?*
Own Communication Management	SelfCorrection	*I mean...*
Partner Communication Management	PartnerCompletion	completion of partner utterance
Discourse Structure Management	DialogueActAnnouncement	*Question.*
	TopicShiftAnnouncement	*Something else.*
Social Obligations Management	Apology	*I'm sorry.*
	Greeting	*Hello!, Good morning.*
	Thanking	*Thanks.*

Table 3: Examples of dimension-specific communicative functions and typical expressions per dimension.

response to an information need that the addressee has signalled ('answers') and those where the motivation to provide information comes from the speaker: he wants the addressee to know or be aware of something, as in a teaching environment, or in the case of a warning ('informing functions'). Both the answering and the informing functions in a number of cases come in two varieties: a 'plain' one and an uncertain one. Especially for answers this is important: when asked a question, a respondent who is uncertain about the correctness or completeness of his answer will often indicate this, verbally and/or nonverbally, with correspondingly different effects on the information state of the addressee. Note that the subcategory of informing functions includes an open subclass of functions where the speaker's goal of informing the addressee of something is further specified in having a certain rhetorical, emotional, or evaluative function. This is one of several points where the DIT^{++} taxonomy has an open subclass.

The category of action-discussion functions corresponds essentially to the classes of commissives and directives from speech act theory; hence these names are used for the two subcategories of this category. Both the commissive and directive subcategories have open subclasses ot communicative functions, to accomodate the wide range of performative verbs that one may wish to distinguish at the semantic level of dialogue acts.

Many commissive and directive acts come in pairs, where one act brings an action into the discussion (proposing, or instructing, or promising,.. to perform that action), and another act is concerned with accepting or rejecting the performance of that action. If the first of these dialogue acts is a directive act, then the second is a commissive act, and vice versa. Different from other taxonomies and theories, the DIT^{++} taxonomy does not have separate functions like Accept Request and Decline Request, but a single function Address Request. The reason for this is that, apart from accepting and rejecting a request, a dialogue participant can also accept (or reject) a request conditionally, or with certain qualifications. (*I will do that only if you....*). This phenomenon occurs more generally for dialogue acts discussing actions, since actions can be done conditionally, repeatedly, with a certain intensity, and so on, and in general can be qualified in many ways, more than a proposition as the topic in information exchange acts. (See Bunt & Schiffrin, 2007, for more discussion on these and related issues.)

4.2 Dimensions

The ten dimensions of DIT^{++} have emerged from an effort to provide a semantics for dialogue utterances across a range of dialogue corpora. Utterances have been identified whose purpose was to address the following aspects of participating in a dialogue: (1) the performance of a task or activity motivating the dialogue; (2) the monitoring of contact and attention; (3) feedback on understanding and other aspects of processing dialogue utterances; (4) the allocation of the sender role; (5) the timing of contributing to the dialogue; (6) the structuring of the dialogue and the progression of topics; (7) the editing of one's own and one's partner's contributions; (8) the management of social obligations. Whether these aspects qualify as dimensions can be determined by checking the applying the above criteria (1).

Take for instance the timing of contributions. Utterances that address this aspect of interacting include those where the speaker wants to gain a little time in order to determine how to continue the dialogue; this function is called Stalling. Speakers indicate this function by slowing down in their speech and using fillers, as in *Ehm, well, you know,...* The observation that dialogue participants exhibit such behaviour means that the category of functions addressing the timing of contributions (which also includes the act of Pausing, realized by means of utterances like *Just a minute, Hold on a second*) satisfies criterion (1-1). Moreover, the devices used to indicate the Stalling function can be applied to virtually *any* kind of utterance, which may have have any other function in any other dimension. Timing therefore satisfies criterion (1-2) as well, and hence qualifies as a proper dimension.

A similar analysis can be applied to the other aspects. Of these, the feedback category (3) should be divided into two, depending on whether a speaker gives feedback on his own processing, or whether he gives or elicits feedback on the addressee's processing; we call these dimensions 'auto-feedback' and 'allo-feedback', respectively (cf. Bunt, 1995). Similarly, the category of dialogue acts concerned with editing one's own or one's partner's contributions, is better split into those concerned with editing one's own speech, called the Own Communication Management (OCM) dimension (using Allwood's terminology - see Allwood, 1997), and those concerned with the correction or completion of what the current speaker is saying, which by analogy we call the Partner Communication Management (PCM) dimension. See the examples of communicative functions within each of these dimensions, with common utterance forms in English, in table 4.1. Dialogue acts with a dimension-specific function are often performed partly or entirely nonverbally, such as positive feedback by nodding, negative feedback by frowning, or turn assignment by direction of gaze. A study by Petukhova (2005), preformed in the context of the EU project AMI (see footnote 4), showed that all the communicative functions of the nonverbal behaviour of participants in AMI meetings could be described adequately in terms of the DIT^{++} functions, and produced a catalogue of nonverbal means (notably head gestures, facial expressions, and gaze behaviour) for expressing DIT^{++} communicative functions, either by themselves or in combination with verbal behaviour.

All in all, this had lead to the following 10 dimensions in the DIT^{++} taxonomy:

1. Task/Activity: dialogue acts whose performance contributes to performing the task or activity underlying the dialogue

2. Auto-Feedback: dialogue acts that provide information about the speaker's processing (perception, interpretation, evaluation, or application) of the previous utterance or some particular previous utterance(s). Note that feedback is called 'positive' here if the processing at the level that is addressed, is or has been successful; negative feedback indicates a processing problem. Note also that 'evaluation' here means that the update information, which has been constructed by successful understanding of a diialogue segment, is evaluated and checked for not leadng to an inconsistent information state, or to an otherwise problematic situation. A positive evaluation leads to a process at the 'execution' level, where the participant's information state is indeed changed, and possibly further action is taken. For instance, the positive evaluation of a question leads to a decision to go ahead and try to answer the question; 'executing' a question means determining the answer to it. Similarly, evaluating an answer is deciding whether its content can be accepted without harm for the information state, and executing the answer is going ahead and integrate its content into the participant's information state. For signalling one's 'evaluation' of information in the sense of forming an *attitude* towards it, such as surprise or disappointment, an open subclass of functions has been added to the positive evaluation feedback function, similar to the open subclasses for informs with a rhetorical or attitudinal function.

3. Allo-Feedback: dialogue acts used by the speaker to express opinions about the addressee's processing (perception, interpretation, evaluation, or application) of the previous utterance or some particular previous utterance(s), or that solicit information about that processing;

4. *Contact Management*: dialogue acts for establishing and maintaining contact;

5. Turn Management: dialogue acts concerned with grabbing, keeping, giving, or accepting the sender role;

6. Time Management: dialogue acts signalling that the speaker needs a little time to formulate his contribution to the dialogue, or that his preparation for producing a contribution requires so much time that the interaction has to be suspended for a while (which may be due to various factors, such as something urgent intervening);

7. Discourse Structuring: dialogue acts for explicitly structuring the conversation, e.g. announcing the next dialogue act, or proposing a change of topic;

8. Own Communication Management: dialogue acts to indicate that the speaker is editing the contribution to the dialogue that he is currently producing;

9. Partner Communication Management: the agent who performs these dialogue acts has the addressee rather than the speaker role, and assists or corrects the dialogue partner in his formulation of a contribution to the dialogue;

10. Social Obligations Management: dialogue acts that take care of social conventions such as welcome greetings, apologies in case of mistakes or inability to help the dialogue partner, and farewell greetings.

5. USING DIT++

The DIT^{++} taxonomy has been and is being used for a variety of purposes:

1. for empirical and theoretical analysis and computational modelling of semantic and pragmatic phenomena in spoken and multimodal dialogue;

2. for annotating dialogues in order to build corpora with well-founded multidimensional annotation of communicative functions;

3. for designing components of dialogue systems, in particular for multimodal input interpretation, dialogue management, and the generation of multifunctional dialogue behaviour in spoken or multimodal dialogue systems;

4. as a well-defined comprehensive, general-purpose taxonomy that unifies and incorporates insights from a range of earlier efforts and projects, it has served as the starting point for an ISO effort to define interoperable concepts for dialogue act annotation.

In this section we briefly consider each of these uses of the DIT^{++} taxonomy.

5.1 Multimodal Dialogue Analysis

Information flow and grounding. Every communicative function in the DIT^{++} taxonomy is formally defined as a particular type of update operation on an addressee's information state. Information states, called 'contexts' in DIT, are viewed as being highly structured, with various components of a structured dialogue context corresponding to various aspects of interacting as reflected in the dimensions of the taxonomy. Depending on its dimension, a dialogue act updates a particular context component; a multifunctional utterance leads to the update of several components. This approach provides good instruments for studying and modelling the flow of information between the participants in a dialogue. Fine-grained models of information flow through the understanding of dialogue behaviour in terms of DIT^{++} dialogue acts have been developed and analysed in Morante (2007), and have resulted in a empirically-based computational model of *grounding* in dialogue (Bunt & Morante, 2007; Bunt et al., 2007).

Semantics of discourse markers. Another pragma-semantic study

*

within	Task	Auto-F.	Allo-F.	Turn M.	Time M.	DS	Contact M.	OCM	PCM	SOM
Task		1.1 (1.2)	0.1 (2.7)	5.6 (8.5)	2.6(3.4)	0.3(0.3)	0(0)	4.3(4.6)	0.3(0.3)	1.5(1.5)
Auto-F.	0.5(0.7)		0(0)	12.7(15.5)	0.5(2.6)	0.3(3.1)	0(0)	0(0)	0(0)	0(0.5)
Allo-F.	0(3.3)	0(0)		23.7(23.7)	1.2(1.5)	0(0)	0(0)	0(15.4)	0(5.1)	0(0)
Turn M.	39.3(40.8)	6.2(12.2)	1.8(6.0)		49.6(60.6)	0.7(1.1)	0(0.3)	2.5(5.9)	0(0.7)	0.4(0.7)
Time M.	34.6(41.7)	0.5(3.5)	0(11.2)	9.1(9.7)		0(0.5)	0(0)	0(4.2)	0(1.4)	0(0.6)
DS	1.7(6.8)	0(6.8)	0(0)	6.7(20.9)	0(1.7)		0(0)	0(1.7)	0(0)	0(8.4)
Contact M.	0(0)	0(0)	0(0)	18.2(18.2)	0(0)	0(0)		0(0)	0(0)	0(0)
OCM	77.9(80.9)	0(0)	0(5.4)	6.5(6.5)	0(8.0)	0(0.9)	0(0)		0(0)	0(0)
PCM	0(0)	0(0)	0(18.2)	27.3 (27.3)	0(0)	0(0)	0(0)	0(0)		0(0)
SOM	0.9(0.9)	0(1.2)	0(0)	1.2(8.3)	0(1.2)	13.9(13.9)	0(0)	0(0)	0(0)	

*

Table 4: Co-occurrences of communicative functions across dimensions in AMI corpus expressed in relative frequency in %, with and without nonverbal behaviour taken into account (in brackets).

based on the multidimensional approach of the DIT^{++} taxonomy is that of the semantics of discourse markers; words or phrases that connect the pieces in a dialogue (or in a monologue), like *but, and, so, well,* etc. It was shown that such words often perform multiple semantic functions, which are well explained in terms of the dimensions and the view of multifunctionality represented in the DIT^{++} taxonomy (Petukhova & Bunt, 2009).

The meaning of nonverbal dialogue behaviour. Petukhova (2005) investigated the applicability of the DIT^{++} taxonomy to nonverbal behaviour in dialogues in the AMI corpus. It was found that the DIT^{++} functions provided fulll coverage for interpreting the nonverbal activity. The nonverbal behaviour may serves four purposes: (1) emphasizing or articulating the semantic content of dialogue acts; (2) emphasizing or supporting the communicative functions of the synchronous verbal behaviour; (3) performing separate dialogue acts in parallel to what was contributed by the partner (without turn shifting); or (4) expressing a separate communicative function in parallel to what the same speaker is expressing verbally. It was recently found (Petukhova 2009, p.c.) that the latter purpose occurs much less than the other three, as witnessed by the fact that the multifunctionality of dialogue segments taking nonverbal behaviour into account shows only a small increase compared to the case where nonverbal behaviour is not taken into consideration.

Multifunctionality and co-occurrence patterns. To generate multifunctional dialogue behaviour in a sensible way, it is important to have qualitative and quantitative knowledge of this phenomenon, and to know which kinds of multifunctional utterances occur in natural dialogue. Studies by Bunt (2007; 2009) and Petukhova (p.c.) have shown that, when a very fine-grained segmentation is applied to dialogue, with very small and possibly overlapping and interleaved functional segments as markables, the average multifunctionality of a markable in spoken dialogue without visual contact amounts to 3.7. With visual contact this is of course higher (and there is an increase of more than 25% of the total number of segments, mostly for participants not in the speaker role providing nonverbal feedback). Table 4.2 (from Petukhova, p.c. 2009) summarizes the co-occurrence data that were found for communicative functions in each pair of DIT^{++} dimensions, with and without taking nonverbal signals into account. (Based on data from the AMI corpus.)[2] Each row in the table describes the relative number of times that an utterance addressing the corresponding dimension, also addressed the dimensions corresponding with the columns. For the most frequently addressed dimensions (the top six rows of the table), the most important cases where nonverbal signals added multifunctionality ate the following:

Task: Auto- and Allo-Feedback, Turn Management, and OCM;
Auto-Feedback: Task, Turn Man., Contact Management, PCM;
Allo-Feedback: Task, Turn Man., Time;
Turn Management: Task, Auto- and Allo-Feedback, Time, Contact Man., OCM;
Time Management: Task, Auto-Feedback. Turn Man., OCM.
Discourse Structure Man.: Task, Auto- and Allo-Feedback, Turn Man.,Contact Man., SOM

The nonverbal signals taken into account here include gaze behaviour and head and hand movements; they do not include facial expressions, which is undoubtedly a rich further source of communicative functionality. It can be observed that the addition of nonverbal signals has an effect for all ten dimensions, the most important effects (in terms of frequency of occurrence) being that nonverbal signals are used for feedback, turn management, and own communication management. These figures also indicate clearly that multifunctionality across dimensions is a very real and important phenomenon in natural dialogue.

5.2 Annotation

The DIT^{++} taxonomy has been applied in manual annotation of dialogues from various corpora: the DIAMOND corpus of two-party instructional human-human Dutch dialogues (1,408 utterances); the AMI corpus of task-oriented human-human multi-party English dialogues (3,897 utterances); the OVIS corpus of task-oriented human-computer Dutch dialogues (3,942 utterances); TRAINS dialogues (in English); and Map Task dialogues both in English and in Dutch. Geertzen et al. (2008) report on the consistency with which naive annotators as well as expert annotators were able to perform annotation, and compares the results. Expert annotators achieve agreements scores of over 90%; naive annotators achieve scores in the order of 60%. .

The LIRICS taxonomy of communicative functions, which is a slightly simplified version of the DIT^{++} taxonomy, was tested in manual annotation of test suites in Dutch, English, and Italian, with very high consistency - see subsection 5.4 and table 5.4.

5.3 Dialogue system building

Dialogue management. The DIT^{++} taxonomy has been used in the design and implementation of the PARADIME dialogue manager, that forms part of the IMIX system for multimodal information extraction; see Keizer & Bunt (2006; 2007). The multidimensional dialogue manager generates sets of dialogue acts (in formal

[2] **A**ugmented **M**ulti-party **I**nteraction (`http://www.amiproject.org/`)

representation) that are appropriate in the current dialogue context, and delivers these to a module for expressing sets of dialogue acts in multifunctional utterances. This opens the opportunity to generate multifunctional utterances in a deliberate and controlled fashion.

Machine recognition of DIT++ functions. A prerequisite for using dialogue acts in a dialogue manager is that th dialogue system is able to recognize dialogue acts sufficiently well. The automatic recognition of dialogue in the DIT^{++} taxonomy (as well as in other taxonomies, such as DAMSL) was investigated for the corpora mentioned above, as well as for dialogues from the Monroe and MRDA corpora. For the various dimensions of the DIT^{++} taxonomy, F_1 scores were found ranging from 62.6 to 96.6%, without tweaking the feature use in the machine learning algorithms. This suggests that the recognition of (multiple) functions in the taxonomy is a realistic enterprise. For details see Geertzen (2009).

5.4 Towards a standard for functional dialogue markup

In 2008 the International Organization for Standards started up the project Semantic Annotation Framework, Part 2: Dialogue acts, which aims at developing an international standard for the markup of communicative functions in dialogue. This project builds on the results of an ISO study group on interoperability in linguistic annotation, of which the European project LIRICS[3] was a spin-off. In the LIRICS project, a taxonomy of communicative functions was defined by simplifying the DIT^{++} taxonomy a little, retaining its dimensions but eliminating the distinction of levels of feedback as well as the uncertain variants of information-providing functions and the informs with rhetorical functions, and excluding some of the low-frequency functions. The resulting LIRICS taxonomy has 23 general-purpose functions (where DIT^{++} has 34 plus 3 open classes) and 30 dimension-specific functions (where DIT^{++} has 55, of which 20 dine-grained feedback functions).

The LIRICS taxonomy was applied by three expert annotators to the LIRICS test suite dialogues in Dutch and English. Unusually high, near-perfect agreement was found between the annotators, as shown in table 5.4 (standard κ-values).

Function category	*Annotator agreement* (κ)
information-seeking	0.97
information-providing	0.98
action discussion	0.99
auto-feedback	0.99
allo-feedback	1.00
interaction management	0.94
social obligations management	0.94

Table 5: LIRICS annotation statistics

The ISO project takes the DIT^{++} and LIRICS taxonomies as point of departure for defining a comprehensive open standard for functional dialogue markup. For the current status of the project see ISO (2009).

6. CONCLUSIONS

In this paper we presented the DIT$++$ taxonomy of communicative functions. We indicated some of its theoretical background and its applications in human and machine annotation and dialogue management and generation. Co-occurence date for communicative functions, indicating the naturally occurring combinations of communicative functions, may be useful for deliberately generating multifunctional dialogue behaviour, which is especially important in multimodal contexts like those of embodied conversational agents, where facial expressions, gestures, and language together should be used to achieve natural forms of multifunctional behaviour.

7. REFERENCES

Ahrenberg, L., N.Dahlbäck & A.Jönsson (1995) Codings Schemes for Studies of Natural Language Dialogue. In: *Working Notes from the AAAI Spring Symposium*, Stanford.

Allen, J. & M. Core (1997) DAMSL: Dialogue Act Markup in Several Layers (Draft 2.1). Technical Report, Multiparty Discourse Group, Discourse Resource Initiative, September/October 1997.

Allwood, J., & E. Ahlsén, J. Nivre, and S. Larsson (1997) *Own Communication Management: Kodningsmnual.* Gothenburg University, Department of Linguistics.

Allwood, J. (2000) An activity-based approach to pragmatics. In H. Bunt & W. Black(eds.) *Abduction, Belief and Context in Dialogue. Studies in Computational Pragmatics.* Amsterdam: Benjamins, 47–80.

Austin (1962) *How to do things with words*. Oxford: Clarendon Press.

Bunt,vH. (1990) Dynamic Interpretation in Text and Dialogue. In H. Bouma (ed.) *Working Models of Human Perception and Cognition*. New York: Academic Press.

Bunt, H. (1995) Dynamic Interpretation and Dialogue Theory. in M.Taylor, D. Bouwhuis & F. Nél (eds.) *The Structure of Multimodal Dialogue, Vol. 2.* Amsterdam: Benjamins, 139–166.

Bunt, H. (2000) Dialogue pragmatics and context specification. In H. Bunt & W. Black(eds.) *Abduction, Belief and Context in Dialogue. Studies in Computational Pragmatics.* Amsterdam: Benjamins, 81–150.

Bunt, H. (2005) A Framework for Dialogue Act Specification. *4th Joint ISO-SIGSEM Workshop on the Representation of Multimodal Semantic Information*, Tilburg, January 2005. `http://let.uvt.nl/research/ti/sigsem/wg`

Bunt. H. (2006) Dimensions in dialogue annotation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006).*

Bunt. H. (2007) Multifunctionality and multidimensional dialogue act annotation. In: E. Ahlsen et al. (eds.)*Communication – Action – Meaning*. Gothenburg University, pp. 237 – 259.

Bunt. H. & R. Morante (2007) The Weakest Link. In V. Matousek & P. Mautner (eds.) *Text, Speech and Dialogue*. Springer, Lecture Notes in Artificial Intelligence 4629, 591–598.

Bunt, H. & Y. Girard (2005) Designing an open, multidimensional dialogue act taxonomy. In C. Gardent & B. Gaiffe (eds.) *DIALOR'05, Proceedings of the Ninth Workshop on the Semantics and Pragmatics of Dialogue* , Nancy, June 2005, 37–44.

Bunt, H. & S. Keizer (2005) Dialogue semantics links annotation fo context representation. In *Joint TALK/AMI Workshop on Standards or Multimodal Dialogue Context*, Edinburgh, December 2005. `http://homepages.inf.ed.ac.uk/olemon/standcon-SOI.html`

Bunt, H., S. Keizer & R. Morante (2007) An empirically-based computational model of grounding in dialogue. In *Proc. SIGDIAL 2007*, Antwerp, 283–290.

Bunt, H. & A. Schiffrin (2006) Methodological aspects of semantic annotation. In *Proceedings LREC 2006*, Genova, May 2006.

[3] `http://lirics.loria.fr`

Bunt, H. & A. Schiffrin (2007) Defining interoperable concepts for dialogue act annotation. In *Proc. 7th International Workshop on Computational Semantics (IWCS-7*, Tilburg, 16–27.

Bunt, H. & L. Romary (2004) Standardization in Multimodal Content Representation: Some Methodological Issues. *Proceedings LREC 2004*, June, Lisbon, 2219–2222.

Carletta, J., A. Isard, S. Isard, J.Kowtko & G. Doherty-Sneddon (1996) HCRC dialogue structure coding manual. Technical Report HCRC/TR-82.

Core, M., & J.F. Allen (1997) Coding dialogues with the DAMSL annotation scheme. In *Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35.

Geertzen, J. (2009) *Dialogue act recognition and prediction.* PhD Thesis, Tilburg University, February 2009.

Geertzen, J.& H. Bunt (2006) A weighted kappa for measuring inter-annottor agreement. In *Proceedings SIGDIAL 2007,* Sydney, Australia.

Geertzen, J., V. Petukhova & H. Bunt (2007) A multidimensional approach to dialogue segmentation and dialogue act classification. in *Proc. SIGDIAL 2007*, Antwerp, 140–147.

ISO (2009) Language resource management – Semantic Annotation Framework - part 2: dialogue acts. ISO document ISO/TC 37/SC 4/N442 rev 02, February 2009. ISO, Geneva.

Keizer, S.& H. Bunt (2006) Multidimensional dialogue management. In *Proceedings SIGDIAL 2006*, Sydney, Australia.

Keizer, S.& H. Bunt (2007) Evaluating combinations of dialogue acts for generation. In *Proceedings SIGDIAL 2007*, Antwerp, Belgium.

Klein, M. (1999) *Standardization efoorts on the leve lof dialogue act in the MATE project.* Available at `http://acl.ldc.upenn.edu/W/W99?W99-0305.pdf`.

Larsson, S. (1998) Coding Schemas for Dialog Moves. *Technical report from the S-DIME project*. See `http://www.ling.gu.se/ sl`

Morante, R. (2007) *Computing meaning in interaction.* PhD Thesis, Tilburg University.

Petukhova, V.V. (2005) Multidimensional interaction of multimodal dialogue acts in meetings. MA thesis, Tilburg University.

Petukhova, V.V. (forthc.) Dimensions of communication: a survey. Technical Report, Tilburg University.

Popescu-Belis, A. (2005) Dialogue Acts: One or More Dimensions? *ISSCO Working Paper 62, ISSCO*.

Searle, J.R. (1969) *Speech Acts*. Cambridge University Press.

Traum, D. & S. Larsson (2003) The Information State Approach to Dialogue Management. In. R. Smith & J. van Kuppevelt (eds.) *Current and New Directions in Discourse and Dialogue.* Dordrecht: Kluwer, 325–353.

8. ANNEX A. THE DIT^{++} TAXONOMY

The DIT^{++} taxonomy in its current version has been stable for the last two years. Occasionally, small improvements have been made in some of the definitions and guidelines. For the latest version see `http://dit.uvt.nl`.

As described above, the full set of communicative functions consists of (a) a taxonomy of general-purpose functions, and (b) one of dimension-specific functions.

General-purpose functions

The full set of general-purpose functions, displayed in Table 6, is a superset of the taxonomy in Table 2.

- *Information Transfer Functions*
  - *Information-Seeking Functions*
    - *Direct Questions*
      - propositional question
        - check question
          - posi-check
          - nega-check
      - set question
      - alternatives question
    - *Indirect Questions*
      - indirect propositional question
      - indirect set question
      - indirect alternatives question
  - *Information-Providing Functions:*
    - *Informing Functions:*
      - inform
        - agreement
        - disagreement
          - correction
        - *Informs with Rhetorical Functions, such as*
          - elaboration
          - justification
          - exemplification
          - . . .
        - *Informs with Attitudinal Functions, such as*
          - warning
          - threat
          - . . .
    - *Answer Functions:*
      - propositional answer
        - confirmation
        - disconfirmation
      - set answer
      - uncertain propositional answer
        - uncertain confirmation
        - uncertain disconfirmation
      - uncertan set answer
- *Action Discussion Functions*
  - *Commissives*
    - offer
      - promise
      - address request
        - accept request
        - decline request
      - address suggestion
        - accept suggestion
        - decline suggestion
    - *other commissives, expressable by means of performative verbs*
  - *Directives*
    - indirect request
      - (direct) request
    - instruct
      - address offer
        - accept offer
        - decline offer
    - suggestion
    - *other directives, such as* advice, proposal, permission, encouragement, urge,..., *expressable by means of performative verbs*

Table 6: DIT^{++} taxonomy of general-purpose communicative functions.

- *Dimension-Specific Communicative Functions*
  - *Task/Domain-Specific Functions*
    - *Functions, expressible either by means of performative verbs denoting actions*
    - *for performing tasks in a specific domain, or by means of nonverbal actions such as*
    - *highlighting, or pointing to something in a picture. For example:*
      - Open Meeting, Suspend Meeting, Resume Meeting, Close Meeting *(in meeting situations)*
      - Bet, Accep tBet, Reject Bet *(in betting situations)*
      - Congratulation, Condolance
      - Hire, Fire, Appoint,... *(in a human resource management domain)*
      - Show, Highlight, Point, List,... *(for performing graphical or multimodal dialogue acts)*
  - *Dialogue Control Functions*
    - *Feedback Functions*
      - *Auto-Feedback Functions*
        - Positive (= Unspecified Positive) Feedback
          - Attention Positive Feedback
            - Perception Positive Feedback
              - Interpretation Positive Feedback
                - Evaluation Positive Feedback
                  - Execution Positive (= Overall Positive) Feedback
        - Negative (= Unspecified Negative) Feedback
          - Execution Negative Feedback
            - Evaluation Negative Feedback
              - Interpretation Negative Feedback
                - Perception Negative Feedback
                  - Attenttion Negative (= Overall Negative) Feedback
      - *Allo-Feedback Functions*
        - *Allo-Feedback-Giving Functions*
          - Positive (= Unspecified Positive) Feedback
            - Perception Positive Feedback
              - Interpretation Positive Feedback
                - Evaluation Positive Feedback
                  - Execution Positive (= Overall Positive) Feedback
        - Negative (= Unspecified negative) Feedback
          - Evaluation Negative Feedback
            - Execution Negative Feedback
              - Interpretation Negative Feedback
                - Perception Negative Feedback
                  - Attention Negative Feedback
        - *Feedback Elicitation Functions*
          - Attention Feedback Elicitation
            - Perception Feedback Elicitation
              - Interpretation Feedback Elicitation
                - Evaluation Feedback Elicitation
                  - Execution Feedback Elicitation

Table 7: Dimension-specific communicative functions, part 1: functions for task performance and feedback.

Dimension-specific functions

The full set of dimension-specific functions is shown in tables 7 and 8, divided over two tables to enable the taxonomy to be represented on paper.) The reader is also referred to the website `http://dit.uvt.nl`, where the definitions of all the communicative functions can be found, plus guidelines for their use in annotation.

- *Interaction Management Functions*
  - *Turn Management Functions*
    - *Turn-unit-initial functions*
      - Turn Accept
      - Turn Grab
      - Turn Take
    - *Turn-unit-final functions*
      - Turn Assign
      - Turn Keep
      - Turn Release
  - *Time Management*
    - Stalling
    - Pausing
  - *Contact Management*
    - Contact Check
    - Contact Indication
  - *Own Communication Management*
    - Error signaling
      - Retraction
        - Self-correction
  - *Partner Communication Management*
    - Completion
    - Correct-misspeaking
  - *Discourse Structure Management*
    - Opening
    - Preclosing
    - Topic Introduction
    - Topic Change Announcement
      - Topic Shift
  - *Social Obligations Management*
    - *Salutation*
      - Initial greeting
      - Return greeting
    - *Self-introduction*
      - Initial self-introduction
      - Return self-introduction
    - *Apologizing*
      - Apology
      - Apology-downplay
    - *Gratitude Expression*
      - Thanking
      - Thanking-downplay
    - *Valediction*
      - Initial goodbye
      - Return goodbye

Table 8: Dimension-specific communicative functions, part 2: functions for Interaction Management and Social Obligations Management

Actions speak louder than words: An intentional look at defining communicative acts for embodied conversational agents

Debora Field
Dept Computer Science
University of Sheffield
Sheffield
S1 4DP, United Kingdom
d.field@sheffield.ac.uk

Allan Ramsay
School of Computer Science
University of Manchester
Manchester
M60 1QD, United Kingdom
allan.ramsay@manchester.ac.uk

ABSTRACT

This paper takes a philosophical look at how one might use insights from examining the nature of action in order to help in defining a set of multi-modal communicative actions for use in and by embodied conversational agents. A key theme of the paper is the central importance of intention in communicative actions. The paper also offers some suggestions on how to conceive of and categorise different types of action (communicative and non-communicative) so as to set a firm foundation before tackling the problem of defining a set of communicative actions for embodied conversational agents. It also gives some more specific practical advice on how to begin going about defining that set.

General Terms

Design

Keywords

Embodied Conversational Agents, Actions, Intention, Communicative Actions, Modality, Pragmatics

1. INTRODUCTION

If you take a group of natural language (NL) engineers[1] and ask them how they might define a set of actions that it might be suitable and useful for Embodied Conversational Agents (ECAs) to be able to perform, you create a marvellous opportunity for a great confusion of terminology and ideas. As we know from the call for papers for this workshop, there is a whole field of research into some objects that NL engineers these days tend to call 'Dialogue Acts'. This field defines 'act' in a particular narrow way that differs dramatically from, say, how a robotics engineer or an AI planning engineer, who generally have no concern with dialogue, would define 'act'. There are different taxonomies of Dialogue Acts (*e.g.*, DAMSL [9] and DIT++ [5] are prominent), but they have an important characteristic in common, owing to the catalysts of their emergence, and the problems they are designed to help with: they consider that the performance of an 'act' constitutes the production of a linguistic sign (typically a spoken or written utterance in a natural language), that this 'act' is performed by an agent who is trying to communicate some idea, and that the mechanics of how the act is performed is of secondary interest—that part is left to the speech synthesis and text display engineers.

This now traditional way of viewing Dialogue Acts is likely to turn out to be inadequate when NL engineers start to consider Function Markup Languages (FMLs) for ECAs. The point of making ECAs compared to non-embodied conversational agents, one presumes, is that there is some added value in the embodiment that would not be there without it. In many cases, judging by the current state of the art in several areas of technology, this added value will constitute modes of perception and production that have so far not been of much interest to NL engineers (*e.g.*, vision, sensation, mobility, dexterity, *etc.*). In attempting to define an FML for ECAs, the possibility of the presence of these new capabilities requires NL engineers, if we are wise, to pay considerable attention to the modality of communicative acts, rather than seeing modality as a relatively uninteresting side-issue. Entering the world of the ECA, we will need to consider different kinds of communicative and non-communicative actions in addition to the traditional Dialogue Act kind of actions, and we will need to be careful how we represent, refer to, categorise, relate, and use these different types of action.

To ground the ideas proposed in the paper, the paper will begin by taking a step back to look at action in general, before considering communicative actions in particular. After this the paper will propose a number of guidelines concerning how one might go about defining a FML for an ECA, which draw on the observations made earlier about action. The tone of the paper will be general, attempting to encompass and consider the predicament of all hypothetical ECAs, rather than focusing on an instantiation of a particular design.

[1] We use the term 'natural language engineers' here loosely to denote researchers (i) who are working towards making robust practical systems to help with real-world situations; (ii) who develop linguistic theories that they realise as computer programs; (iii) who make computational models to understand human thought processes; (iv) who use computer programs to develop and hone linguistic theories.

Cite as: Actions speak louder than words, D. Field, A. Ramsay, *Proc. 8th AAMAS (2009) Workshop 'Towards a Standard Markup Language for Embodied Dialogue Acts'*, Decker, Sichman, Sierra and Castelfranchi (eds.), May, 10–15, 2009, Budapest, Hungary, pp. XXX-XXX.

2. EVENTS, ACTIONS, AND AGENTS

Philosophers have long discussed the nature of action, along with those subjects that are intimately related to it, among them intention, belief, problem solving, intelligence, free will, responsibility, behaviour, and causality, to name a few. 'Action' can be understood in a very general sense as change through time. A hypothetical universe in which time passed but nothing ever changed would be a universe without action. In philosophy, the meaning of 'action' is more specific, and generally denotes an act which is executed by an agent, where an agent is understood to be a person (or other being) who possesses the ability to choose between options, and has the ability to do what he chooses [11]; actions are thus distinguished from other types of events, for example, events which simply happen without being intended or caused by an agent, like when an apple falls from a tree.

With regard to events that are unintentionally caused by agents—for example, causing accidental damage—questions arise as to whether these events are actions or not. This is because, to qualify as an agent, it is generally thought that one must have intention in relation to one's action. There appears to be no consensus on what intention is, but here is one definition from the field of NL engineering to help give the general flavour: intention is "a conduct-controlling pro-attitude. . . intimately related to endeavoring and action" [4, p. 30].

Arguments continue over the nature of intention, over what counts as an internal representation of information, and over determinism. Was my husband exercising his free will when he decided to marry me, or did choice not come into the matter, and he simply responded to chemically and electrically signalled information? Are female thornbug treehoppers who, in spite of the genetic risks from incest, mate with their brothers and thus avoid contact with a particular disease, acting with intention [20]? Is phototaxis[2] in plants action?

3. THE CENTRALITY OF INTENTION

What is notable about discussions in philosophy concerning action is that the central defining characteristic around which all other considerations revolve is *intention.* In contrast, what is notable about discussions in NL engineering about Dialogue Act taxonomies is that speaker intention, though of interest to some, is usually *not* treated as the defining characteristic of Dialogue Acts, that place being taken by the syntax and semantics of linguistic surface form.

Many of us NL engineers are familiar and comfortable with the pragmatics/semantics dichotomy, feeling at home with the idea that sentences outside a context and without utterers nevertheless have meanings (semantic meaning), while sentences inside particular contexts made by particular utterers have extra meanings on top of semantic meaning—pragmatic meanings.[3] We even sometimes tend to view the pragmatic extra meaning as somehow optional, peripheral, not central or essential to the meaning of a sign.

Sentences and even individual words in isolation do not, however, have meanings, rather like the notion that trees falling down in woods containing no observers do not make sounds. Meanings are found in the minds of speakers and hearers, not in sentences, and they depend very much on context ([14, p. 12]). At home, for me, 'cat' usually denotes a type of domesticated house pet, and 'the cat' points to my pet cat, Scrumptious. At work, for my car mechanic, 'cat' usually denotes a car exhaust part, and 'the cat' means 'the catalytic converter that you know I am referring to' (after *ibid* p. 27–28). But although pragmatics ideas have been around for at least 130 years [24, 14, 32, 22, 27], we still seem to be entrenched in the idea that meaning is what you find in language itself, rather than in the minds of its users.

The bias of this attitude suddenly comes more sharply into focus when we start to consider the non-linguistic communicative acts that ECAs might do, or be able to perceive, in their multimodal ways. (In fact, it is not just that ECA's *might* do these acts, it is surely the case that many of them are being designed specifically to be able to make and perceive non-linguistic communicative signs.)

Imagine we have made an ECA that has vision, and that is able to perceive when a person is standing in front it, looking at it, and waving. What is the meaning of this *waving* action? As with words and sentences, its meaning is in the minds of the communicators, not in the waving itself. To know what the waving means to these communicators in this context, we need to know more about this context, including these communicators. Is the person waving performing a communicative act of greeting (waving hello)? Or one of departing (waving goodbye)? Or does she have no communicative intention at all, and is, for example, brushing away a wasp? Or perhaps testing the ECA's camera?[4] To know what a waving hand 'means', and whether it 'means' anything at all, you need contextual information that tells you what it might have been *intended* to mean, if anything.

4. COMMUNICATIVE ACTIONS

Let us begin to think about different modalities of communicative action. To express meaning using linguistic signs, a speaker might, for example, say to a hearer, "My name is George" in order to communicate to the hearer that her name is George. Or she might leave a written note for her 'hearer' ('reader' seems more appropriate) saying "Your dinner is in the dog" in order to communicate dissatisfaction to him. But a person not only has words (and other linguistic signs) at her disposal, but other modalities of a non-linguistic kind.[5] For example, to express dissatisfaction *non*-linguistically, she may poke out her tongue at her 'hearer' (observer, now) using the modality of facial gesture. To

[2] "**phototaxis**: the bodily movement of a motile organism in response to light" (Concise Oxford Dictionary, 10th Edition, 1999).

[3] There is as yet no consensus on precisely how to specify what pragmatics is; different scholars have defined pragmatics in different ways. There is, however, general agreement that it concerns meaning conveyed by the use of language in context, rather than by the words that are actually said. Some respected definitions include [30, p. 729], [18, p. 32], [10, p. 11], [29, Section 1.0].

[4] We are limiting our example interpretations to those appropriate to British culture. No doubt these interpretations are not appropriate for many cultures. We would be wise, of course, to always bear in mind cultural differences when investigating the science of gesture.

[5] There is, of course, well-publicised research that argues that the greater percentage of our effort at communication is non-linguistic, being found in subconsciously manifested facial expression and gesture. In this section we do not taking these into consideration, we consider only action carried out with intention.

draw her observer's attention to some object, she may point to it, using the modality of bodily gesture. Or to communicate gratitude to her observer, she may eat some food that she finds disgusting. Or to communicate total exasperation to her hearer, she may walk out of the room, slamming the door. (These last two examples constitute complex series of non-linguistic actions carried out with an intention to communicate a specific message. We believe that this *series-of-complex-non-linguistic-actions* modality does not yet have a name.)

As we have just seen, the terms 'speaker' and 'hearer' cause confusion when mapped onto the domain of multi-modal communication. Hereon, therefore, we will replace them with the modality-neutral terms 'doer' and 'perceiver', where 'doer' means 'one who performs the action in question' (be it communicative or non-communicative) and 'perceiver' means 'one who perceives the doer in question's action'.

Note that an action may communicate information without the 'doer' intending this, for example, smiling is often done without any communicative intent, and often done without even a perceiver. However, when a smile is seen by a perceiver, it communicates information to the perceiver which may be of interest to him, even though it may well not have been the intention of the doer to communicate information to the perceiver.

Despite the fact that actions like smiling communicate information, we would not classify these as 'communicative actions', because they are not performed with the intention to communicate. We see them more as manifestations of behaviour which happen to be of particular interest to others for some reason. Smiling, and many other behavioural manifestations, are nevertheless of great interest to many who are building dialogue systems, for example, those whose intention is to monitor the emotional state of the user during a dialogue, and respond appropriately to it [33]. Many ECAs, we imagine, will want to be able to recognise user smiles, and to understand what user smiles might tell them about the user's feelings and desires, and therefore smiles as actions (but, we would suggest, not communicative actions) will need to be included in their action taxonomies.

5. ACTION TYPES AND THE EMBODIED CONVERSATIONAL AGENT

Having introduced action and communicative actions in general, here now is a minimal set of different types of action that we believe it would be wise to be able to distinguish between when developing FMLs for ECAs (which will vary, of course, depending on the particular domain and capabilities of the ECAs, to be discussed):

1. Actions that are performed by agents without the intention to achieve them, like accidentally spilling a glass of wine, or accidentally insulting someone by addressing them by their offensive office nick-name. (Some philosophers may argue that the lack of intention renders these events, while others will argue that the presence of an agent who caused each one renders it an action.)

2. Actions that are performed with the intention to achieve non-communicative goals, like connecting a cable to a source of electricity (a goal one might expect a mobile, motile ECA to need to achieve on a regular basis), or moving from one room to another.

3. Facial gestures and expressions that are exhibited with the intention to achieve communicative goals, like poking out one's tongue, and winking.

4. Facial expressions that communicate information about the doer, but not necessarily with intention, like smiling, and frowning.

5. Gestural actions that are performed with the intention to achieve communicative goals, like pointing, and shrugging one's shoulders.

6. Complex series of actions (non-gestural) that are performed with the intention to achieve communicative goals, like eating food that one finds disgusting in order to communicate gratitude, or buying and giving an expensive gift to communicate appreciation.

7. Linguistic actions that are performed with the intention to achieve communicative goals, like a written Conventional-Opening (DAMSL) "Hi John, how are you?" (very common in online chat discourse), or a spoken *Indirect Propositional Question* (DIT++) "Do you know if we have any carrots left?", or a signed head nod or shake in American Sign Language [19].[6]

8. Linguistic actions that are performed without the intention to communicate, like talking to a teddy bear

6. CONSTRAINTS TO HELP IN THE DEFINITION OF AN FML FOR ECAS

This is all very well, you say, but my ECA has no limbs with which it can gesture, and so your list is not relevant. Our response is that your objection brings us straight to the first of a handful of conditions that we believe it would be wise to impose on any set of actions that one is attempting to define for an ECA.

Just before that, we would like at this point to introduce the term 'Multi-modal Communicative Action' (MCA) to denote all actions performed with the intention to communicate, regardless of whether they are linguistic or not. This set would include 3, 5, 6 and 7 above. We would also like to introduce the term 'Non-Communicative Action' (NCA). By 'Dialogue Act' we denote only the set 7 above.

6.1 The set of system MCAs should be partly determined by the nature of the embodiment and capabilities of the ECA

A question that [28] raises concerning how to define a set of Dialogue Acts (not MCAs) is, "Can the same taxonomy be used for different kinds of agents?" (p. 19). With respect to MCAs we would propose that the taxonomy of MCAs should be defined with very close attention to the domain and behaviour of the ECA. ECAs are an emerging

[6] We include sign language under 7 on several grounds, despite the fact that the principal modality of communication appears similar to gesture and facial expression. Papers in collaboration with Ronnie Wilbur (Speech, Language and Hearing Sciences, Purdue University, IN, USA) are forthcoming on this topic, of which space precludes discussion in this paper.

technology, and they will doubtless take many forms. We might choose, for example, to embody our currently non-embodied spoken conversational agent in a plastic, roughly human-shaped, human-sized bust that has no capabilities over and above the program that we have lodged inside it (similar to Autom[7] [17]). Such an ECA, though embodied, would not be able to perceive or produce any MCAs that were not spoken linguistic actions, and there would be no need to consider MCAs apart from Dialogue Acts. In contrast, imagine a similar ECA but with the addition of vision. Now it could theoretically perceive all eight types of action, if it possessed strategies for being able to recognise and label them. Such an ECA would require a taxonomy of MCAs and NCAs, but only with regard to perceiving them in the user, not producing them, which brings us to our next point.

6.2 ECA MCAs should be considered separately from user MCAs, and production MCAs should be considered separately from perception MCAs

In any ECA–User pairing, the capabilities and behaviours of each party are likely to be very different from each other's (unless, of course, both parties are ECAs, and they are the same kind of ECA, which for now we will assume is likely to be a relatively rare scenario). It is, then, eminently sensible to bear in mind the difference in system and user capabilities when defining a set of MCAs. The system's capabilities have a great bearing on which MCAs are appropriate for its own perception and production. If the system is an Autom-like, plastic, human-shaped shell, with a spoken conversational agent lodged inside it, and with the single added perceptive modality of vision, its own *actions* will necessarily be restricted to type 7 above,[8] while it may have strategies for *perceiving* and responding appropriately to user facial expressions, even though it cannot produce facial expressions itself. So while the ECA needs to be able to recognise and label user facial-expression MCAs, and know how to respond to them in speech, there is no point in it considering them when deliberating over its own behaviour.

This may seem like a somewhat obvious point to make, but a tendency to think of MCAs as one homogenous set is likely to be not uncommon among NL engineers and linguists who study Dialogue Acts in the context of human-human dialogue corpora.

6.3 The set of MCAs should be partly determined by the nature of the activity

Another question that [28] raises is "Can the same taxonomies be used for different kinds of activities? (p. 17). Our response to this with respect to ECAs is, 'Yes, it can, but it might not work very well'. A better approach, we would suggest, is that the nature of the activity of the ECA be key to determining the set of MCAs. Imagine we are building an ECA whose main function is to take care of a dog by talking to it, soothing it, playing with it, feeding it, letting it into the garden, and generally being a human-like companion to it.[9] A good ECA for this domain would be one that can interact with a dog in a human-like way, and this would require it to be able to recognise signs of significant doggy behaviour like tail wagging, barking, licking, and whining, including communicatively less interesting behaviours such as sleeping, eating, and sulking. Let us assume that the ECA has vision and hearing (as well as speech, controllable moving forelimbs and digits, and lateral mobility) and is able detect when each of these doggy behaviours is being exhibited by its doggy user—and let us call it Cruella.

Let us also imagine that there is a second mechanically and technically identical ECA whose main function is to entertain a child by talking to it, soothing it, playing with it, feeding it, and so on, and let us call this ECA Nanny. Despite the identicality of the design of their hardware and capabilities, an appropriate taxonomy of MCAs for Cruella will be a very different taxonomy from an appropriate one for Nanny, because dogs' behaviour and needs are very different from children's.

6.4 The set of user MCAs should include only those that are discernible by the ECA, and the set of system MCAs should include only those that are performable by the ECA

It may sound like a rather obvious point to make to say that user MCAs should be discernible by the ECA, but this point may resonate strongly with some NL engineers. From its earliest foundations, the discipline that led to the emergence of taxonomies of Dialogue Acts defined speaker acts in terms of conditions that cannot be perceived by another being—because they are thoughts that remain private to the individual. If George's husband says to her "I'll clean the bathroom for you before they arrive", Searle [26, p. 49] tells George that this is an act of 'promising' if her husband *intends* to clean the bathroom—and only if eight other conditions are also met, most of which concern the private, unobservable thoughts of the speaker—but it is *not* an act of 'promising' if he does not intend to do the predicated act.

How does George know whether her husband intends to do this predicated act or not? She cannot observe his intentions, so she cannot know, even if he tells her explicitly that it his intention to do it. She might be able to have a good guess at whether he will do what he has said he will do, based on his previous performance. But this is a quite different problem from thinking that she has to recognise the speech act of 'promising' in order to work out what this utterance means, and how to respond to it.

The act of 'informing' raises a similar problem but from the perspective of the doer rather than the recipient. If, for example, an effect of the act of 'informing' is that the perceiver comes to believe that what was expressed is true (a popular idea since [8] and [2]), it is very difficult for you or an ECA to plan to perform and 'inform', because neither of you can ensure your hearers will believe what you tell them if they do not want to believe it, even if you torture them.

[7] `http://robotic.media.mit.edu/projects/robots/autom/overview/overview.html`

[8] Although one may conceive of this ECA as *seeing* things, we would argue in the context of this paper that its seeing is not an action, on the grounds that it is a passive receiving of visual data, not an action executed with intention.

[9] You may consider that such a robot would not, by definition, be an ECA, on the grounds that dogs do not make conversation. We would argue that dogs do communicate their desires and feelings to people through communicative acts, and that they attach meaning to many spoken words. They also hugely enjoy human company and communicative interaction, and so we feel such a theoretical robot is hypothetically admissable as an ECA.

This point is not just a theoretical one, but has repercussions if we want to build systems that try to work out how to converse on the basis of what the system believes are the user's beliefs (which change at every conversational turn). We need to bear in mind that neither we nor our ECAs can plant beliefs into the minds of our perceivers, and so we need to take into account, among many other epistemic possibilities, the possibility that perceivers may not believe what doers say, and that a doer may himself privately not believe his own utterance, and be attempting to deceive.[10]

You may think that the need for an ECA to have to worry whether a user might be attempting to deceive it is far-fetched. But there are clear cases where an ECA might be a desirable tool where deception might also not be unexpected. An example is health-oriented conversational systems that try to encourage people to eat better, or take more exercise, or stop smoking, say, in which the ECA acts as a kind of health expert who encourages better habits in the user. Health professionals report that compliance with their advice in such situations can often be poor [3], and users can often be dishonest in reporting their activities [23]. Any ECA attempting to take the place of the health professional will need to be prepared for this.

Of course, we need approaches to the problem of not being able to read the thoughts of our interlocutors that enable us to proceed in spite of this difficulty. One approach we have previously discussed is to assume that in a purely neutral context where neither party has any specific views on the reliability or cooperativeness of the other, it is nonetheless rational for a doer to produce utterances that he believes and for a perceiver to believe that this is what he is doing (see [25] for detailed discussion of the grounds and implementation).

6.5 The set of user MCAs should include only those that add strategic value for the ECA

A related point to our last point is this one: there is no gain in having acts in your taxonomy that do not increase the ECA's strategic capabilities. The whole point of recognising and labelling acts at all, from the engineer's point of view, is so that the system can recognise and categorise user behaviour, and use its analysis to work out how to act itself. The process of recognition of acts should buy the ECA information that it does not already know, and cannot more easily get from elsewhere, and that will help it to determine its own actions. There is no point in the system putting labels on phenomena at run-time if the labels have no use.

We could, for example, put elegant taxonomies of varies kinds into our ECAs that don't achieve anything from the ECA's point of view, and that would not result in any degradation if they were removed. An example would be any theory of action based on STRIPS operators [13] in which identifying the operator brought no insight to the system that it had not already had before the action was performed (see [6] and [25] for discussion of how certain approaches to implementing speech acts can fall into this trap).

7. CONCLUSION

This paper has attempted to use insights from AI planning and pragmatics to give direction to how one might begin to define suitable sets of multi-modal communicative acts for embodied conversational agents. In response to the tone of the call for papers for this workshop, the aim of this paper has been to be thought-provoking, to stir up debate, and to hasten practical progress with implementations of ECAs.

7.1 Acknowledgments

This work was funded by *Companions*[31], European Commission Sixth Framework Programme Information Society Technologies Integrated Project IST-34434.[11]

8. REFERENCES

[1] J. Allen, J. Hendler, and A. Tate. Editors. *Readings in planning*, 1990. San Mateo, California: Morgan Kaufmann.

[2] J. F. Allen and C. R. Perrault. Analyzing intention in utterances, 1980. *Artificial Intelligence* 15: 143–78. Reprinted in [16], pp. 441–58.

[3] M. H. Becker and L. A. Maiman. Sociobehavioral determinants of compliance with health and medical care recommendations, 1975. Medical Care **13**(1):10–24.

[4] M. E. Bratman. What is intention?, 1990. In [7], pp. 15–31.

[5] H. Bunt. Dynamic interpretation and dialogue theory, 2000. In: M. M. Taylor, D. G. Bouwhuis and F. Neel (eds.) *The Structure of Multimodal Dialogue*, Vol. 2, Amsterdam: John Benjamins, 2000, pp. 139–166.

[6] P. R. Cohen and H. J. Levesque. Rational interaction as the basis for communication, 1990. In [7], pp. 221–55.

[7] P. R. Cohen, J. Morgan, and M. E. Pollack. Editors. *Intentions in communication*, 1990. Cambridge, Massachusetts: MIT.

[8] P. R. Cohen and C. R. Perrault. Elements of a plan-based theory of speech acts, 1979. *Cognitive Science* 3: 177–212. Reprinted in [16], pp. 423–40.

[9] M. Core and J. Allen. Coding dialogs with the DAMSL annotation scheme, 1997. In *Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, pp. 28–35, Cambridge, MA.

[10] S. Davis. Editor. *Pragmatics: A reader*, 1991. Oxford: Oxford University Press.

[11] A. D. Donagan. *Choice: The essential element in human action*, 1987. London: Routledge.

[12] D. Field and A. Ramsay. Sarcasm, deception, and stating the obvious: Planning dialogue without speech acts, 2004. *Artificial Intelligence Review* **22**(2): 149–171.

[13] R. E. Fikes and N. J. Nilsson. STRIPS: A new approach to the application of theorem proving to problem solving, 1971. *Artificial Intelligence* 2: 189–208. Reprinted in [1], pp. 88–97.

[14] G. Frege. Der gedanke. Eine logische Untersuchung, 1918. In *Beiträge zur Philosophie des deutschen Idealismus, I*, pp. 58–77. Translated by A. M. and Marcelle Quinton as 'The thought: A logical inquiry',

[10] See [12] for discussion of formal additions to the maxims of Grice's Cooperative Principle [15] to allow for the fact that people often do not abide by the Gricean maxims, but violate them.

[11] http://www.companions-project.org/

Mind 65, 1956, pp. 289–311. Reprinted in [21], pp. 9–30.

[15] H. P. Grice. Logic and conversation, 1975. In P. Cole and J. Morgan (eds) *Syntax and semantics, Vol. 3: Speech acts*, pp. 41–58. New York: Academic Press.

[16] B. J. Grosz, K. S. Jones, and B. L. Webber. Editors. *Readings in natural language processing*, 1986. Los Altos, California: Morgan Kauffmann.

[17] C. Kidd and C. Breazeal. Designing a sociable robot system for weight maintenance, 2006. In *Proceedings of IEEE Consumer Communications and Networking Conference (CCNC-06)*. Las Vegas, NV, Vol. 1, pp. 253–257.

[18] S. C. Levinson. *Pragmatics*, 1983. Cambridge textbooks in linguistics. Cambridge: Cambridge University Press.

[19] S. K. Liddell. *American Sign Language Syntax*, 1980. Mouton Publishers, 1980.

[20] P. A. D. Luca and R. B. Cocroft. The effects of age and relatedness on mating patterns in thornbug treehoppers: inbreeding avoidance or inbreeding tolerance?, 2008. *Behavioral Ecology and Sociobiology* Vol. 62, No. 12, pp. 1869-1875.

[21] P. Ludlow. Editor. *Readings in the philosophy of language*, 1997. Cambridge, Massachusetts: MIT.

[22] C. Morris. Foundations of the theory of signs, 1938. In O. Neurath, R. Carnap and C. Morris. Editors. *Foundations of the unity of science: toward an international encyclopedia of unified science*, Vol. 1, pp. 77–137. London and Chicago: University of Chicago Press.

[23] L. S. Muhlheim, D. B. Allison, S. Heshka, and S. B. Heymsfield. Do unsuccessful dieters intentionally underreport food intake?, 1998. *International Journal of Eating Disorders*, Vol. 24, Issue 3, pp. 259–266.

[24] C. S. Peirce. How to make our ideas clear, 1878. In *Popular Science Monthly* 12: 286–302. Reprinted in N. Houser and C. Kloesel. Editors. *The essential Peirce: Selected philosophical writings, Vol. 1 (1867–1893)*, 1992, pp. 124–41. Bloomington and Indianapolis: Indiana University Press.

[25] A. Ramsay and D. Field. Speech acts, epistemic planning and Grice's maxims, 2008. *Journal of Logic and Computation* **18**: 431–457.

[26] J. R. Searle. What is a speech act?, 1965. In M. Black. Editor. *Philosophy in America*, pp. 221–39. Allen and Unwin. Reprinted in J. R. Searle. Editor. *The philosophy of language*, 1991, pp. 39–53. Oxford: Oxford University Press.

[27] P. Strawson. On referring, 1950. *Mind* **59**: 320–44. Reprinted in [21], pp. 335–59.

[28] D. Traum. Questions for dialogue act taxonomies, 2000. *Journal of Semantics* **17**(1):7–30.

[29] J. Verschueren, J.-O. Östman, and J. Blommaert. Editors. *Handbook of pragmatics: Manual*, 1995. Amsterdam: Benjamins.

[30] A. Weiser. Deliberate ambiguity, 1974. In *Papers from the 10th Regional Meeting of the Chicago Linguistics Society*: 723-731.

[31] Y. Wilks. Artificial companions, 2005. *Interdisciplinary Science Reviews*, June, Vol. 30, pp. 145–152.

[32] L. Wittgenstein. Some remarks on logical form, 1958. Exerpt from *The blue and brown books*, Oxford: Basil Blackwell. Reprinted in [21], pp. 31–47.

[33] S. Worgan and R. Moore. Enabling reinforcement learning for open dialogue systems through speech stress detection, 2008. In *Proceedings of the Fourth International Workshop on Human-Computer Conversation*, Bellagio, Italy.

Functional Description of Multimodal Acts: A Proposal

Kristinn R. Thórisson
CADIA / School of Computer Science
Reykjavik University
Kringlunni 1, 103 Reykjavik, Iceland
thorisson@ru.is

Hannes H. Vilhjalmsson
CADIA / School of Computer Science
Reykjavik University
Kringlunni 1, 103 Reykjavik, Iceland
hannes@ru.is

ABSTRACT

Architectures for controlling communicative humanoids have been many and varied. Planning systems for multimodal behavior still require significant efforts to design and implement; this could be alleviated to some extent through the use of a common platform. In this paper we outline an approach to multimodal action generation following the SAIBA framework. The proposal focuses on planning at a medium-level of action abstraction – what we refer to as a *functional* level – building on our prior efforts in creating systems capable of human-like multimodal behavior. Starting from a high-level initial goal or releasing mechanism we assume that surface behavior can be generated in continuous incremental steps at multiple levels of abstraction, as outlined by SAIBA, progressing over time in a depth-first manner, towards an actual executed behavior. The paper proposes a starting point for a language intended to represent/describe functional aspects of communicative action. We argue that among key aspects that must be addressed for this to be successful are temporal constraints, prioritization and classification of function.

Categories and Subject Descriptors

D.3.0 [**Programming Languages**]: General – *Standards.* D.3.2 [**Programming Languages**]: Language Classifications – *Very high-level languages.*

General Terms

Design, Standardization, Languages, Theory

Keywords

Multimodal Communication, Realtime, Communicative Humanoids, Functional Markup Language, SAIBA, Multimodal Acts, Embodied Agents.

1. INTRODUCTION

The SAIBA framework [1] is motivated by a need to enable collaboration in building communicative humanoids. Second, it is motivated by push towards more sophisticated multimodal communicative planning, and third, by a hope for easier construction of multimodal skills for multimodal characters, whether physical robots or virtual. Towards this end, SAIBA proposes a modular approach to the "planning pipeline"[1] for realtime multimodal behavior. There are at least two important modular splits in this respect. The first is between a representation *language that describes* an action/set of actions and *the engine/mechanism that realizes* these, according to a specification written in this language. Another split – or set of splits – proposed by SAIBA is between lowest-level behaviors ("animation level"), a medium-level representation typically called "behavior" level, and a higher level called the "functional" level. These levels correspond roughly to what have sometimes been called the *primitive/servo level*, *e-move level* and *task level*, respectively, in the robotics community [2].

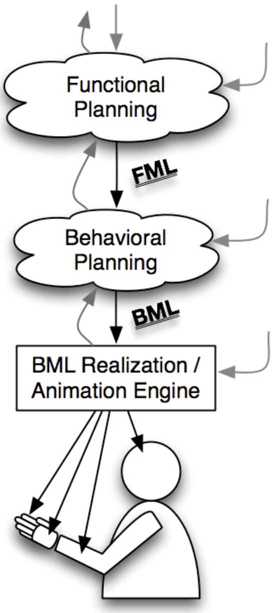

Figure 1: The planning levels envisioned by the SAIBA framework, showing where FML and BML fit in. Upwards gray arrows indicate feedback; gray side arrows imply that other input to the planning mechanisms could come from elsewhere in the system.

At the first split, SAIBA proposes the Behavior Markup Language (BML) as the representation language for describing human movement at the level of behaviors [1][3]. BML can thus serve as

[1] As we assume ample feedback loops in this system via perceptual mechanisms it is, strictly speaking, not a pipeline. This is an important point that often is overlooked. However, the pipeline model works reasonably as a first approximation.

the input to a basic animation engine. An example of an engine that can realize BML as realtime-executed multimodal actions is SmartBody [4], which we have incorporated into the relatively simple-to-use CADIA BML Realizer[2].

The idea behind the split between representation language and realization engine is to enable those researchers who desire to focus on a particular level of planning to stick to a certain level of detail. The language describing the desired outcome at a particular detail level can be represented in a common way between researchers, making easier the collaboration on – and competition between – proposed mechanisms. This allows construction of alternative planning mechanisms at particular levels of abstraction, and thus exploration of different ways of producing certain behavior phenomena, without having to solve the mechanism for the whole field, as the representational languages provide an API that allows modular sharing of solutions for different parts of the architecture. This also enables the comparison between realization mechanisms from different research labs. Further benefits to such a modular scheme are discussed in [1][5][6].

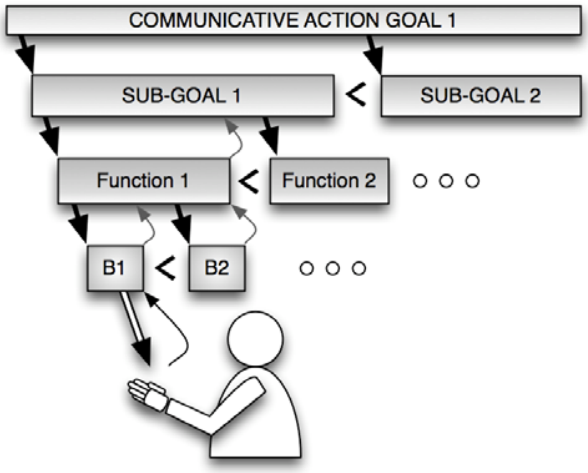

Figure 2: Successive refinement of goals (spanning long durations of time) into functional descriptions (fat arrows) and then into behaviors that can be executed (hollow arrow) as actual movements. Only sequential dependencies between actions are depicted here. Feedback about the actual implementation of each chunk is provided to the next level up (narrow bent arrows).

Looking at things from a descriptive, static perspective, BML describes human multimodal behavior at a particular level of abstraction. For runtime systems BML provides a "human-readable description level", which is assumed to be the output of a behavior-level engine. Several prototypes of such engines exist in our research community. The *input* to such an engine, however, has not been specified in the context of the SAIBA framework/consortium. SAIBA proposes that the input to this (abstract) engine should be in the form of a language that represents functional aspects of the movements. The idea here is that this could be captured in a Functional Markup Language, FML [7].

In this paper we will not discuss the mechanisms that produce FML automatically – this will be the topic of a future paper – here we will focus on the design of FML. We propose an outline of what FML could contain, and present a starting point for its creation.

Although BML and FML will very likely share ideas, especially related to temporal issues and synchrony, a number of issues will necessarily be different between these two languages. In particular, time and temporal dependencies in FML will certainly be represented in a coarser way than in BML (just as BML represents time more coarse-grained than the final (frame-based) animation level), as "plans" at this level span larger chunks of time and can thus be seen as providing a "rough outline" or specification for the next planning level below. This process of iterative/depth-first[3] construction, as proposed by the SAIBA framework, is represented in Figure 2.

A few words are required to clarify background assumptions. First, we look at dialogue as a continuous, realtime process, turntaking being a case in point, which requires dynamically negotiated role acceptance including who has turn, how multimodal "signals" are interpreted and used, how willing the parties are in trying to understand each other, go along with premises put forth, etc. [7]. (Most often such negotiation is implicit and goes unnoticed by dialogue participants.) We also assume agent-orientation: A participant is at any point in time in charge of *only his own end* of such a realtime activity, but of course in all except the most extreme cases has an ability to affect the other party in many ways by his own behavior.

2. WHAT IS A FUNCTION?

In the present context, by "function" we mean *the effect that an action is intended to have* in a particular multimodal communicative interaction, either on the body of the actor him/herself and/or on the mind/body of the interlocutor(s). What is typically expressed here are inner states such as affect or agreement, and those related to management of the interaction itself, including the exchange of turns. The realization of these functions relies on the coordination of a wide range of behaviors including prosody, vocal fillers, head motion, body posture and eye gaze, all of which are specified further at the behavior-level.

3. FML: AN OUTLINE

FML must describe the effect that an intended action or plan should have on the environment, most obviously the agent itself, that must express that function; in line with the SAIBA framework it leaves out, however, morphological considerations that are intended to be composed at runtime by one or more (mostly as-of-yet unspecified) engines/mechanisms. As with BML this is the research part: The language does not specify *how* an

[2] http://cadia.ru.is/projects/bmlr/

[3] It should be noted that SAIBA does not specify that planning needs to proceed in a top-down manner; it is well conceivable that higher levels take proposed BML as input and generate FML as a way to make sure that the behavior to be performed does not work against the agent's goals at that point in time (example: I really want to scratch my head but my boss has told me not to). The same could be done for producing a hypothesis for what another person's behavior means at the functional level; in this case the BML is used to describe the surface form of the actions observed; the FML would represent hypotheses for what it is intended to achieve.

FML specification is turned into actual expressed behavior, whether through BML or some other way.

We propose that FML be based on labels which refer to basic functions of multimodal communicative actions/goals, and constraints on those functions, which themselves can be divided into functional categories. An example is the communicative function to express moderate happiness: The function *express_happiness* has a constraint describing its amount, say "medium" or “0.5”.

There are several things that need to be taken into account for FML to be successful. First, we must provision for coarse-grained temporal constraints. Second, we need a prioritization scheme so that an FML Engine can be given instructions as to how to solve conflicting functions. Such a scheme could be represented as any other type of constraints on functions. Third, we need to classify functions in to groups that help a human designer use FML; Thórisson's scheme in the Ymir/Gandalf system [8] of splitting them into *Topic Functions* and *Envelope Functions* provides an example of a step in this direction.

3.1 Temporal Constraints

Temporal constraints at the functional level of description tend to be much more coarse-grain than those at the lower levels termed “behavior” and “execution”. At the execution level one has to deal with frames and milliseconds; at the intermediate behavior level we deal with temporal relationships of bits of multimodal events such as e.g. gaze, grasps and body stance; at the functional level we specify temporal relationship between what we could call “plan chunks”. These chunks refer to parts of a plan that implements the form for e.g. a set of inter-related propositions to be expressed, for instance directions on how to get from one city to another. Each such plan chunk will typically consist of several multimodal acts at the behavioral level. To take an example, in a full plan consisting of several chunks intended to help someone decide where to take a walk in the forest, pointing at a map and saying “You start here [deictic gesture; looking at map], and walk all the way through the forest [tracing with finger], and end up here [finger stops], will take you approximately 1 hour [gazing back at interlocutor]” would be one plan chunk containing (roughly) three BML chunks.

To specify temporal relationships between functional plan chunks we propose to start with simple synchronization primitives such as:

must_end_before(a,b,T)
execute_anytime_during(a,b,T)
start_immediately_after(a,b)
start_sometime_after(a,b,T)
start_together(a,b,...z,T)
end_together(a,b,...z,T)

These are relatively self-describing; *a* and *b* are plan chunks whose relation is described with the primitive, where *b* is the reference; *T* is an optional parameter that describes a maximum boundary or tolerance which can be provided by the designer or even computed dynamically, based on context.

3.2 Prioritization Scheme

In the Ymir/Gandalf system Thórisson [8] proposed three main levels of prioritization: *Reactive*, *Process Control* and *Topic*, each one of a lower priority than the prior, respectively. If a reactive behavior is required while a behavior of a different priority is executing, the reactive behavior takes precedence. If a process control-level behavior is requested while a topic-level plan is being executed the latter one will have to yield. This prioritization scheme has been proposed as a cognitive theory of human dialogue organization [7]. More importantly for the present discussion, a key goal of such a prioritization scheme is to enable a designer to stop worrying, to some extent, about unwanted interactions between conflicting behaviors.

Generally speaking, BML maps to a reactive level while FML to the process control level (and partly the Topic level). Compared to these priority levels in Ymir architecture, however, this mapping is not 1:1 because Ymir separated a Behavior Lexicon from perception-driven decision/planning mechanisms while SAIBA proposes a different split, as already described in the Introduction. Nonetheless, the comparison can provide a rough sketch for prioritization scheme in FML.

3.3 Classification

The classification of behavioral functions will aid the designers of multimodal dialog systems at different levels. At the highest level, the designer will see a rough outline of the human communicative capacity of a system by noting what general kinds of function specification are available. At a much lower level, a designer can expect that functions within a certain category will share some specification characteristics (such as types of constraints) or share a representation of common plan chunks or structures (such as turns or participants).

Choosing a classification scheme that embraces all prevailing perspectives on communicative function is not easy, but we have to start somewhere. The research community has more or less come to an agreement about the existence of a category of communicative functions that serve to *coordinate* a multimodal dialog. The functions in this category have been called envelope, interactional or management functions [8][10][11]. Examples gathered from a range of multimodal dialog projects are shown in Table 1a (these tables are replicated from [10]).

Table 1a: ENVELOPE/INTERACTION FUNCTIONS

Function Category	Example Functions
Initiation / Closing	*react, recognize, initiate, salute-distant, salute-close, break-away*
Turntaking	*take-turn, want-turn, give-turn, hold-turn*
Speech-Act	*inform, ask, request*
Grounding	*request-ack, ack, repair, cancel*

Another category covers the actual content that gets exchanged during a dialog. Given that the envelope functions are doing a proper job in an ongoing dialogue, topic functions have a better chance of being achieved. Typically this is the deliberate exchange of information, which gets organized and packaged in information chunks[4] that facilitate uptake/interpretation in an interlocutor. Another set of function examples have been gathered for Table 1b.

Table 1b: TOPIC/CONTENT FUNCTIONS

Function Category	Example Functions
Discourse Structure	*topic, segment*
Rhetorical Structure	*elaborate, summarize, clarify, contrast*
Information Structure	*rheme, theme, given, new*
Propositions	*any formal notation (e.g. "own(A,B)")*

It has been suggested that a useful distinction could be made between functions that carry deliberate intent and those that merely give off behavior involuntarily [5]. Examples of such functions are shown in Table 1c.

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*

This classification, along with each of the named functions, is a proposal for actual FML tags, which can be discussed further at this workshop.

4. CONCLUSIONS

SAIBA is an important effort in coordinating and advancing research on multimodal behavior generation and the specification of FML is a key element. For FML to be successful, three things in particular have to be taken into account: (1) Temporal constraints at a coarser level of granularity than the BML level; (2) A prioritization scheme that helps arbitrate conflicts and supports reactivity; (3) Classification of FML tags into categories that help human designers make sense of communication capabilities as well as for identifying groups of functions with similar parameterization. These notes should serve as seeds for a discussion that is highly relevant to the kinds of real-time multimodal dialog systems being built at CADIA.

REFERENCES

[1] Kopp, S., B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. R. Thórisson, H. Vilhjálmsson: Towards a Common Framework for Multimodal Generation: The Behavior Markup Language. *Proceedings of Intelligent Virtual Agents (IVA '06)*, August 21-23. Also published in Springer Lecture Notes in Computer Science (2006)

[2] Herman, M. and Albus, J.: REAL-TIME HIERARCHICAL PLANNING FOR MULTIPLE MOBILE ROBOTS. In *Proc. DARPA Knowledge-Based Planning Workshop*, Austin , Texas, December 1987, 22-1 – 22-10 (1987)

[3] Vilhjalmsson, H., Cantelmo, N., Cassell, J. et al.: The Behavior Markup Language: Recent Developments and Challenges. *Proceedings of Intelligent Virtual Agents (IVA'07)*, Vol. LNAI 4722. Springer (2007) 99-111

[4] Thiebaux, M., Marshall, A., Marsella, S., and Kallmann, M., SmartBody: Behavior Realization for Embodied Conversational Agents, in *Proceedings of Autonomous Agents and Multi-Agent Systems* (AAMAS) (2008)

[5] Heylen, D., Kopp, S., Marsella, S., Pelachaud, C., Vilhjalmsson, H.: The Next Step Towards a Functional Markup Language. *Proceedings of Intelligent Virtual Agents (IVA'08)*, Springer (2008)

[6] 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)

[7] Thórisson, K. R. Natural Turn-Taking Needs No Manual: Computational Theory and Model, from Perception to Action. In B. Granström, D. House, I. Karlsson (Eds.), *Multimodality in Language and Speech Systems*, 173-207. Dordrecht, The Netherlands: Kluwer Academic Publishers. (2002)

[8] Thórisson, K. R. Gandalf: An Embodied Humanoid Capable of Real-Time Multimodal Dialogue with People. *First ACM International Conference on Autonomous Agents*, Mariott Hotel, Marina del Rey, California, February 5-8, (1997) 536-7

[9] Vilhjalmsson, H. and Thórisson, K.R. A Brief History of Function Representation from Gandalf to SAIBA, in the *proceedings of the 1st Function Markup Language Workshop at AAMAS*, Portugal, June 12-16, (2008)

[10] Vilhjalmsson, H. Representing Communicative Function and Behavior in Multimodal Communication, A. Esposito et al. (eds.), *Multimodal Signals: Cognitive and Algorithmic Issues, Lecture Notes in Artificial Intelligence*, Vol. 5398. Springer (2009)

[11] Thórisson, K. R. and Jonsdottir, G. R.: A Granular Architecture for Dynamic Realtime Dialogue. *Proceedings of Intelligent Virtual Agents (IVA'08)*, Tokyo, Japan, September 1-3 (2008)

[4] In other work we have used the concept of "thought unit" as the smallest unit that is ready be turned into the equivalent of a BML-specified behavior [11].

List of authors