Adaptive Expressiveness – Virtual Conversational Agents That Can Align to Their Interaction Partner

Hendrik Buschmeier, Kirsten Bergmann and Stefan Kopp
Sociable Agents Group, CITEC, Bielefeld University
PO-Box 10 01 31, 33501 Bielefeld, Germany
{hbuschme, kbergman, skopp}@techfak.uni-bielefeld.de

ABSTRACT

Speakers in dialogue tend to adapt to each other by starting to use similar lexical items, syntactic structures, or gestures. This behaviour, called *alignment*, may serve important cognitive, communicative and social functions (such as speech facilitation, grounding and rapport). Our aim is to enable and study the effects of these subtle aspects of communication in virtual conversational agents. Building upon a model for autonomous speech and gesture generation, we describe an approach to make the agent's multimodal behaviour adaptive in an interactive manner. This includes (1) an activation-based microplanner that makes linguistic choices based on lexical and syntactic priming, and (2) an empirically grounded gesture generation such that linguistic priming parallels concordant gestural adaptation. First results show that the agent aligns to its interaction partners by picking up their syntactic structures and lexical items in its subsequent utterances. These changes in the agent's verbal behaviour also have a direct influence on gestural expressions.

Categories and Subject Descriptors

I.2.0 [**Artificial Intelligence**]: General—*Cognitive simulation*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Language Generation*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Natural Language*

General Terms

Design, Experimentation, Human Factors

Keywords

Verbal and non-verbal expressiveness, Modelling natural language, Multimodal Interaction, User-adaptated interaction, Interactive alignment

1. INTRODUCTION

Alignment of interlocutors is a ubiquitous and much described phenomenon in human interaction. When speaking in dialogue, for instance, speakers and listeners rapidly begin to converge on the same vocabulary [7], they tend to use similar syntactic structures [5], they adapt the rate and other features of their speech to one another [12] and they mimic the other's gestures and body posture [15]. That said, they align much of their behaviour.

Causes and functions of alignment are manifold – and an issue of hot debates. We base our work on Pickering and Garrod's interactive alignment model [20] which assumes alignment to be an automatic process, driven mainly by implicit priming of underlying (linguistic) representations. The focus of this approach is primarily on the cognitive function of alignment, i.e., how it facilitates language processing and enables fluid production. Another aspect often stressed, is the communicative function of alignment where understanding becomes easier due to the growing shared vocabulary and beliefs of speaker and listener [10]. Finally, there is also a social function. Interlocutors sharing a vocabulary feel connected, as they 'speak the same language'. In this way alignment also facilitates rapport and effects (and is affected by) a mutual positive evaluation between interaction partners.

It stands to reason that these functions make alignment potentially beneficial for human-agent interaction. The cognitive function might allow more efficient language generation and understanding on the side of the agent and might thus enable fast turn-taking and deep embodied feedback. The communicative function relates to and supports the key notion of grounding and is considered to be at the core of successful verbal exchange. Increasingly getting into the focus of researchers in human-agent interaction is the social function of alignment. Here, recent work has demonstrated that such qualities of an agent can make users feel better understood, relieve them from social anxiety [14], or make them rate the agent more persuasive and positive [1].

We have started to model alignment capabilities in virtual conversational agents for two main reasons. First, we want to unravel and understand the mechanisms that are at work when agents are engaged in social interaction. Second, we want to study the effects and affordances of such adaptations on the interaction between human users and artificial agents.

In the present work, we make first steps towards virtual agents that can align interactively to their interaction partners in their verbal and gestural behaviour. After reviewing related work (Section 2), we present empirical evidence for adaptive expressiveness by giving an overview of findings from the literature and results of a corpus analysis investigating the connection between syntactic constructions and gestural representation technique (Section 3). Following this,

Cite as: Adaptive expressiveness – Virtual conversational agents that can align to their interaction partner, Buschmeier H., Bergmann, K. and Kopp, S., *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, van der Hoek, Kaminka, Lespérance, Luck and Sen (eds.), May, 10–14, 2010, Toronto, Canada, pp. 91-98

we describe our priming/activation-based model of language microplanning and our architecture for autonomous speech and gesture generation (Section 4; [8, 4]). In Section 5 we present a novel integration of both modules, which results in linguistic alignment paralleled by concordant adaptation of gestural behaviour. Furthermore, we report on results of produced agent behaviour that aligns to user input. In Section 6 we conclude by discussing the current model and laying out further extensions

2. RELATED WORK

Several computational models of linguistic alignment have been created over the last years. Isard et al. [13], for example, present a 'massive over-generation' approach to modelling alignment and individuality in natural language generation. Their system generates a huge number of alternative sentences and evaluates each one with a trigram model consisting of a default language model computed from a large corpus and a cache model derived from the user's utterance. The cache model represents the aligned structures and is combined with the default model. Another approach is based on the Dynamic Syntax formalism, which uses the same representations and mechanisms for parsing and generation of natural language [21]. The implementation extends the formalism with a model of context consisting of two distinct representations: a record of the semantic trees generated and parsed so far and a record of the transformation actions used for the construction of these semantic trees. Alignment is then created through *re-use* of trees and actions. Both systems have not yet been employed in dialogue systems or virtual agents and, therefore, do not lend themselves for studying the effects of machine alignment on human users. In contrast, the following two systems have both been evaluated in interactive settings.

The SPaRKy text generation system [26] also uses an over-generation and rank approach to adapt surface utterance and text structure to the users' individualities. It generates text snippets according to a single user's personal preference, which is collected prior to system usage by ranking a large number of sentences. Unfortunately, its approach to user-adaptation as well as its time-consuming generation method make SPaRKy unusable for on-line interaction with virtual agents. Finally, de Jong et al. [11] present a virtual guide that is adaptive to its users' levels of politeness and formality. The system analyses several features of a user's utterance and generates a reply in the same register. Lexical and syntactic alignment is said to occur automatically because the lexical items and syntactic constructions to choose from are constrained by the linguistic style adopted. Both systems do not model alignment in detail since they focus on general adaptation to the user. Furthermore, although implemented in an embodied virtual agent, de Jong et al. did not extend the effects of politeness, formality or alignment to the agent's overall behaviour including gestures.

Work on speech and gesture generation, on the other hand, did not look at user-adaptation so far, but deals with the more basic problem to gain flexible yet convincing expressiveness, usually by employing a fixed gesture repertoire or model-based approaches. Recently, Neff et al. [19] proposed a data-based approach to generate character-specific gestures by capturing the individuality of human speakers. Based on statistical gesture profiles learned from annotated multimodal behaviour, their system takes arbitrary texts as input and produces synchronised conversational gestures in the style of a particular speaker.

In conclusion, current speech and gesture generation approaches cannot be used to model user-adaptation and alignment phenomena, either because they are not yet able to produce a sufficient range of behavioural variety to allow for unrestricted adaptation or they do not account for all the necessary generation levels (meaning, structure, form). In recent work we have proposed a production architecture that integrates model-based and data-based methods to overcome these weaknesses [3]. In this paper we describe how this model can be further extended to enable interactive alignment to the user.

3. EMPIRICAL BASIS

A growing body of research shows that alignment is a prevalent phenomenon in human interaction, with important functions for language processing, communication and social interaction. It is assumed to take place on all levels of linguistic representations, but due to the 'hidden' nature of semantic representations or situation models, directly observable evidence has been found only for phonetic, phonological, lexical and syntactic representations (cf. [20]).

Alignment phenomena can also be observed in human-computer interaction. An extensive review [17] recently concluded that humans do align – at least on some levels – when interacting with computers and that this behaviour could also be beneficial for virtual agents. Human-machine alignment is, especially regarding verbal behaviour, even likely to be stronger than alignment between humans and is mediated by efforts to enhance communicative success [6]. Unfortunately, the studies concerning natural language focus only on users aligning with a computer when interacting with ordinary user interfaces. Still, a question is how humans react when they communicate with embodied virtual agents that align to them – possibly in multiple modalities. This is largely unanswered, since there is no comprehensive system which is able to flexibly adapt to a user's conversational traits.

Apart from linguistic alignment, many studies show that people also align their bodily behaviour [9, 18]. This is usually called 'mimicry' or 'imitation' – depending on which aspects of an action are taken over. The same is found in the realm of communicative gestures, where speakers align their gestural forms and thus their meaning-form mapping [15]. This kind of gestural alignment, however, is only one kind of influence affecting gesture formation. Gestures heavily interact with speech production, where speech influences gesturing and vice versa [16]. Further, recent findings indicate that a gesture's form is crucially influenced by a number of contextual variables such as the visuo-spatial feature of its referent (accounting for a gesture's iconicity), the discourse context and even the previously performed gesture [4]. This suggests that gestures cannot go unaffected when speech changes due to linguistic alignment.

Empirical evidence, moreover, indicates that the production of gestures is shaped by concomitant speech: e.g., the packaging of content for gestures parallels linguistic information packaging [16] and gestures compensate for verbal encoding problems [2]. Since these findings concern several stages in the production process (such as content planning, speech semantics and lexical access) we hypothesised that there may also be a relationship between syntactic construc-

Table 1: Noun phrase patterns in the corpus (N= 4156; Part-of-speech tags according to the Stuttgart-Tübingen Tagset (STTS) for German).

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

tions and gesture production, and that such a relationship may be effected in conversational alignment letting gesture follow linguistic adaptation. Since it would also have to be considered for an adaptive multimodal production system, we investigated this kind of relation in a corpus analysis.

3.1 Corpus Analysis

As in previous work, our analysis is based on a corpus of speech and gesture use in spatial description tasks (25 dialogues, 4961 iconic/deictic gestures, 39435 words; cf. [4]). In the scope of the work reported here, we concentrate on noun phrases that were identified in 15 corpus transcripts by automatic Part-of-Speech tagging ([23]; Table 1 gives a list of the most common NP patterns found in our corpus).

3.1.1 Inter-subjective correlations

Our first analysis aimed to correlate NP patterns with gesture use. 37.8% of the 4156 NPs used are accompanied by gestures, and it turned out that there is a significant correlation between these two variables: for NP patterns (1)–(10) gestures co-occur significantly more often than expected, whereas in combination with patterns (11)–(15) the number of gestures is decreased ($\chi^2 = 248.9, df = 14, p < .001$). Thus, gestures are preferentially used accompanying noun phrases consisting of articles, adjectives, cardinals and nouns. In contrast, for noun phrases consisting only of pronouns and nouns, gesture use is less frequent.

In a second analysis we investigated the relation of NP patterns and gestural representation techniques. As concerns the latter, we distinguish the following five categories: *'indexing':* pointing to a position within gesture space; *'placing':* an object is placed or set down within gesture space; *'shaping':* an object's shape is contoured or sculpted in the air; *'drawing':* the hands trace the outline of an object's shape; *'posturing':* the hands form a static configuration to stand as a model for the object itself. Other less frequent techniques and combinations of techniques are counted as *'other'*.

All in all, we found a significant relationship between the two variables ($\chi^2 = 160.8, df = 70, p < .001$). On closer inspection, different gestural representation techniques co-occur with particular NP patterns in a significant way. For the patterns (1)–(4), i.e., patterns consisting of determiners, adjectives and nouns, the number of shaping gestures is significantly increased in comparison with expectation. NPs consisting of determiner and noun (5), come along with posturing gestures significantly more often than expected. Moreover, for cardinals (NP patterns (6/7)) placing gestures and for demonstrative pronouns (11) posturing gestures are used more often than expected. Furthermore, indexing gestures are frequently used along with personal pronouns (14), and for reflexive personal pronouns (15) the number of shaping gestures is significantly increased.

3.1.2 Individual Differences

To investigate in how far the correlations described above depend on the individual, we repeated both kinds of analysis under consideration of the individual speakers. Concerning the interrelation of gesture use and NP patterns, we found that the correlation of the two variables is highly significant for four speakers ($p < .001$), at least significant for another six speakers ($p < .05$) and no correlation is present in the data of five speakers. Similarly, the highly significant correlation between use of NP patterns and gestural representation techniques is present in five speakers' data ($p < .001$), for four speakers the correlation is still significant ($p < .05$), and in six speakers' data no correlation is present at all. Thus, similar to other factors influencing gesture use (cf. [4]), speakers in dialogue vary in how strong the link between NP patterns and gesture use is. Consequently, speakers should also vary in how far linguistic alignment may be reflected in their gestures. Nevertheless, such an effect should be present for the majority of speakers in our study.

4. COMPUTATIONAL MODELLING OF SPEECH & GESTURE PRODUCTION

4.1 Alignment-Capable Speech Formulation

Our starting point to build an agent that can align its communicative behaviour to a human, is to provide a natural language generation system whose flexibility can be exploited for simulating linguistic alignment effects. We focused on the stage of microplanning (put simply, the problem of turning meaning into linguistic form) and have developed the alignment-capable microplanner SPUD *prime*. The major difference between SPUD *prime* and the systems described in Section 2 is that it incorporates flexible priming/activation mechanisms. Thus, it can account for many alignment effects found in human communication, which do not only manifest in an utterance's surface form, but also in activation of underlying linguistic representations. We now give a short overview of SPUD *prime* and the results of its evaluation; further details can be found in [8].

4.1.1 Model and Implementation

For SPUD *prime* we adopt a simplified view of priming, where it results in two basic activation mechanisms: *temporary* and *permanent activation* which are both in accordance with empirical findings for alignment [22]. We call the former 'recency of use effects' and the latter 'frequency of use effects'.

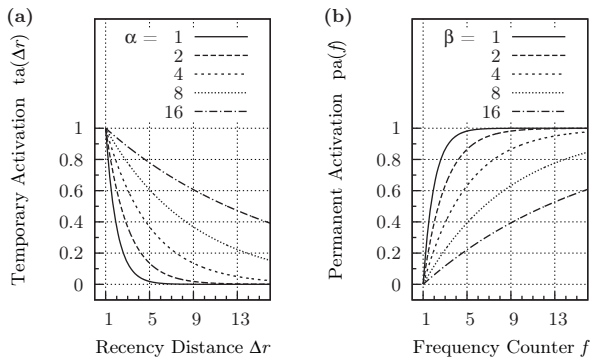

Figure 1: Plots of the mathematical models of recency and frequency effects: (a) temporary activation depending on the recency of priming; (b) permanent activation depending on the frequency count. Both plots are shown for different values of the slope parameters α and β.

Since corpus studies indicate that the repetition probability of primed syntactic structures depends logarithmically on the distance between priming and usage [22], we model recency of use effects by a general *exponential decay* function, modified to meet the needs for modelling activation decay of primed structures:

$$ta(\Delta r) = \exp\left(-\frac{\Delta r - 1}{\alpha}\right), \quad \Delta r \in \mathbb{N}^+;\ \alpha > 0; \quad ta \in [0,1] \tag{1}$$

$ta(\Delta r)$ is the temporary activation value of a linguistic structure depending on the distance Δr between the current time T and the time r at which the structure was primed. α determines the function's slope. A plot of $ta(\Delta r)$ with different values for α is given in Figure 1a.

To prevent frequency of use effects from leading to an ever increasing activation value, we model them with a general *exponential saturation* function, again modified to meet the requirements for modelling permanent activation of linguistic structures. This reflects the finding that the frequency effect is inversely connected to the recency effect [22].

$$pa(f) = 1 - \exp\left(-\frac{f - 1}{\beta}\right), \quad f \in \mathbb{N}^+;\ \beta > 0; \quad pa \in [0,1] \tag{2}$$

Here, the permanent activation value $pa(f)$ is not a function of time but a function of the frequency-counter f attached to each linguistic structure. Whenever a structure is primed, its counter is increased by the value of 1. Again, the function's slope is determined by a parameter, in this case β. A plot of $pa(f)$ with different slope parameters is given in figure 1b.

We combined both priming models by using a weighted linear combination of temporary and permanent activation:

$$ca(\Delta r, f) = \nu \cdot ta(\Delta r) + (1-\nu) \cdot pa(f), \quad 0 \le \nu \le 1; \quad ca \in [0,1] \tag{3}$$

Different values of ν allow different forms of alignment. With a value of $\nu = 0.5$ recency and frequency effects are equally important, with a value of $\nu = 1$ alignment depends on recency only, and with a value of $\nu = 0$ alignment is governed solely by frequency. Being able to adjust the influence of the different sorts of priming on alignment is crucial as it has not yet been empirically determined to what extent recency and frequency of use affect alignment.

The computational alignment model will not only consider alignment between interlocutors (interpersonal- or *other-alignment),* but also to oneself (intrapersonal- or *self-alignment*). Self-alignment is accounted for by the same priming-based mechanisms. To this end, four counters are attached to each linguistic structure: one for recency of use by the system itself (Δr_s), one for recency of use by the interlocutor (Δr_o), one counter for frequency of use by the system itself (f_s) and one for frequency of use by the interlocutor (f_o).

The overall activation value of the structure is modelled as a linear combination of the combined activation value $ca(\Delta r_s, f_s)$ and the combined activation value $ca(\Delta r_o, f_o)$ from equation (3):

$$act(\Delta r_s, f_s, \Delta r_o, f_o) = \mu \cdot ca(\Delta r_s, f_s) + (1-\mu) \cdot ca(\Delta r_o, f_o), \quad 0 \le \mu \le 1; \quad act \in [0,1] \tag{4}$$

Again, by changing the factor μ, smooth interpolation between pure self-alignment ($\mu = 1$) and pure other-alignment ($\mu = 0$) is possible, which can account for different empirical findings or human individual differences.

This priming-based model of alignment has been implemented by extending the integrated microplanning system SPUD *lite* [24], which is a lightweight Prolog re-implementation of the SPUD microplanning system [25], based on the context-free tree rewriting grammar formalism TAGLET. SPUD *lite* carries out the different microplanning tasks (lexical and syntactic choice, referring expression generation and aggregation) at once by treating microplanning as a search problem. During generation it tries to find an utterance which is meeting the constraints set by its input. This is done by searching the search space, spanned by the linguistic grammar rules and the knowledge base, until a goal state is found. Non-goal states are preliminary utterances and are extended by one linguistic structure in each search step until a syntactically complete utterance is found which conveys all the specified communicative goals.

Our alignment-capable microplanner SPUD *prime* extends SPUD *lite* in several ways. First, we altered the predicate for initial TAGLET trees by adding counters for self/other-recency/frequency values. Second, we have created a mechanism that enables SPUD *lite* to change the recency and frequency information attached to the initial trees on-line during generation. Finally and most importantly, the activation values of the initial trees – calculated with equation (4) – are considered during generation. Thus, in addition to the evaluation measures used by SPUD *lite*'s heuristic state evaluation function, the mean activation value

$$\overline{act}(S) = \frac{\sum_{i=1}^{N} act_{t_i}(\Delta r_{s_{t_i}}, f_{s_{t_i}}, \Delta r_{o_{t_i}}, f_{o_{t_i}})}{N}$$

of the N initial trees $\{t_1, \ldots, t_N\}$ of a given search state S is taken into account as a further evaluation measure. Hence, when SPUD *prime* evaluates (otherwise equal) successor search states, the one with the highest mean activation value is chosen as the next current state. The result of this is that aligned and highly activated structures are preferentially used in utterances generation.

4.1.2 Evaluation of SPUD prime

In previous work [8], we evaluated SPUD *prime* on t ken dialogue corpora that were collected in experim alignment in task-oriented dialogue. Pairs of parti played a 'Jigsaw Map Game', where they took turn structing each other to place objects on a table rel the previously placed objects. We simulated the di from the experiments – letting SPUD *prime* act as the two participants – and recorded whether it ge object names that match those the participants pr For each participant we did this for a fixed set of p SPUD *prime*'s parameter space, covering different t alignment behaviours.

For the first corpus of seven dialogues SPUD *prim* account for a mean of 89.8% of the target object (Min = 66.7%, Max = 100.0%, SD = 8.2%) which improvement of 24.6% on the baseline condition (ali switched off), where 65.3% of the target object nam be generated correctly.

The second experiment was a slightly revised ve the first one. For its corpus of 12 dialogues SPUD could account for a mean of 81.9% of all target object (Min = 56.3%, Max = 100.0%, SD = 12.2%) wh an improvement of 17.3% on the baseline conditio 64.3% of the target nouns could be generated correc

For each simulated speaker there were usually man in SPUD *prime*'s parameter space that led to a n number of mismatches between the generated object and the object names the speaker produced in the ment. Thus, different alignment behaviours could ha at work in a speaker to accomplish the same result. F more, the speakers' mean points in parameter spa produced a minimal number of mismatches also indicating that individual differences exist in thei ment behaviours. Overall, participants in the first ment tended to align to themselves, whereas particip the second experiment tended to align to their interl This was not surprising due to the differences in the imental setups and it can be concluded that SPUD retraced them successfully.

4.2 Gesture Formulation

Iconic gesture production in humans is influenced by several factors. Apparently, iconic gestures communicate through iconicity, i.e., their physical form depicts object features such as shape or spatial properties. Recent findings indicate that a gesture's form is also influenced by a number of contextual constraints and the use of more general gestural representation techniques such as shaping or drawing. In addition, inter-subjective differences in gesturing are pertinent. There is, for example, wide variability in how much individuals gesture when they speak. Similarly, inter-subjective differences are found in preferences for particular representation techniques or low-level morphological features such as handshape or handedness [4].

To investigate the challenge of considering general and individual patterns in gesture use, we have proposed GNetIc, a gesture net specialised for iconic gestures [3], in which we model the process of gesture formulation with Bayesian decision networks (BDNs) that supplement standard Bayesian networks by decision nodes. This formalism provides a representation of a finite sequential decision problem, combining probabilistic and rule-based decision-making. Each decision to be made in the formation of an iconic gesture (e.g., whether or not to gesture at all or which representation technique to use) is represented in the network either as a decision node (ruled-based) or as a chance node with a specific probability distribution. Factors which contribute to these choices (e.g., visuo-spatial referent features) are taken as input to the model. Individual as well as general networks are learned from annotated corpora by means of automated machine learning techniques and supplemented with rule-based decision making (see the *Gesture Formulator* in Figure 2).

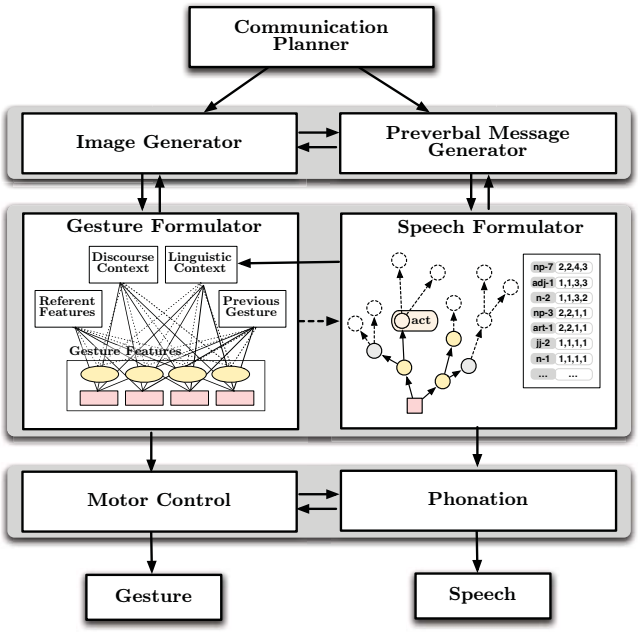

Figure 2: Overall production architecture with a focus on microplanning: The gesture formulator box shows the GNetIc decision network; the speech formulator box shows a search tree, evaluated with the activation function, and linguistic representations with recency and frequency counters.

Learning a Bayesian network from a sample of data cases comprises two tasks. First, the structure of the Bayesian network is learned using the constraint-based NPC algorithm. Second, as the network structure is found, maximum likelihood estimates of parameters are computed employing the EM algorithm. Certain variables of a complete gesture specification that cannot be learned from data, but are subject to inter-subjective regularities (e.g., to employ a certain handshape for a certain referent shape feature) are modelled as if-then rules in the decision nodes. So far, we have incorporated three different factors into this model: referent features, discourse context and the previously performed gesture. A prediction-based evaluation of this account to gesture formulation, in which we compared generated gestures with empirically observed counterparts, yielded very promising results [3].

4.3 Overall Production Architecture

Both production modules described so far, i.e., the alignment-capable microplanner SPUD *prime* and the GNetIc approach to gesture formulation, have been integrated into

an overall production architecture which is psycholinguistically inspired [16] and characterised by a close interplay between all production modules. As outlined in Figure 2, it consists of interacting, modality-specific modules at each of three stages: (1) *Image Generator* and *Preverbal Message Generator* are concerned with content planning; (2) *Gesture Formulator* and *Speech Formulator* turn content into form; (3) *Motor Control* and *Phonation* realise it as synchronised speech and gesture animations. All modules operate concurrently and proactively on a central working memory, realised as a globally accessible, structured blackboard on which the overall production process evolves. In this way, interaction among the modules realises content planning and microplanning in an interleaved and interactive manner.

5. ADAPTIVE MULTIMODAL EXPRESSIVENESS

How can the interaction between SPUD *prime* and GNetIc allow for modelling adaptive multimodal expressiveness? Our approach is to make particular gesture formulation choices depending on the linguistic context in terms of the chosen noun phrase pattern. For this purpose the gesture networks are learned from individual speakers' data enriched with information about the syntactic construction used in a gesture's concomitant speech. As expected – due to our empirical results presented in Section 3 – links between the variable 'NP pattern' and gesture features emerged for some of the speakers' data – but not for all of them. During the online production process of multimodal utterances, the Speech Formulator provides information about the NP pattern chosen for the Gesture Formulator. Here, the information is taken as evidence in GNetIc and, thus, exerts influence on the gesture generation choices.

Due to the interplay of our activation-based microplanner making linguistic choices based on lexical and syntactic priming, and the gesture generation module capable of being influenced by different variables, our system is now not only able to model linguistic alignment, but – as a novel feature – also to have it paralleled by concordant gestural adaptation. As an example consider Figure 3, which displays a decision network for one particular speaker. In addition to influences from referent features, discourse context and the previously performed gesture, the noun phrase pattern used in speech has an impact on the gesture generation choices.

5.1 First Results

To illustrate interaction between speech and gesture formulation processes, we will now walk through three generation examples to show how the agent aligns to its interaction partner by picking up lexical choices and syntactic structures in its subsequent utterances. These changes in the agent's verbal behaviour also have a direct impact on gestural expressions. Each example starts upon the arrival of a message from the Communication Planner which specifies the communicative intent to mention two landmarks:

```
introduce_lm(landmark-1, landmark-2).
```

Based on this communicative intention, the Image Generator activates the imagistic descriptions of all objects involved in the communicative goal and the Preverbal Message Generator starts by selecting the following propositions:

```
private(inst(landmark-1, church)).
```

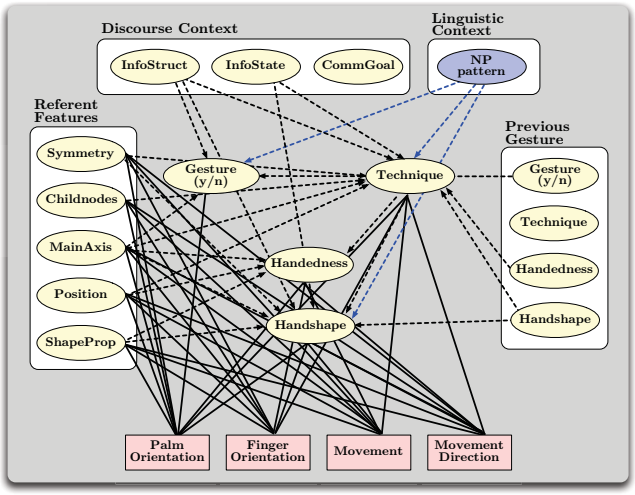

Figure 3: A GNetIc decision network (learned from one speaker's data) including NP pattern as an influencing variable for gesture generation choices.

```
private(relpos(landmark-1, left)).
private(inst(landmark-2, church)).
private(relpos(landmark-2, right)).
```

These propositions and the communicative goal are then accessed by the Speech Formulator in each of the following three example cases.

Example without priming

In the first case the Speech Formulator processes the knowledge base and the communicative intention without any priming. One could assume for instance that the dialogue partner has asked *'Welche Sehenswürdigkeiten gibt es auf dem Platz?'* ('Which sights are there on the square?'). Due to the generation grammar specified for our domain of application, SPUD *prime* generates the following verbalisation: *'Es gibt zwei Gebäude.'* ('There are two buildings.'). The evidence available (referent features, discourse context, previous gesture and linguistic context) is propagated through the BDN in the Gesture Formulator resulting in a posterior distribution of probabilities for the values in each chance node for gesture features. Depending on the network used (we have chosen a speaker's network in which there is a strong relation between NP pattern and gesture use), the kind of NP pattern planned by the Speech Formulator ('CARD NN' in this case) in the verbal utterance is considered in the process of decision making. The network is accessed for each of the two referents resulting in a left-handed placing gesture for `landmark-1` and a right-handed placing gesture for `landmark-2`. Further features of the gesture are also chosen in the network. Finally, the temporal relation of speech and gesture is determined on the basis of semantics: *'zwei Gebäude'* ('two buildings') refers to both referents, thus, both gestures are produced in temporal synchrony with the noun phrase (see Figure 4a).

Example with lexical priming

Now the Speech Formulator receives the same knowledge base and the same communicative intention as in the first example, but this time with lexical priming. That is, we as-

Figure 4: Example utterances generated on-the-fly from the same communicative goal but in different alignment conditions (see text for English translation): (a) simple noun phrase construction with two-handed placing gesture produced without previous priming; (b) conjunction of two noun phrases along with indexing gestures in temporal sequence resulting from syntactic priming.

sume the agent's interlocutor uttering a sentence such as *'Ich würde gerne eine Kirche sehen.'* ('I'd like to see a church.') prior to the generation process. Based on such an utterance generation is primed to use the lexem *'Kirche'* ('church'). This is passed to the system as additional input data. The output produced by SPUD *prime* this time is *'Es gibt zwei Kirchen.'* ('There are two churches.'). Since the syntactic construction of the utterance ('CARD NN') is not changed in comparison with the first example, the BDN in the Gesture Formulator plans the same gestural behaviour as before (see Figure 4a), but along with a different verbal phrase.

Example with syntactic priming

To demonstrate the effect of syntactic priming on speech and gesture production, we assume an interaction partner uttering *'Die Kirche hat oben ein Fenster und unten eine Tür.'* ('The church has a window at the top and a door at the bottom.') beforehand. This results in priming of the syntactic structure 'ADV NP *und* ADV NP'.

Now, SPUD *prime* generates the following utterance from the same given communicative goal: *'Es gibt rechts eine Kirche und links eine Kirche.'* ('There is one church on the right and one church on the left.'). The Gesture Formulator receives the different syntactic construction as input ('ART NN'). This fact changes the production choices such that a different representation technique is planned for both gestures, in this case indexing. Due to the different syntax of this sentence in comparison with the two previous examples, the right-handed indexing gesture for the right church is to co-occur with *'[rechts] eine Kirche'* ('one church [on the right]') while the left-handed indexing gesture is to co-occur with *'[links] eine Kirche'* ('one church [on the left]'). The resulting multimodal behaviour is displayed in Figure 4b.

6. DISCUSSION AND CONCLUSION

This paper presented work that makes first and novel steps towards virtual agents that can interactively align to their interaction partners in their verbal and gestural behaviour. This is achieved by endowing a flexible model for speech and gesture production with abilities for being biased in its linguistic as well as gestural choices. Our first results demonstrate that the interplay of our activation-based microplanner and the gesture generation module enables a conversational virtual agent to align – lexically, syntactically and with concordant gestural behaviour – to a user's utterance in an interactive manner. Some points, however, merit closer inspection.

First, in our current model linguistic alignment is seconded by concordantly adapted gestures, such that an empirically observed coupling between both modalities persists throughout an interaction. While the 'linguistic route' currently happens to be predominant in mediating alignment, both modalities incorporated in the system should, in principle, have the potential to model both individual style and priming/activation-based alignment. Taking individual differences into account in speech formulation is in principle supported through setting the alignment model's parameters to certain values. For the work presented here, it was just not feasible to calculate these for the speakers in our corpus of speech and gesture data. Likewise, the priming-based alignment model is also applicable to gesture generation, for instance, by increasing or decreasing the prior probabilities within the decision network. In ongoing work, we are developing a probabilistic model for embodied gesture perception tailored to yield activations of motor components at a level that is also being used in the gesture production framework (cf. [17]). Connecting the two is left to future work.

Another important point to note is that the work presented here employs linguistic knowledge in form of a grammar (defining both lexicon and syntax) that is attuned for language generation, albeit extended to activations within the grammar. These activations are assumed to result from processing of verbal user input. Interactive alignment, hence, suggests shared (or at least connected) linguistic representations for both tasks, generating and parsing/understanding language, such that activations from hearing an interlocutor using a certain lexical entry or syntactic constructions makes this item more likely to be used in one's own deliveries. However, while symmetrical theoretical models exist (e.g., Dynamic Syntax [21]), natural language processing systems usually employ different representations for generation and for understanding, since both tasks face different challenges and are thus often modelled independently. Our present system employs two essentially equivalent representations for each lexical item and syntactic construction, one optimised for understanding, the other for generation. Since both can be addressed by the same identifier, reception and production, respectively, always prime both representations. In ongoing work we are currently investigating in how far a TAG-based grammar can be developed for dual use, and how the mechanisms behind SPUD *prime* can as well be employed for parsing and interpreting speech. Experiences so far hint to the fact that grammar formalisms as the one employed here do not to scale too well. This may suggest adopting a different format like construction grammars some of which are developed for bi-directional use.

Finally, we want to point out that we currently model gestural behaviour to follow the (possibly aligned) linguistic choices such that empirically observed correlations between speech and gesture are retained. However, one must be cautious to note that we do not necessarily assume a causal influence of lexical or syntactic choices onto gestural behaviour. Such a link could just as well be mediated via the semantic level, i.e., gestures can reflect different activations of visuo-spatial mental imagery that may result from priming of mental representations of language semantics. Our production architecture is prepared to model such

aspects of cross-modal interaction which, however, calls for sufficient natural language understanding capabilities.

Overall, the complexity of the phenomenon tackled here and of its presumable cognitive underpinnings open up exciting perspectives for research on conversational agents and their appropriate behaviour in interaction with human users. One remaining question, for example, is still how communicative virtual agents with adaptive expressiveness affect the human-agent interaction [17]. We are confident that the work presented here provides promising steps in this direction and may ultimately yield conversational agents that behave in appropriate and acceptable ways when interacting with human users.

7. ACKNOWLEDGMENTS

This research is supported by the German Research Foundation (DFG) in the Center of Excellence 277 in 'Cognitive Interaction Technology' (CITEC) as well as in the Collaborative Research Center 673 'Alignment in Communication'.

8. REFERENCES

[1] J. N. Bailenson and N. Yee. Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological Science*, 16:814–819, 2005.

[2] J. Bavelas, J. Gerwing, C. Sutton, and D. Prevost. Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, 58:495–520, 2008.

[3] K. Bergmann and S. Kopp. GNetIc – Using Bayesian Decision Networks for iconic gesture generation. In *Proceedings of the 9th International Conference on Intelligent Virtual Agents*, pages 76–89, Amsterdam, The Netherlands, 2009.

[4] K. Bergmann and S. Kopp. Increasing expressiveness for virtual agents – Autonomous generation of speech and gesture in spatail description tasks. In *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems*, pages 361–368, Budapest, Hungary, 2009.

[5] J. K. Bock. Syntactic persistence in language production. *Cognitive Psychology*, 18:355–387, 1986.

[6] H. P. Branigan, M. J. Pickering, J. Pearson, and J. F. McLean. Linguistic alignment between people and computers. *Journal of Pragmatics*, in press.

[7] S. E. Brennan and H. H. Clark. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22:1482–1493, 1996.

[8] H. Buschmeier, K. Bergmann, and S. Kopp. An alignment-capable microplanner for natural language generation. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 82–89, Athens, Greece, 2009.

[9] T. L. Chartrand and J. A. Bargh. The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, 76:893–910, 1999.

[10] H. H. Clark. *Using Language*. Cambridge University Press, Cambridge, UK, 1996.

[11] M. de Jong, M. Theune, and D. Hofs. Politeness and alignment in dialogues with a virtual guide. In *Proceedings of the 7th International Conference on Autonomous Agents and Multiagent Systems*, pages 207–214, Estoril, Portugal, 2008.

[12] H. Giles and N. Coupland. *Language: Contexts and Consequences*. Wadsworth Publishing, 1991.

[13] A. Isard, C. Brockmann, and J. Oberlander. Individuality and alignment in generated dialogues. In *Proceedings of the 4th International Natural Language Generation Conference*, pages 25–32, Sydney, Australia, 2006.

[14] S. Kang, J. Gratch, N. Wang, and J. H. Watt. Does the contingency of agents' nonverbal feedback affect users' social anxiety? In *Proceedings of the 7th International Conference on Autonomous Agents and Multiagent Systems*, pages 120–127, Estoril, Portugal, 2008.

[15] I. Kimbara. On gestural mimicry. *Gesture*, 6:39–61, 2006.

[16] S. Kita and A. Özyürek. What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, 48:16–32, 2003.

[17] S. Kopp. Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors. *Speech Communication*, in press.

[18] J. Lakin, V. Jefferis, C. Cheng, and T. Chartrand. The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Psychological Science*, 19:816–822, 2003.

[19] M. Neff, M. Kipp, I. Albrecht, and H.-P. Seidel. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics*, 27:1–24, 2008.

[20] M. J. Pickering and S. Garrod. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169–226, 2004.

[21] M. Purver, R. Cann, and R. Kempson. Grammars as parsers: Meeting the dialogue challenge. *Research on Language and Computation*, 4:289–326, 2006.

[22] D. Reitter. *Context Effects in Language Production: Models of Syntactic Priming in Dialogue Corpora*. PhD thesis, University of Edinburgh, 2008.

[23] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, 1994.

[24] M. Stone. Lexicalized grammar 101. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 77–84, Philadelphia, PA, 2002.

[25] M. Stone, C. Doran, B. Webber, T. Bleam, and M. Palmer. Microplanning with communicative intentions: The SPUD system. *Computational Intelligence*, 19:311–381, 2003.

[26] M. Walker, A. Stent, F. Mairesse, and R. Prasad. Individual and domain adaptation in sentence planning for dialogue. *Journal of Artificial Intelligence Research*, 30:413–456, 2007.