Appl Intell (2012) 37:305–320
DOI 10.1007/s10489-011-0329-5

Focus tree: modeling attentional information in task-oriented human-machine interaction

Milan Gnjatović · Marko Janev · Vlado Delić

Published online: 7 December 2011

Abstract This paper introduces a new model of attentional state in task-oriented human-machine interaction. It integrates three lines of research: (i) neurocognitive understanding of the focus of attention in working memory, (ii) the notion of attention related to the theory of discourse structure in the field of computational linguistics, and (iii) investigation of a corpus that comprises recordings of spontaneous speech-based human-machine interaction. The underlying idea was to make a computationally appropriate representation of attentional information that imitates the function of a focus of attention in human perception. The introduced model addresses both the research questions of storage and processing of attentional information. Finally, the paper illustrates the model for concrete interaction domains, and discusses its implementation within a prototype spoken dialogue system.

Keywords Focus tree · Attentional information · Human-machine interaction · Cognition · Utterance processing

1 Introduction

This paper introduces a new model of attentional state in task-oriented human-machine interaction (HMI). It integrates three lines of research: (i) neurocognitive understanding of the focus of attention in working memory [11, 43], (ii) the notion of attention related to the theory of discourse structure in the field of computational linguistics [23], and (iii) investigation of a corpus that comprises recordings of spontaneous speech-based HMI [20]. The underlying idea was to make a computationally appropriate representation of attentional information that imitates the function of a focus of attention in human perception. While first two research lines address the research question of modeling attentional information in general, the third line of research (i.e., the corpus) provides a specific, data-driven view of the focus of attention in HMI. However, it should not be understood that the introduced approach is restricted to the observed corpus only—the corpus is used here for illustration purposes.

To the extent that the model is computationally appropriate, the discussion is concentrated on the research problem of robust automatic processing (i.e., recovering semantic information) of different syntactic forms of spontaneously uttered users' commands with no explicit syntactic expectations. Forcing users to always produce utterances that follow rules of a preset grammar would be too restrictive and not well accepted. It cannot be expected that users will always behave cooperatively and produce utterances that fall within the application's domain, scope and grammar. Thus, there is a need to enable the system to handle flexible mapping relations between the spontaneously produced user's commands and the system's actions (cf. [55]). Attentional information has been already recognized as essentially important for processing of utterances in discourse [23, 40]. In this paper, we discuss how the introduced model of attentional state may be used to process spontaneously uttered users' commands. This represents an integration of our previous work on processing users' commands and designing adaptive dialogue strategies [17–19, 21].

M. Gnjatović (✉) · V. Delić
Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia
e-mail: milangnjatovic@yahoo.com

M. Janev
Mathematical Institute of the Serbian Academy of Sciences and Arts, Kneza Mihaila 36, 11001 Beograd, p.p. 367, Serbia

2 Underlying concepts

From the methodological point of view, the introduced computational model of attentional state is inspired by human information processing system. Thus, this approach belongs to the interdisciplinary fields of natural computing [30] and brain informatics [56]. From the conceptual point of view, it is in line with an understanding of computation as interaction, i.e., the joint manipulation of concepts and actions by human and software agents [45]. The following subsections introduce relevant underlying concepts in more detail.

2.1 Focus of attention in cognitive models of working memory

Working memory is fundamentally related to human cognition, and there is a variety of models and theories that reflect diverse perspectives on working memory. We do not aim here to provide a complete overview of different cognitive models of working memory (for detailed overviews cf. [7, 49, 53]), but rather to highlight some widely accepted perspectives on the relation between working memory and attention that are relevant for our discussion.

(i) The concept of attention is closely related to working memory. Human cognition strongly relies on the ability to appropriately filter and organize information for further use [7, p. 172]. Attention is considered as a filtering mechanism that limits the amount of information in a memory store [49, pp. 16–18]. In some traditional conceptions of working memory [2, 8], attention and memory are considered distinct. In contrast to them, current process models of working memory incorporate attention control functions [3, 5–7, 11, 42, 43]. Based on them, Bledowski et al. [7] provided a review of basic operations on the contents in working memory. One of them is updating the focus of attention with the selected item. Considering cognitive models of working memory, they devote special attention to the models introduced by Cowan [11] and Oberauer and Lange [43]. In his embedded-process model, Cowan proposes that working memory is a functional state that allows a direct access to activated part of long-term memory. Based on this model, Oberauer and Lange [43, p. 104] (cf. also [42, p. 412]) conceptualize working memory as a concentric structure of representations with three functionally distinct regions (see Fig. 1):

- The activated part of long-term memory holds representations that are activated above baseline through a match with perceptual input or through spread of activation in long-term memory.
- The region of direct access holds a limited number of activated representations that are temporarily bound to a common cognitive coordinate system. Such a common coordinate system can be a temporal context, a spatial context, etc. These representations are available for ongoing cognitive, goal-directed processes.
- The focus of attention holds a representation from the region of direct access that is directly affected by a cognitive, goal-directed operation. At any time, a single representation is selected to be in the focus of attention.

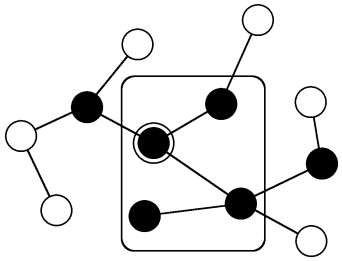


Fig. 1 A concentric model of working memory, adopted and adjusted from the original work of Oberauer [42, p. 412]. *Nodes* and *lines* represent a network of long-term memory representations. *Black nodes* represent activated representations. The region of direct access (*big oval*) holds a limited number of activated representations. Within this region, one representation is selected to be in the focus of attention (*small oval*)

Following Oberauer [42, p. 412], retrieving an item from working memory means bringing this item into the focus of attention. For him, the focus of working memory has a function with respect to memory that is equivalent to the function of a focus of attention in perception. This observation is in line with findings that spatial attention and internal representations in working memory are closely interrelated [1, 22, 41, 54].

(ii) The capacity of working memory is limited. Although it is widely accepted that the amount of information in working memory is limited [12, 36], there is often little agreement between researchers on which mechanisms constrain the capacity of working memory [49, pp. 10–12]. Diverse factors that are considered to underlay the capacity limitations include: inhibitory mechanisms [44, 51], processing speed [47], domain-specific storage of information [16], limited amount of activation in the system [15, 28], etc. Recent work suggests two important determinants of capacity of working memory: attention control and basic memory abilities [53]. In attention based theories, individual differences in capacity of working memory is primarily determined by attention control, i.e., maintaining task-relevant information in an active state in conditions of interference or competition [14, 29]. In memory based theories, capacity of working memory is primarily determined by ability to access information from long-term memory [38]. Finally, in the dual-component model [52, 53], capacity of working memory is jointly determined by both attention control and memory abilities.

(iii) Storage-oriented and processing-oriented aspects of working memory. Working memory is commonly conceptualized as a system for simultaneous storage and processing of information [42, p. 411]. The storage aspect relates

to the scope of the focus [13], including basic mechanisms (i.e., encoding, maintenance, and retrieval of information) and representation of information in working memory [49, p. 6]. The processing aspect relates to attention control [14]. Cognitive models of working memory differ in how they address this simultaneity mechanism. The multiple-component model of working memory introduced by Baddeley and colleagues [4, 5] comprises specialized components: two temporary memory systems (the phonological loop and the visuospatial sketchpad) used to actively maintain memory traces, and a supervisory system (the central executive) involved in control of the working memory system. The dual-component model of working memory introduced by Unsworth and Engle [52] suggests a probabilistic cue-dependent search component (secondary memory) and a dynamic attention component. The concentric model of working memory introduced by Oberauer [42] and embedded-process model introduced by Cowan [11] distinguish between the activated part of long-term memory and the focus of attention. However, while much research in the neurosciences was devoted to the storage function, the processing aspect remains underspecified [7, p. 177].

2.2 Focus of attention in the theory of discourse structure

The second line of research lies within the field of computational linguistics, and relates to the theory of discourse structure introduced by Grosz and Sidner [23, pp. 177–182]. They introduce a model of attentional state as one of the components of general discourse structure. For them, attentional state contains information about the objects, properties, relations and discourse intentions that are most salient at any given point. They model the attentional state by a set of *focus spaces*. They call the collection of focus spaces available at any one time the *focusing structure*, and the process of manipulating spaces—*focusing*. In the focusing process introduced by Grosz and Sidner, a focus space is assigned to each discourse segment. A focus space contains entities that are salient in the given discourse segment, e.g., entities that have been mentioned explicitly or introduced implicitly in the process of producing or comprehending the utterances [23, p. 179]. This is illustrated in Fig. 2 that is adopted and adjusted from the original work of Grosz and Sidner [23, pp. 180–1].

The intentional structure of the discourse in the given example, including relationships among discourse segment purposes, is represented in the dominance hierarchy on the left in the figure. Discourse segment DS1 dominates discourse segments DS2 and DS3. The focusing structure is given on the right in the figure. Each of these discourse segments is tied to a focus space. The state of focusing when discourse segment DS2 is being processed is given in the first part of Fig. 2. Being the most salient, focus space FS2 is

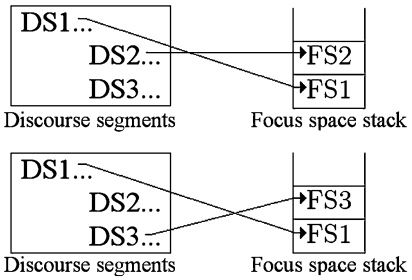


Fig. 2 Discourse segments and focus spaces, adopted and adjusted from the original work of Grosz and Sidner [23, p. 181]

positioned on the top of the stack. Focus space FS1, assigned to the dominating discourse segment DS1, is also accessible, although less salient. When discourse segment DS3 is being processed, focus space FS2 has been popped from the focus space stack, and focus space FS3 has been pushed onto it.

Grosz and Sidner provide concrete, well-elaborated examples that illustrate their theory of discourse structure [23, pp. 182–192]. They note that their theory, although still incomplete, does provide a solid basis for investigating both the structure and meaning of discourse, as well as for constructing discourse-processing systems. However, they also suggest that one of research problems of primary importance that remain to be further explored is investigation of alternative models of attentional state [23, p. 202]. In the Grosz and Sidner's theory, the focusing structure is parasitic upon the intentional structure, i.e., the relationships among discourse segment purposes determine pushes and pops [23, p. 180]. On the other hand, an observation that is widely accepted in Conversational Analysis is that intentionality is not given at the beginning of a conversation and that it evolves as the conversation proceeds [23, 46, 48]. Grosz and Sidner's theory is consistent with this observation—the focusing structure, since it is determined by the intentional structure, is not given a priori and also evolves as the discourse proceeds. However, from the technical aspect of developing a conversational agent, this property makes the focus stack somewhat inflexible in topic management since it requires the pushing and popping of focus spaces in a particular order, and once that focus space has been popped from the stack, it cannot be referred to except by reintroducing it [26, p. 88].

(iv) The notion of the focus tree. In attempt to provide a (more) unified account of focus phenomena, McCoy and Cheng [34] introduce a tree-structured discourse model—the *focus tree*. Their point of departure is that, during the course of conversation, the participants (including conversational agents) focus their attention on some subset of their knowledge. The focus tree is a hierarchical structure of interrelated concepts that represents a subset of the agent's world knowledge and contains those concepts that are in the focus of attention. In addition, at any point in a coherent conversation, the focused knowledge represents the knowledge that

is likely to be included next in the interaction between the user and the system [34, p. 104]. Therefore, with respect to the dialogue coherence, the focus tree both constrains and enables prediction of what is likely to be talked about next [27, p. 633]. With respect to the topic management, it supports a more flexible management of focus shifts by allowing different strategies to traverse the tree structure.

Although the work of McCoy and Cheng was primarily concerned with text generation, the concept of the focus tree is applied and adjusted in various approaches to dialogue management in HMI. Jokinen et al. [27] introduce the topic tree—a probabilistic topic model intended to be used as a context model for spoken dialogue systems. Stede and Schlangen [50] use hierarchical organization of topics combined with weights representing discourse history in information seeking chats. Kirschner [31] applies the focus tree to represent dialogue context in interactive question answering systems. Moeller [37] uses the concept of the focus tree to provide plans that ensure the generation of coherent domain-oriented dialogues. Hovy and McCoy [25] propose to use focus trees together with Rhetorical Structure Theory trees. However, we do not aim here to provide a complete overview of different models based on the concept of the focus tree, but rather to discuss important properties that are relevant for our discussion.

(v) Each node of the focus tree represents a specific conversational topic. Conversational topics represent concepts that are currently in the focus of the dialogue participants, e.g., concepts that has already been mentioned in the ongoing dialogue [31]. An important implication is that conversational topics may be assigned to individual dialogue acts. For example, in the topic tree introduced by Jokinen et al. [27], nodes of the tree correspond to topics which represent clusters of the words expected to occur at a particular point of the dialogue. Jokinen et al. propose an algorithm that uses the information structure of the dialogue act to link it to a topic (i.e., a node) in the tree.

(vi) Branches of the focus tree indicate possible shifts of the focus of attention. In other words, branches of the tree indicate focus shifts that are cognitively easy to process, and that can be expected to occur in dialogues, i.e., focus transitions from a node to its children or siblings are considered to be more likely than shifts to nodes in separate branches [27]. Therefore, the focus tree constrains focus shifts and enables prediction of ensuing focus shifts (e.g., focus shifts that are likely to be expected in the course of conversation) [34].

(vii) The focus tree should not contain task structure or intentional relationships. In the theory of discourse structure, the focusing structure and the intentional (e.g., task) structure are introduced as interrelated but distinct. Grosz and Sidner argue that conflation of these two structures is a misinterpretation of their theory, since it prevents a theory from accounting adequately for certain aspects of discourse [23, pp. 180–2]. Therefore, the focus tree, which is also a conceptualization of attentional state, should not contain information about task structure or intentional relationships. This distinction is not just of theoretical interest, but is also very important from the aspect of developing a conversational agent. Lecœuche et al. propose to separate reasoning tools and dialogue managers in conversational agents, and to make them interact and constrain each other [32, pp. 23–4]. They note that reasoning tools are usually driven by an agenda of task to perform, and should not be aware of dialogue management strategies needed to ensure a natural dialogue. This is particularly important if we want to reuse the same reasoning tool in different environments where interaction rules may differ from those of natural language. Thus, Lecœuche et al. introduce a dialogue manager driven by focus rules in order to ensure that the spoken interaction between the user and the system follows human dialogue conventions. A conceptually similar approach can be found in a robotic architecture for HMI applications introduced by Mohammad and Nishida [39, p. 149]. Every process in this parallel architecture has two attributes that control its contribution to the behavior of the robot: actionability (i.e., the activation level of a process) and attentionality (i.e., the relative attention that should be given to this process). Mohammad and Nishida differentiate between these two attributes to provide a mechanism for implementing attention focusing. In their words, it allows the robot to select the active processes depending on the general context (the actionability value) and to assign the computation power according to the exact environmental and internal condition (the attentionality value). Both these approaches to attention focusing described in [32, 39] are in line with the distinction between the focusing structure and the intentional structure in the theory of discourse structure.

Finally, it should be mentioned that the discussion on the appropriateness of tree-structured models of attentional state is just a part of a wider discussion on the research question of defining a computationally appropriate structure to represent human knowledge (for an insight into recent developments cf. also [33, 35]). Considering this research question, Li and Tsai [33, p. 68] emphasize two major concerns: how to formulate human cognition, and how to make it in a form that users can visualize. They discuss that human cognition is believed to have generally hierarchical properties, and that, therefore, hierarchical structures (i.e., trees) are more appropriate to describe human cognition rather than non-hierarchical structures. Related to the latter concern, Li and Tsai note that human visualization capacity is limited with regard to information processing, and that structures with large numbers of objects or links between them may reduce the effectiveness of knowledge navigation and visualization. However, this concern is not of critical importance

in our approach, since we use a tree structure to represent activated representations in working memory whose number is inherently limited (cf. Sects. 2.1 and 3).

2.3 Focus of attention in the NIMITEK corpus

This section considers attentional information for the model of commands contained in the NIMITEK corpus of affected behavior in speech-based HMI produced by Gnjatović and Rösner [20]. It contains 15 hours of audio and video recordings produced during a refined Wizard-of-Oz study designed to induce emotional reactions. Ten healthy native German speakers (seven female, three male; ages 18 to 27, mean 21.7) participated in the study. All dialogues were transcribed. The number of dialogue turns is 1,847 for the subjects and 1,846 for the wizard (i.e., the simulated system). The average number of words per turn is 17.19 for the subjects (with standard deviation 24.37), and 8.53 for the wizard (with standard deviation 8.70). The subjects' lexicon contains about 900 lemmata (i.e., root forms). Evaluation of the corpus showed that 98.09 percent of subjects' verbal dialogue acts were spontaneously produced. The class of spontaneously uttered commands is the most represented class—there are 6798 (74.79 percent) commands in the NIMITEK corpus. Therefore, we performed a corpus-based investigation of a typology of spontaneously uttered users' commands with respect to the propositional content. We used a corpus-specific classification of commands based on the inspection of the NIMITEK corpus and on observations on the structure of spoken language made by Campbell [9, 10].

Considering the structure of spoken language, Campbell differentiates between two types of content that are often simultaneously signalled in spontaneous spoken language: propositional content and affect. He introduces the notions of *fillers* and *wrappers* to denote parts of utterances that relate to these two types of content. Keeping his observations in mind, we conducted an inspection of commands from the NIMITEK corpus. As expected, they often contain words or phrases (i.e., fillers) that explicitly relate to entities from the currently salient focus space. We illustrate this for the interaction scenario when the subjects solve the Tower of Hanoi puzzle. Some typical examples of the users' commands are:

- The two on the three. (Die Zwei auf die Drei.)
- The next disk. (Den nächsten Ring.)
- Rightwards. (Nach rechts.)

In general, a fully formulated command in the Tower of Hanoi puzzle is expected to contain following information: which disk should be moved, and to which peg it should be moved. In the first command, the subject uses the phrase "the two" to refer to the second disk, and the phrase "the three" to refer to the third peg. It should be noted that choice of lexical items may introduce ambiguity, e.g., isolated from the surrounding dialogue context, the phrase "the two" may (and does) relate to the second disk or to the second peg. In addition, the subjects often assume that the graphical interface represents a non-linguistic context shared between them and the system. Consequently, they use ellipses [24], a form of grammatical cohesion where they omit to utter information that is already known by the system and, in the same time, bring new information in the focus of attention. For example, the second and third commands are elliptical (i.e., they contain information only about the disk and only about the peg, respectively) and cannot be interpreted without taking the context into account. Thus, in the third command, the subject instructs that a previously selected disk should be moved on the next peg on the right. This command also illustrates a form of lexical cohesion. The subject uses an adverb ("rightwards") to specify a peg, although a nominal phrase might appear as more appropriate. Therefore, lexical items that relate to the same entity from the currently salient focus space do not have to match grammatically.

In contrast to this, commands may also contain some additional information that does not directly relate to propositional content (e.g., phrases of courtesy, etc.). Some examples from the NIMITEK corpus are:

- I would like to put *the smallest disk* on *the three*. (Ich würde gern' *die kleinste Scheibe* auf *die Drei* legen.)
- *The middle disk* please on *the number two*. (*Den mittleren Ring* bitte auf *die Nummer Zwei*.)

Words and phrases that relate to propositional content (i.e., fillers) are given in italics. From the system's point of view, fillers are important for understanding propositional content. We consider fillers in a restricted scope—to denote parts of utterance that carry new or salient information. We refer to them using the term—*focus stimuli*.

Finally, the subjects often use negation in attempt to correct the system's behavior:

- *Don't rotate*, but move to right. (*Nicht drehen*, sondern nach rechts schieben.)
- But *not picture one*. (Aber *nicht Abbildung Eins*.)
- *Not on the first* ... (*Nicht auf der erste* ...)

Phrases that express negation are given in italics. In this strategy for recovering from non-understandings, the subjects try to help the system by explicitly referring to entities from the current interaction domain that should not be in the focus of attention. Therefore, we refer to these phrases using the term—*negative reinforcement stimuli*.

Inspection of the NIMITEK corpus shows that 80.45 percent of all spontaneously produced commands contains focus stimuli or negative reinforcement stimuli. Thus, this class of commands has a central position in our approach to processing users' commands. On the surface level, utterance chunks that we refer to as stimuli are non-recursive phrases.

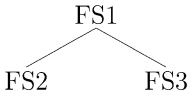
Propositional content is expressed by frequent insertion of chunks (i.e., stimuli) that explicitly relate to entities from the currently salient focus space. The order of stimuli within an utterance is flexible, while the word order within them is rather fixed. This allows the robust processing of different syntactic forms of spontaneously produced users' utterances (from syntactically very simple utterances to verbose utterances) with no explicit syntactic expectations. However, on the level of dialogue structure, stimuli carry information related to the attentional state. Therefore, in our approach, focus stimuli and negative reinforcement stimuli represent incoming linguistic stimuli that may change the focus of attention. The system should be able to attend to these stimuli in an appropriate way. While Sects. 2.1 and 2.2 primarily motivate the research question of storing attentional information (i.e., the focusing structure), this section motivates the question of updating focus of attention with respect to linguistic stimuli. In terms of cognitive models, stimuli are attentively encoded into working memory. Similarly, one of the aims of this paper is to provide an algorithm for encoding linguistic stimuli into the focusing structure (i.e., the focus tree).

3 Foundations of a new model of attentional state

This section introduces the idea underlying a new model of attentional state, based on the focus tree, that encapsulates the important properties of both cognitive models of working memory and the focusing structure in the theory of discourse structure. There are two main guidelines for developing a model of attentional state: (i) storage of attentional information should reflect the principle that available entities are distinguished with respect to their access status, and (ii) the model should include also processing of attentional information.

(i) One of the most obvious principles that can be found in both Oberauer and Lange's cognitive model of working memory and Grosz and Sidner's theory of discourse structure is that available entities are distinguished with respect to their access status. However, in the cognitive model, entities (i.e., long-term memory representations) are organized in a network, without further specifications of its actual topology that would reflect this principle. On the other hand, in the theory of discourse structure, the hierarchical organization is strictly defined. Grosz and Sidner use a stack structure to represent the dynamic nature of the attentional state. The stacking and manipulating of focus spaces reflects the relative salience of the entities in each space [23, p. 180]. We use the concept of the focus tree to define a topological structure of attentional information that reflects the principle that available entities are distinguished with respect to their access status.

Fig. 3 A simple focus tree



(ii) Processing of attentional information remained underspecified both in cognitive models of working memory and in the theory of discourse structure. In the neurosciences, the working memory is commonly conceptualized as a system for simultaneous storage and processing of information. However, while much research was devoted to the storage function, less attention was devoted to the processing aspect (cf. Sect. 2.1). On the other hand, related to the theory of discourse structure, we briefly discussed advantages of the focus tree over the focus space stack with respect to topic management (cf. Sect. 2.2). The focus tree enables a more flexible topic management, and both constrains and enables prediction of what is likely to be talked about next. However, although these observations seem promising, the research issue that remains to be addressed is to introduce concrete algorithms for processing of attentional information.

These two guidelines are addressed in more detail in Sects. 4 and 5, respectively. The rest of this section presents foundations of a new model of attentional state based on the focus tree.

The focus tree encapsulates structural relations between focus spaces that correspond to the dominance hierarchy of discourse segments. In the example given in Fig. 2, discourse segment DS1 dominates discourse segments DS2 and DS3. Therefore, we introduce structural relations according to which focus space FS1 "dominates" focus spaces FS2 and FS3. A simple focus tree that corresponds to the given example is shown in Fig. 3.

The focus tree also preserves the idea of recursive development of the focus space stack. In the focus tree, the dynamical nature of attentional state is represented by placing the focus of attention on one of the nodes in the focus tree. This comparison is illustrated in Fig. 4. It shows the sequence of states of the focus space stack (introduced by Grosz and Sidner) and the corresponding focus of attention in the focus tree (introduced in our approach) during the processing of the discourse segments in the given example. Before the segments are processed (State 1), the focus stack is empty and the focus of attention is not placed on any of the nodes in the focus tree. When segment DS1 is being processed (State 2), focus space FS1 is positioned on the stack. In the focus tree, it is represented by placing the focus of attention on node FS1 (the node is represented in oval). Processing of segment DS2 (State 3) pushes focus space FS2 on the top of the stack. Corresponding to the fact that this focus space is the most salient at the moment, the focus of attention is shifted on node FS2 in the focus tree. After segment DS2 has been processed (State 4), focus space FS2 has been

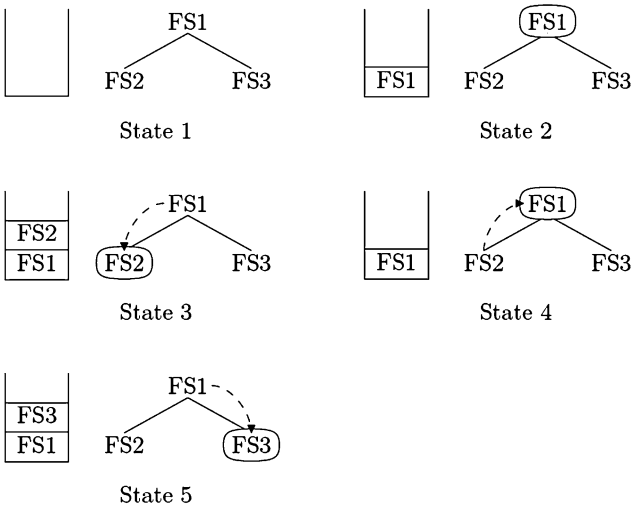


Fig. 4 Comparison between focus stack and focus tree

popped from the stack. Focus space FS1 is now on the top of the stack and the focus of attention is again placed on node FS1 in the focus tree. Processing of segment DS3 (State 5) gives rise to focus space FS3—pushing focus space FS3 on the stack is represented by placing the focus of attention on node FS3 in the focus tree.

For a given interaction domain, the focus tree is determined beforehand and fixed. It can be summarized that, for a given discourse structure, the focus tree encapsulates the set of all possible states of the focus space stack. States of the focus space stack are denoted by the position of the focus of attention in the focus tree. A node that carries the current focus of attention corresponds to a focus space that is placed on the top of the stack, its parent node corresponds to a focus space that is placed below, and so on—all ancestor nodes correspond to focus spaces contained in the stack, where the root node of the focus tree corresponds to a focus space placed on the bottom of the stack. However, whereas the focus space stack represents a collection of available focus spaces at the given moment and, thus, can be dynamically changed, the focus tree is determined beforehand and fixed. Therefore, to construct a focus tree, all entities that may become salient during the processing of the discourse segments must be known in advance. This also implies that the number of nodes in the focus tree is limited. We find a justification for these implications in cognitive models of working memory. It is widely accepted that the capacity of working memory is limited, e.g., in terms of Oberauer and Lange, the region of direct access in long-term memory holds a limited number of activated representations that are available for ongoing cognitive, goal-directed processes. In our approach, the focus tree contains entities that become salient during the processing of discourse segments. There is a clear analogy—the focus tree represents the region of direct access in long-term memory and, thus, has a limited number of nodes. In addition, following Oberauer and Lange, representations in the region of direct access are bound temporarily to a common cognitive coordinate system. Therefore, it is justifiable to assume that representations are known in advance. One may argue that, from the speaker's point of view, representations are part of long-term memory and, thus, inherently known in advance. However, it does not hold for a dialogue system. A dialogue system may a priori recognize as relevant for the interaction only those entities that are obviously present in the given context of interaction (e.g., a spatial context, etc.). Another implication of the analogy between the region of direct access in long-term memory and the focus tree is that a node in the focus tree represents a single long-memory representation. This is in line with Oberauer and Lange who note that a single representation is selected to be in the focus of attention—at every moment of interaction, the current focus of attention is placed on exactly one node in the focus tree.

4 Storage function: constructing a focus tree

We illustrate construction of a focus tree for a concrete interaction domain taken from the NIMITEK corpus—the Tangram puzzle. After inspection of subjects' commands from the corpus, we differentiate among four *focus classes* whose instances form attentional information. They are given in the following list, starting from the most general focus class and ending with the most specific:

- Task focus—Focus instances contained in this class relate to the tasks given to the subjects in the Wizard-of-Oz study (e.g., the Tangram puzzle, the Tower of Hanoi puzzle, the Grid puzzle, etc.).
- Object focus—Focus instances contained in this class relate to graphical objects that can be manipulated in the given tasks (e.g., Tangram pieces, disks in the Tower of Hanoi puzzle, tiles in the Grid puzzle, etc.).
- Action focus—Focus instances contained in this class relate to actions that can be performed over selected objects. E.g., for the Tangram puzzle, there are two focus instances contained in this focus class that relate to actions of translation and rotation, respectively.
- Direction focus—Focus instances contained in this class relate to further specification of actions that can be performed over selected objects. E.g., in the context of the Tangram puzzle, for the action of translation there are four focus instances that relate to direction (up, down, left and right), and for the action of rotation there are two focus instances that relate to direction (clockwise and counterclockwise).

These focus classes are interrelated—an instance of a more specific focus class is a sub-focus of an instance of

the immediately preceding more general focus class. We shortly explain the sub-focus relation: focus instance g_1 is a sub-focus of focus instance g_2 if focus instance g_1 cannot become salient in the given dialogue without g_2 being also salient in the same moment (g_2 may be explicitly mentioned in the dialogue or implicitly introduced into the dialogue context). For example, a focus instance representing an action over a Tangram piece is a sub-focus of a focus instance representing that Tangram piece because we have to specify a piece before we can perform an action over it. It is important to note that a sub-focus relation is a kind of semantic relation and not determined with the syntactical structure of users' utterances. Due to this property, it is possible, as we discuss below, to utilize sub-focus relations to process the user's commands of different syntactic forms.

Sub-focus relations are illustrated in the simplified focus tree for the Tangram puzzle given in Fig. 5. Each instance of the focus classes is represented by a node in the focus tree. Each node, except the root node, represents a sub-focus of its parent node. The root node represents the most general focus instance. Nodes at the same level of the tree belong to the same focus class. For the purpose of easier representation and without loss of generality, we reduce the number of Tangram pieces to two: the triangle (△) and the square (□). It means that we show only a part of the "bigger" focus tree including all seven Tangram pieces. Table 1 provides short descriptions of all focus instances in this focus tree.

At every moment of interaction, the current focus of attention is represented by exactly one node in the focus tree. Mapping of an ensuing user's command onto the focus tree is performed with respect to the position of the current focus of attention. Also, the user's command may change the focus of attention. This is considered in the next section in more detail.

Table 1 Focus instances in the simplified focus tree for the Tangram puzzle

Focus instance	Focus class	Description of focus instance
tangram	task	Tangram puzzle
△	object	triangle
$tran_1$	action	translation of △
↑	direction	upward translation of △
←	direction	leftward translation of △
↓	direction	downward translation of △
→	direction	rightward translation of △
rot_1	action	rotation of △
↶	direction	counterclockwise rotation of △
↷	direction	clockwise rotation of △
□	object	square
$tran_2$	action	translation of □
⇑	direction	upward translation of □
⇐	direction	leftward translation of □
⇓	direction	downward translation of □
⇒	direction	rightward translation of □
rot_2	action	rotation of □
⟲	direction	counterclockwise rotation of □
⟳	direction	clockwise rotation of □

5 Processing function: transition of the focus of attention

For the purpose of defining the algorithm for transition of the focus of attention, we introduce the following abbreviations:

- f—a focus stimulus,
- r—a negative reinforcement stimulus,
- g—a focus instance (i.e., a node) in the focus tree that represents f, which we write as: $g \leftrightarrow f$,
- $R(g)$—rank of a node g, defined as the length of the path from the root node to node g.

To give an example: the command "don't move, but rotate to the left" includes two focus stimuli, f_1 = "rotate" and f_2 = "to the left", that belong to the action and direction focus classes, respectively, and a negative reinforcement stimulus r_1 = "don't move". Observed out of the interaction context, focus stimulus f_1 can be represented by nodes $\{rot_1, rot_2\}$, while focus stimulus f_2 can be represented by nodes $\{\leftarrow, \curvearrowleft, \Leftarrow, \circlearrowleft\}$ in the focus tree given in Fig. 5.

In addition, we make the assumption that all focus stimuli contained in a command represent focus instances that belong to different focus classes. This assumption might seem too strong. For example, the command "move triangle and square rightwards" contains two focus stimuli, "triangle" and "square", that relate to the object focus class. However, such commands may be appropriately divided in a sequence of commands whose all focus stimuli relate to different focus classes (e.g., "move triangle rightwards", "move square rightwards") before they are further processed. In other words, the issue of compositionality may be addressed, when necessary, independently of the proposed algorithm for transition of the focus of attention. Therefore, under such conditions, the aforementioned assumption appears justifiable.

The algorithm for transition of the focus of attention introduced in this section is of a general nature and illustrated for spontaneously uttered commands from the NIMITEK corpus. As mentioned above, mapping of a command onto the focus tree is performed with respect to the position of the current focus of attention. Let g_c be the node representing the current focus of attention, and let C be a command that comprises the following focus stimuli $f_1, f_2, \ldots, f_n$, where:

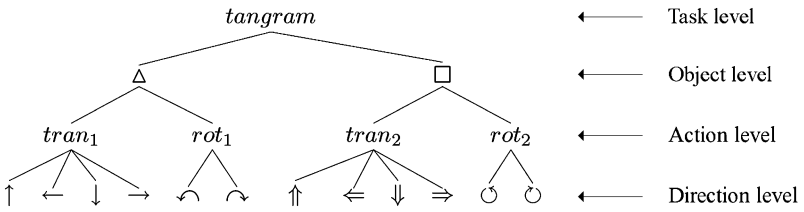


Fig. 5 The simplified focus tree for the Tangram puzzle

$n \geq 1$, all focus stimuli relate to different focus classes, f_1 is the most general focus stimulus, and f_n is the most specific focus stimulus in command C. In addition, command C may optionally contain a set of negative reinforcement stimuli $\{r_1, r_2, \ldots, r_k\}$. If the set of negative reinforcement stimuli is not empty, it signals potential problems in communication, and, thus, the current focus of attention is placed on the root node of the focus tree before the algorithm is applied. The underlying idea of the algorithm could be summarized as follows. In the first step, a temporary focus of attention is positioned on a node that represents the most general focus stimulus from command C, i.e., f_1. There can be more than one node satisfying this condition. Thus, the selection of one of them is determined by the position of node g_c, as discussed below. In succeeding steps, a temporary focus of attention is iteratively transited over nodes that represent focus stimuli $f_2, f_3, \ldots, f_n$, following the rule that, for all i, j, where $1 \leq i < j \leq n$, the node representing the focus stimulus f_j is a descendant of the node representing focus stimulus f_i. The new focus of attention is placed on a node representing the most specific focus stimulus from command C, i.e., f_n.

Generally, for a given current focus of attention, command C can be mapped to different sets of nodes in the focus tree. In each of these steps there might be more candidate nodes for a temporary focus of attention. The transition of a temporary focus of attention may branch with each focus stimulus from command C, consequently resulting in more candidate nodes for the new focus of attention. One of these candidates is to be selected to represent the new focus of attention. It is a matter of dialogue context and dialogue strategy which candidate node will be selected (examples of dialogue strategies are given in [19, 21]).

In order to describe transition of the focus of attention in more detail, we distinguish between two cases. The first case is when each focus stimulus f_i from command C can be represented by some of the descendant nodes of node g_c representing the current focus of attention, or by node g_c, i.e.:

$$(\forall f_i \in C)(\exists g_j \in \textit{descendant-or-self}(g_c))(g_j \leftrightarrow f_i) \qquad (1)$$

In this case, the mapping of command C is restricted to the sub-tree determined by node g_c as its root node. In the first step, candidate nodes for representing the most general focus stimulus f_1 are selected only among nodes in this sub-tree. Other nodes in the focus tree are not taken into consideration. Since selection of a new temporary focus of attention in succeeding steps is always limited to the sub-tree determined by the current temporary focus, final candidate nodes for the new focus of attention are selected among descendant nodes of node g_c (including also node g_c).

The second case is when not all focus stimuli from command C can be represented by some of the descendant nodes of node g_c, or by node g_c, i.e.:

$$(\exists f_i \in C)(\forall g_j \in \textit{descendant-or-self}(g_c))\neg(g_j \leftrightarrow f_i) \qquad (2)$$

In this case, a temporary focus of attention is first placed on the *closest* antecedent node g_{temp} of node g_c that satisfies the condition that each focus stimulus from command C can be represented by some of the nodes from the sub-tree determined by node g_{temp} as its root node, i.e.:

$$\begin{aligned} & g_{temp} \in \textit{antecedent}(g_c) \\ \wedge\ & R(g_{temp}) = max\{R(g_i) | g_i \in \textit{antecedent}(g_c) \\ \wedge\ & (\forall f_i \in C)(\exists g_j \in \textit{descendant-or-self}(g_i))(g_j \leftrightarrow f_i)\} \end{aligned} \qquad (3)$$

Command C is then mapped within the sub-tree determined by node g_{temp}, as described in the first case. Both these cases are encapsulated in the recursive algorithm given in Fig. 6. For a given command C, the first call of this algorithm is to be realized with the following arguments: the node representing current focus of attention g_c, and the most general focus stimulus from command C, i.e., f_1, as showed in Fig. 7. After all recursive calls of this algorithm are finished, candidate nodes for the new focus of attention are accumulated in the set variable `focus_candidates`.

At this point of the algorithm, negative reinforcement stimuli—if any—are taken into account in order to filter the set variable `focus_candidates`. All nodes from this set that can be related, or whose ancestors in the focus tree can be related, to some of negative reinforcement stimuli $\{r_1, r_2, \ldots, r_k\}$ are removed from the set. If the set is left empty after this filtering, it signals that command C is semantically irregular. A command may be semantically irregular in two cases. First, if it includes at least two focus stimuli f_i and f_j for which holds: $i \leq j$ and the focus instance that represents f_j is not a sub-focus of the focus instance that represents f_i, e.g., "translate clockwise". And second, if it includes a focus stimulus and a negative reinforcement stimulus that relate to the same focus instance, e.g., "rotate but don't rotate".

Otherwise, if the set of candidate nodes is not left empty, one of them is selected to represent the new focus of attention. In Sect. 2.2, we discussed that the focus tree should not contain task structure or intentional relationships. In our approach, the focus tree model is intentionally separated from a reasoning tool that takes information about task structure and intentional relationships into account. This separation is in line with our intention to introduce an attentional state model of a general nature, while such a reasoning tool is task-oriented. Therefore, in the algorithm given in Fig. 7, function `choose_from` represents an interface between the module that implements the concept of the focus tree and an external reasoning tool that takes information about task structure and intentional relationships into account. Here, we abstract away from aspects of this interface that are task-oriented and, thus, less important for the discussion in this paper. In Sect. 6, we discuss a prototype system that also includes a task-oriented reasoning tool.

```
procedure get_focus_candidates(g_arg, f_i)
begin
  if (f_i = f_1) ∧ ({r_i,...,r_k} ≠ ∅)
  then
    g_arg := root_node;

  if (∀f_j, i ≤ j ≤ n)(∃g_l ∈ des_or_self(g_arg))(g_l ↔ f_j)
  then
  begin
    S := {g | g ∈ des_or_self(g_arg) ∧ g ↔ f_i};
    for each g ∈ S do
      if (f_i = f_n)
      then
        focus_candidates.add(g)
      else
        get_focus_candidates(g, f_{i+1})
  end
  else
    if (f_i = f_1) ∧ (R(g_arg) > 0)
    then
      get_focus_candidates(parent(g_arg), f_i)
    else
      Exit();
end
```

Fig. 6 Algorithm for identifying candidate nodes for the new focus of attention. Function des_or_self(g_{arg}) returns a set that contains node g_{arg} and all its descendant in the focus tree

Let us illustrate these algorithms for the following sequence of commands:

(C_1:) triangle to right but don't rotate ... (C_2:) now to right rotate ... (C_3:) to left ... (C_4:) upwards ...

From the user's point of view, the interaction in this sequence could be summarized as follows. In the first command, the user selects the triangle. Afterwards, he assumes that the selected object is a part of the shared knowledge between him and the system. Thus, until the end of the given

```
procedure change_focus()
begin
  focus_candidates.empty();
  get_focus_candidates(g_c, f_1);
  R := {r_i,...,r_k};

  if (R≠∅) then
    for each g ∈ focus_candidates do
      if (R ∩ ancestor_or_self(g) ≠∅)
        focus_candidates.remove(g);

  if (focus_candidates≠∅)
  then
    g_c := choose_from(focus_candidates)
  else
    report_irregular_command();
end
```

Fig. 7 Transition of the focus of attention. Function `choose_from` represents an interface between the module that implements the concept of the focus tree and an external reasoning tool that takes information about task structure and intentional relationships into account

sequence, he instructs only actions that should be performed over the selected object, without explicitly referring to the selected object itself. The introduced algorithm for transition of the focus of attention supports system's decision making processes when it is confronted with such user inputs.

At the beginning of this sequence, the current focus of attention is placed on the root node of the focus tree. Relevant parts of the focus tree are represented in Fig. 8 for commands C_1 and C_2, and in Fig. 9 for commands C_3 and C_4. Changes of a temporary focus of attention are marked with dashed arrows. Nodes representing the temporary focus of attention during the mapping of a command are positioned in ovals, while nodes representing the new focus of attention after a command has been mapped are positioned in boxes.

Command C_1 contains two focus stimuli: f_1 = "triangle" and f_2 = "to right", and a negative reinforcement stimulus "don't rotate". In the given focus tree, focus stimulus f_1 can be represented only by node △, while focus stimulus f_2 can be represented by four different nodes {→, ↷, ⇒, ↻} (cf. Table 1). For the starting focus of attention placed on the root node, the condition (1) is satisfied, i.e., all focus stimuli from command C_1 can be represented by some of the descendant nodes of the node representing the current focus of attention. Therefore, all changes of the temporary focus of attention are directed towards more specific focus instances. In the first iteration, when focus stimulus f_1 is being mapped, the temporary focus of attention is placed on node △. In the second iteration, mapping of the focus stimulus f_2 is restricted to the sub-tree determined by node △ as its root node. There are just two nodes in this sub-tree that are candidates to represent focus stimulus f_2: {→, ↷}. Since there are no more focus stimuli in command C_1 to be mapped, one of these nodes should be selected to represent the new

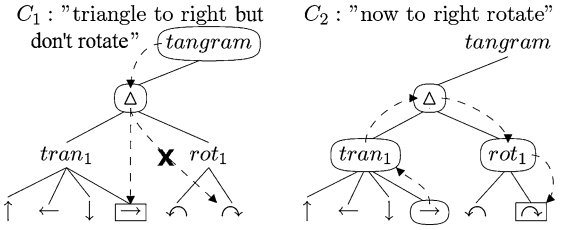


Fig. 8 Transition of the focus of attention for commands C_1 and C_2

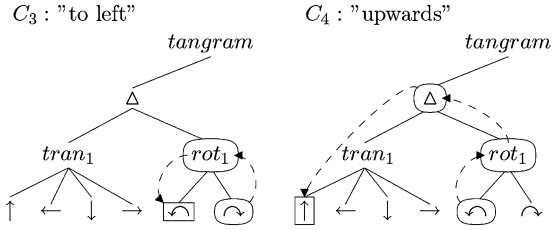


Fig. 9 Transition of the focus of attention for commands C_3 and C_4

focus of attention. Nevertheless, the algorithm now takes the negative reinforcement stimulus "don't rotate" into account. This stimulus relates to node rot_1, which is the parent of node $\curvearrowright$, so node $\curvearrowright$ is removed from the set of candidate nodes. Therefore, node $\rightarrow$ is selected to represent the new focus of attention (cf. left part of Fig. 8).

Command C_2 contains two focus stimuli: $f_3 =$ "to right" and $f_4 =$ "rotate". In the focus tree, focus stimulus f_3 can be represented by nodes $\{\rightarrow, \curvearrowright, \Rightarrow, \circlearrowright\}$, while focus stimulus f_4 can be represented by nodes $\{rot_1, rot_2\}$. However, these focus stimuli cannot be mapped immediately. Keeping in mind that the current focus of attention is placed on node $\rightarrow$ after command C_1 has been processed, command C_2 satisfies condition (2), i.e., not all focus stimuli from command C_2 can be represented by some of the descendant nodes of node $\rightarrow$. Thus, the temporary focus of attention should be iteratively moved towards higher levels of the focus tree until we reach a node whose descendant nodes can represent all focus stimuli from the command, i.e., a node that satisfies condition (3). So, the temporary focus of attention is first placed on the parent node of the node representing the current focus of attention—node $tran_1$. Since condition (3) is still not satisfied, the temporary focus of attention is moved one level higher in the focus tree and placed on node $\triangle$. Now, when condition (3) is satisfied, focus stimuli f_3 and f_4 can be mapped within the sub-tree determined by node $\triangle$ in a similar way as focus stimuli in command C_1 have been mapped. It is important to note that, although focus stimulus f_3 comes before focus stimulus f_4 in the utterance, focus stimulus f_4 is first mapped because it is more general. When focus stimulus f_4 is being mapped, the temporary focus of attention is placed on node rot_1. When focus stimulus f_3 is being mapped, the temporary focus of attention is placed on node $\curvearrowright$. Since now all focus stimuli contained in command C_2 have been mapped, the new focus of attention is placed on this node (cf. right part of Fig. 8).

Mapping of commands C_3 and C_4 is performed in the same manner as command C_2. Transition of the focus of attention is illustrated in Fig. 9.

Illustrating the introduced algorithms, we considered focus stimuli and negative reinforcement stimuli that were contained in a command, but we did not explain how these stimuli were extracted from the command, and how focus stimuli were ordered from the most general to the most specific. Moubaiddin and Obeid [40, p. 149] argue—in line with our discussion in [17, pp. 39–42] and [18]—that users often produce incomplete and grammatically ill-formed utterances. Therefore, a language interface must be able to cope with such dialogue phenomena. They suggest keyword spotting as a possible solution for robust parsing of such utterances. We build upon this approach as follows. To each node in the focus tree, a set of phrases (i.e., stimuli) that represent its focus instance is assigned. For example, the focus instance that is represented by node $\square$ in the focus tree for the Tangram puzzle may be correlated with the following focus stimuli: {square, yellow square, quadrangle, ...} (in German: {Quadrat, gelbes Viereck, Viereck, ...}). When mapping users' commands onto the focus tree, the system takes as input a textual version of the command delivered by the speech recognizer. The focus stimuli and negative reinforcement stimuli are then automatically derived from a given command, i.e., the system detects phrases that (might) relate to certain focus instances. After all focus stimuli are related to focus instances (i.e., nodes in the focus tree), they are order from the most general to the most specific according to the following rule. Focus stimulus f_i is more general than focus stimulus f_j if the node that represents f_i is at the higher level of the focus tree then the node that represents f_j. We recall that we introduced the assumption that all focus stimuli contained in a command relate to focus instances that belong to different focus classes, i.e., to focus instances that are positioned at different levels of the focus tree. In addition, sets of stimuli assigned to different focus instances are not necessary disjoint sets. For example, discussing how command $C_1 =$ "triangle to right" is processed, we stated that the focus stimulus "to right" could be assigned to four different nodes in the given focus tree, and illustrated how the introduced algorithm copes with such a situation.

6 Implementation example

This section illustrates the implementation of the introduced model of attentional state in the NIMITEK prototype spoken dialogue system for supporting users while they solve problems in a graphics system. The dedicated task is the 3-disk version of the Tower-of-Hanoi puzzle (cf. Figs. 10 and 11).

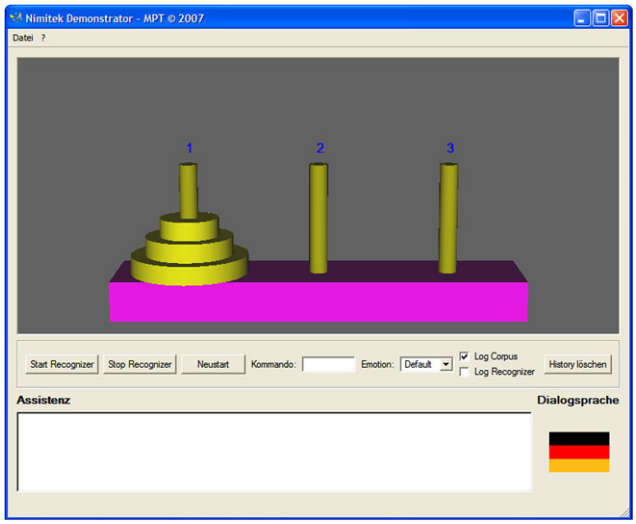


Fig. 10 The 3-disks version of the Tower of Hanoi puzzle—screen display of the NIMITEK prototype system represents the initial state of the puzzle

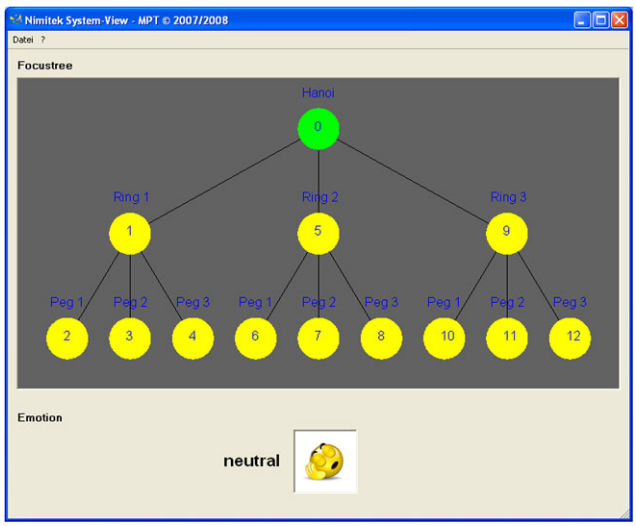


Fig. 11 The focus tree for the 3-disks version of the Tower of Hanoi puzzle—screen display of the NIMITEK prototype system corresponds to the initial state of the puzzle

The adaptive dialogue strategy of the NIMITEK system is introduced in more detail by Gnjatović and Rösner [17, 19]. Here, we illustrate some aspects of the implementation that are particularly related to the model of attentional state.

We give an analysis of an actual dialogue between the user and the prototype system that took place during the testing of the system. The dialogue fragment is given in Fig. 12. Utterances produced by the user and the system are written in italic, and descriptions of non-verbal actions performed by the system (e.g., moving a disk, etc.) are underlined.

The transition of the focus of attention is performed according to the introduced algorithm as this dialogue fragment unfolds. At the start of the dialogue fragment, the focus

$User_1$:	*The smallest disk* ... (*Den kleinsten Ring* ...)
$System_2$:	selects the smallest disk
$User_3$:	... *on the three.* (... *auf die Drei.*)
$System_4$:	puts the selected disk on the third peg
$User_5$:	*I would like to move the middle plate to position two.* (*Ich möchte gerne die mittlere Scheibe auf Position Zwei schieben.*)
$System_6$:	moves the middle disk to the second peg
$User_7$:	*The next disk* ... (*Den nächsten Ring* ...)
$System_8$:	selects the biggest disk
$User_9$:	... *on the three.* (... *auf die Drei.*)
$System_{10}$:	*This move is not possible. You can only move a disk from the top. You cannot place a larger disk onto a smaller one.* (*Dieser Schritt ist nicht möglich. Sie können nur einen oben liegenden Ring wählen. Sie dürfen keinen Ring auf einen kleineren legen.*)
$User_{11}$:	*Help.* (*Hilfe.*)
$System_{12}$:	*Which peg do you want to select?* (*Welchen Stab möchten Sie auswählen?*)
$User_{13}$:	*I said help!* (*Ich sagte Hilfe!*)
$System_{14}$:	*The last move drew you back from the solution. I propose that you select another disk.* (*Der letzte Schritt führte von der Lösung weg. Ich schlage vor, dass Sie einen anderen Ring auswählen.*)

Fig. 12 Dialogue between the user and the NIMITEK prototype system

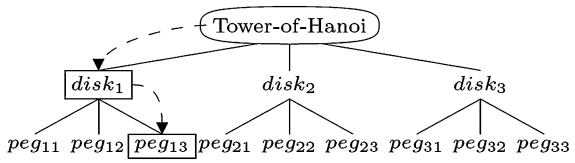


Fig. 13 Transition of the focus of attention for commands $User_1$ and $User_3$

of attention is placed on the root node of the focus tree (cf. Fig. 13). The system recognizes that command $User_1$ contains only one focus stimulus "the smallest disk" that can be represented by node $disk_1$. This node is a descendant of the node representing the current focus of attention. Therefore, the system selects the smallest disk on the graphical display ($System_2$) and, simultaneously, places the new focus of attention on node $disk_1$. Command $User_3$ also contains only one focus stimulus "on the three" that relates to the third peg on the graphical display. However, this focus stimulus may be represented by three different nodes in the focus tree: peg_{13}, peg_{23} and peg_{33}. Following the introduced algorithm, node peg_{13} should be selected, since it is a descendant node of the node representing the current focus of attention. Thus, the system places the selected smallest disk on the third peg on the graphical display ($System_4$), and the new focus of attention on node peg_{13}. The transition of the focus of attention for commands $User_1$ and $User_3$ is illustrated in Fig. 13.

Command $User_5$ contains words that cannot be recognized by the speech recognition module. The textual version of this command as it is recognized is: *<not recognized> the middle disk <not recognized> position two*. Still, the

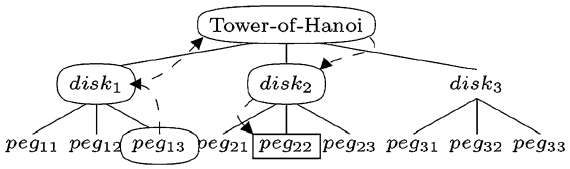


Fig. 14 Transition of the focus of attention for command User_5

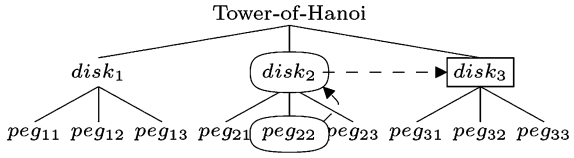


Fig. 15 Transition of the focus of attention for command User_7

system recognizes all focus stimuli that are contained in the command. The first focus stimulus, "the middle disk", may be represented by node $disk_2$. The second focus stimulus, "position two", relates to the second peg on the graphical display and may be represented by three nodes: peg_{12}, peg_{22} and peg_{32}. However, these focus stimuli cannot be represented by descendant nodes of the node caring the current focus of attention, i.e., node peg_{13}. According to the introduced algorithm, the focus of attention is iteratively moved towards higher levels of the focus tree to the closest antecedent node whose descendant nodes can represent all focus stimuli from the command—the root node. Then, similarly as explained for first two user commands, the focus of attention is first placed on node $disk_2$ and then to its child node peg_{22}. On the graphical display, the system selects the middle disk and places it on the second peg (System_6). The transition of the focus of attention for command User_5 is illustrated in Fig. 14.

Command User_7 is context-dependent. It contains a nominal phrase that cannot be uniquely related to a node in the focus tree. In other words, a disk should be selected, but it is not explicitly specified which disk should be selected. Therefore, the contextual information should be taken into account. The current focus of attention and the structure of the focus tree enable the system to process such a command. Intuitively, it is clear that the previously selected disk was the middle disk, and that the phrase "the next disk" relates to the biggest disk. The processing of this command can be formalized in a more general manner. The system traverses the focus tree in preorder, starting from the node that represents the current focus of attention. The new focus of attention is placed on the first node that satisfies the following conditions: (i) it belongs to the object focus class, and (ii) the selected node is not the starting node. The simplified transition of the focus of attention is illustrated in Fig. 15. After processing command User_7, the new focus of attention is placed on node $disk_3$. On the graphical display, the system selects the biggest disk (System_8).

In command User_9, the user instructs the system to place the biggest ring to the third peg, which is an illegal command. Therefore, the system informs the user that it is not allowed to place a larger disk onto a smaller one (System_{10}). In command User_{11}, the user explicitly asks for help. Following its dialogue strategy, the system assumes that the user has a problem to formulate a valid command. The actual support intervention is determined by the current focus of attention. The system tries to help the user to formulate a command that will place the focus of attention on one of the descendant nodes of the node carrying the current focus of attention. Since the current focus of attention is positioned on node $disk_3$, the system asks the user on which peg does he want to place the selected ring (System_{12}). When the user repeats the request for help in command User_{13}, the system provides support of higher intensity determined by the state of the puzzle. It proposes the user to select another disk (System_{14}).

7 Discussion

The introduced model of attentional state in HMI differs from existing models in several aspects.

(i) Processing aspect of updating focus of attention. Processing of attentional information remained underspecified both in cognitive models of working memory [7, p. 177] and in the theory of discourse structure [23, p. 202] (including also several conceptualizations of the focus tree [31, p. 137]). In contrast to them, the model of attentional state introduced in this paper addresses the research questions of storage and processing of attentional information in an integrated manner. With respect to the storage of information, we introduced the focus tree—a topological structure of attentional information. With respect to the processing of information, we provided algorithms for transition of focus of attention, and illustrate them for concrete interaction domains.

(ii) Analytical and generative aspects of the dialogue behavior. While existing models of attentional state are often focused either on interpretation (cf. [23]) or on generation (cf. [34]) of utterances in discourse, the model of attentional state introduced in this work is intended to be used both for interpretation of user commands and generation of system dialogue acts. In the dialogue fragment provided in Fig. 12, the system both interprets the user's commands (cf. system response in System_2, System_4, System_6, System_8) and generates coherent dialogue acts (cf. system response in System_{10}, System_{12}, System_{14}).

(iii) Aspects of generalizability. The model of attentional state was illustrated for the Tangram and Tower of Hanoi puzzles. A question that arises is to what extent can this approach be generalized. We discuss this question from two

points of view: the engineering point of view and the linguistic point of view.

The engineering point of view considers primarily implementation aspects. The proposed modeling method and algorithms are not a priori related to some specific predefined task. The introduced algorithms are independent of the structure of the focus tree and of the content of the phrasal lexicon. For a given task (e.g., Tangram puzzle, Tower of Hanoi puzzle, etc.), the structure of the focus tree and the sets of stimuli that are assigned to focus instances are defined in input XML files, independently of the implementation of the algorithms. This means that the implementation of the proposed model of attentional state within the dialogue management module in the NIMITEK prototype system is independent of:

- changes of the structure of the focus tree (e.g., a change from the Tangram puzzle to the Tower of Hanoi puzzle, etc.),
- changes of the vocabulary (e.g., changing the size of the vocabulary by extending or redefining sets of phrases, changing the language of the vocabulary by translating phrases from German into English, etc.).

These changes do not require a change in the implementation of the model, but just a redefinition of input XML files. From the engineering point of view, this gives a relatively high level of generalizability of the proposed model—the given task can be relatively easy redefined or extended.

The linguistic point of view considers the question to which types of dialogue can this approach be applied. In this paper, we concentrated on spoken human-machine interaction in the specific case where some kind of display with a graphical interface is involved. A graphical interface that is present in interaction with the dialogue system NIMITEK represents a non-linguistic context shared between the user and the system that influences the user's focusing processes. This suggests that our approach to modeling attentional information is appropriate for the class of spoken dialogue systems that are intended to control a subclass of graphical user interfaces, e.g., manipulating with graphical entities represented on the display, controlling graphical menus, solving graphically-based tasks including spatial reasoning, etc. Nevertheless, the introduced model is not restricted only to this class of spoken systems. It is also applicable to some interaction domains that include verbal-only interfaces. For example, the introduced model of attentional state was also implemented within the adaptive dialogue manager in the system Contact [21]. This system is primarily intended to be used by the visually impaired—it reads aloud textual contents (e.g., news, articles, etc.) from various newspapers and web sites over the telephone line. In contrast to the NIMITEK system, the dialogue system Contact does not include a graphical interface that would help to establish a non-verbal context. However, entities that become salient during the interaction (e.g., newspapers, sections, articles, etc.) can still be organized in a focus tree, and sub-focus relations between them are also part of the common knowledge: e.g., each newspaper contains one or more sections, each section contains one or more articles, etc. Based on this common knowledge, the user can mentally construct a "quasipictorial" representation of the focus tree used by the system. This mental representation serves as a non-linguistic context shared between him and the system. We recall that this explanation is in line with recent findings that spatial attention and mental images are closely interrelated [1, 22, 41, 54].

Finally, for the purpose of completeness, we state that the introduced approach to modeling attentional information is not limited only to verbally uttered commands. It supports also non-verbal dialogue acts produced by the user (e.g., using a mouse or a keyboard, etc.). Such non-verbal actions may also change the attentional state. For example, if the user was allowed to use a mouse to select a graphical piece represented on the screen, he would thereby unambiguously specify that the current focus of attention should be placed on the node in the focus tree that represents the selected piece.

It can be summarized that two basic requirements are needed to apply the proposed modeling methodology. First, all "relevant" focus instances must be known in advance. And second, it must be possible to define sub-focus relations between them. These requirements may appear to be too restrictive for a general case of unrestricted dialogue. However, we primarily consider task-oriented HMI. In such "practical" interaction domains, the aforementioned requirements appear to be adequate. This was already discussed in Sect. 3.

8 Conclusion

This paper introduced a new model of attentional state in task-oriented HMI. It integrates three lines of research: (i) neurocognitive understanding of the focus of attention in working memory, (ii) the notion of attention related to the theory of discourse structure in the field of computational linguistics, and (iii) investigation of a corpus that comprises recordings of spontaneous speech-based HMI. The underlying idea was to make a computationally appropriate representation of attentional information that imitates the function of a focus of attention in human perception. To the extent that the model is computationally appropriate, the discussion was concentrated to the research problem of robust automatic processing of different syntactic forms of spontaneously produced users' commands with no explicit syntactic expectations.

The introduced model differs from existing models in several aspects. First, it addresses the research questions of storage and processing of attentional information in an integrated manner. Second, the model is intended to be used both for interpretation of user commands and generation of system dialogue acts. And third, the proposed modeling method is intended to be sufficiently general both from the engineering and linguistic points of view. Finally, the paper discussed an implementation of the introduced model within a prototype spoken dialogue system and gives an analysis of an actual dialogue fragment between the user and the system that took place during the testing of the system.

The introduced model has an important role in several lines of our research. These include defining and implementing adaptive dialogue strategies aimed to support the user to overcome problems with the interface language, and to handle miscommunication on the conversational, the intentional, and the signal levels. One of the planned research directions is exploring possibilities to apply this conceptualization in the scope of therapeutic HMI with language-impaired patients.

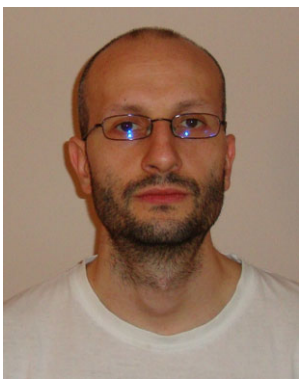
Acknowledgements The presented study is performed as part of the projects "Design of Robots as Assistive Technology for the Treatment of Children with Developmental Disorders" (III44008) and "Development of Dialogue Systems for Serbian and Other South Slavic Languages" (TR32035), funded by the Ministry of Education and Science of the Republic of Serbia. The responsibility for the content of this paper lies with the authors.

References

1. Awh E, Jonides J (2001) Overlapping mechanisms of attention and spatial working memory. Trends Cogn Sci 5(3):119–126
2. Atkinson RC, Shiffrin RM (1968) Human memory: A proposed system and its control processes. In: Spence KW, Spence JT (eds) The psychology of learning and motivation: Advances in research and theory, vol 2. Academic Press, New York, pp 89–195
3. Baddeley AD (1993) Working memory or working attention? In: Baddeley AD, Weiskrantz L (eds) Attention: Selection, awareness, and control. A tribute to Donald Broadbent. Oxford University Press, New York, pp 152–170
4. Baddeley AD, Hitch GJ (1974) Working memory. In: Bower GH (ed) The psychology of learning and motivation: Advances in research and theory, vol 8. Academic Press, New York, pp 47–89
5. Baddeley AD, Logie RH (1999) Working memory: The multiple component model. In: Miyake A, Shah P (eds) Models of working memory. Cambridge University Press, New York, pp 28–61
6. Bledowski C, Rahm B, Rowe BR (2009) What 'works' in working memory? Separate systems for selection and updating of critical information. J Neurosci 29:13735–13741
7. Bledowski C, Kaiser J, Rahm B (2010) Basic operations in working memory: Contributions from functional imaging studies. Behav Brain Res 214(2):172–179
8. Broadbent DE (1958) Perception and communication. Pergamon Press, New York
9. Campbell N (2006) On the structure of spoken language. In: Proceedings of the 3rd international conference on speech prosody 2006, Dresden, Germany, 4 pages
10. Campbell N (2007) Towards conversational speech synthesis; lessons learned from the expressive speech processing project. In: Proceedings of the sixth ISCA workshop on speech synthesis (SSW6), Bonn, Germany, pp 22–27
11. Cowan N (1988) Evolving conceptions of memory storage selective attention, and their mutual constraints within the information processing system. Psychol Bull 104(2):163–191
12. Cowan N (2001) The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behav Brain Sci 24:87–185
13. Cowan N, Fristoe NM, Elliot EM, Brunner RP, Saults JS (2006) Scope of attention, control of attention, and intelligence in children and adults. Mem Cogn 34:1754–1768
14. Engle RW, Kane MJ (2004) Executive attention, working memory capacity, and a two-factor theory of cognitive control. In: Ross B (ed) The psychology of learning and motivation, vol 44. Elsevier, New York, pp 145–199
15. Engle RW, Cantor J, Carullo JJ (1992) Individual differences in working memory and comprehension: A test of four hypotheses. J Exp Psychol Learn Mem Cogn 18:972–992
16. Ericsson KA, Kintsch W (1995) Long-term working memory. Psychol Rev 102:211–245
17. Gnjatović M (2009) Adaptive dialogue management in human-machine interaction. Verlag Dr Hut, Munich
18. Gnjatović M, Rösner D (2007) An approach to processing of user's commands in human-machine interaction. In: Proceedings of the 3rd language and technology conference (LTC'07), Adam Mickiewicz University, Poznan, Poland, pp 152–156
19. Gnjatović M, Rösner D (2008) Adaptive dialogue management in the NIMITEK prototype system. In: Proceedings of the 4th IEEE tutorial and research workshop perception and interactive technologies for speech-based systems (PIT'08), Lecture notes in computer science, vol 5078. Springer, Berlin, pp 14–25
20. Gnjatović M, Rösner D (2010) Inducing genuine emotions in simulated speech-based human-machine interaction: The nimitek corpus. IEEE Trans Affect Comput 1:132–144
21. Gnjatović M, Pekar D, Delić V (2011) Naturalness, adaptation and cooperativeness in spoken dialogue systems. In: Toward autonomous, adaptive, and context-aware multimodal interfaces, theoretical and practical issues, Lecture notes in computer science, vol 6456. Springer, Berlin, pp 298–304
22. Griffin IC, Nobre AC (2003) Orienting attention to locations in internal representations. J Cogn Neurosci 15(8):1176–1194
23. Grosz B, Sidner C (1986) Attention, intentions, and the structure of discourse. Comput Linguist 12(3):175–204
24. Halliday M (1994) An introduction to functional grammar, 2nd edn. Edward Arnold, London
25. Hovy E, McCoy K (1989) Focusing your RST: A step toward generating coherent, multisentential text. In: Proceedings of the 11th annual conference of the cognitive science society, Ann Arbor
26. Jokinen K (2009) Constructive dialogue modelling: Speech interaction and rational agents. Wiley, New York
27. Jokinen K, Tanaka H, Yokoo A (1998) Context management with topics for spoken dialogue systems. In: Proceedings of COLING-ACL 1998, pp 631–637
28. Just MA, Carpenter PA (1992) A capacity theory of comprehension: Individual differences in working memory. Psychol Rev 99:122–149
29. Kane MJ, Conway ARA, Hambrick DZ, Engle RW (2007) Variation in working memory capacity as variation in executive attention and control. In: Conway ARA, Jarrold C, Kane MJ, Miyake A, Towse JN (eds) Variation in working memory. Oxford University Press, Oxford, pp 21–48
30. Kari L, Rozenberg G (2008) The many facets of natural computing. Commun ACM 51(10):72–83

31. Kirschner M (2007) Applying a focus tree model of dialogue context to interactive question answering. In: Proceedings of the ESSLLI'07 student session, Dublin, Ireland, pp 135–147
32. Lecœuche R, Robertson D, Barry C, Mellish C (2000) Evaluating focus theories for dialogue management. Int J Hum-Comput Stud 52(1):23–76
33. Li S-T, Tsai F-C (2010) Constructing tree-based knowledge structures from text corpus. Appl Intell 33(1):67–78
34. McCoy K, Cheng J (1991) Focus of attention: Constraining what can be said next. In: Paris CL, Swartout WR, Moore WC (eds) Natural language generation in artificial intelligence and computational linguistics. Kluwer Academic, Norwell, pp 103–124
35. Mendonça M, de Arruda LVR, Neves F (2011) Autonomous navigation system using event driven-fuzzy cognitive maps. Appl Intell. doi:10.1007/s10489-011-0320-1
36. Miller GA (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychol Rev 63:81–97
37. Moeller J-U (1996) Domain related focus-shifting constraints in dialogues with knowledge based systems. In: Adorni G, Zock M (eds) Trends in natural language generation: An artificial intelligence perspective. Springer, Berlin, pp 188–204
38. Mogle JA, Lovett BJ, Stawski RS, Sliwinski MJ (2008) What's so special about working memory? An examination of the relationship among working memory, secondary memory, and fluid intelligence. Psychol Sci 19:1071–1077
39. Mohammad Y, Nishida T (2010) Controlling gaze with an embodied interactive control architecture. Appl Intell 32(2):148–163
40. Moubaiddin A, Obeid N (2009) Partial information basis for agent-based collaborative dialogue. Appl Intell 30(2):142–167
41. Nobre CA, Coull TJ, Maquet P, Frith DC, Vandenberghe R, Mesulam MM (2004) Orienting attention to locations in perceptual versus mental representations. J Cogn Neurosci 16(3):363–373
42. Oberauer K (2002) Access to information in working memory: Exploring the focus of attention. J Exp Psychol Learn Mem Cogn 28(3):411–421
43. Oberauer K, Lange EB (2009) Activation and binding in verbal working memory: A dual-process model for the recognition of nonwords. Cogn Psychol 58(1):102–136
44. O'Reilly RC, Braver TS, Cohen JD (1999) A biologically-based computational model of working memory. In: Miyake A, Shah P (eds) Models of working memory. Cambridge University Press, New York, pp 375–411
45. Rahwan I, McBurney P (2007) Argumentation technology. IEEE Intell Syst 22(6):21–23
46. Roulet E (1992) On the structure of conversation as negotiation. In: Mey JL, Parret H, Verschueren J (eds) (On) Searle on conversation. John Benjamins, Philadelphia, pp 91–99
47. Salthouse TA (1996) The processing-speed theory of adult age differences in cognition. Psychol Rev 103:403–428
48. Searle J (1992) Conversation. In: Mey JL, Parret H, Verschueren J (eds) (On) Searle on conversation. John Benjamins, Philadelphia, pp 7–29
49. Shah P, Miyake A (1999) Models of working memory: An introduction. In: Miyake A, Shah P (eds) Models of working memory: Mechanisms of active maintenance and executive control. Cambridge University Press, New York, pp 1–26
50. Stede M, Schlangen D (2004) Information-seeking chat: Dialogue management by topic structure. In: Proceedings of the 8th workshop on semantics and pragmatics of dialogue, CATALOG 04, Barcelona
51. Stoltzfus ER, Hasher L, Zacks RT (1996) Working memory and aging: Current status of the inhibitory view. In: Richardson JTE, Engle RW, Hasher L, Logie RH, Stoltzfus ER, Zacks RT (eds) Working memory and human cognition. Oxford University Press, New York, pp 66–88
52. Unsworth N, Engle RW (2007) The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. Psychol Rev 114:104–132
53. Unsworth N, Spillers GJ (2010) Working memory capacity: Attention control secondary memory, or both? A direct test of the dual-component model. J Mem Lang 62(4):392–406
54. Wheeler ME, Treisman AM (2002) Binding in short-term visual memory. J Exp Psychol Gen 131:48–64
55. Xu Y, Ohmoto Y, Okada S, Ueda K, Komatsu T, Okadome T, Kamei K, Sumi Y, Nishida T (2010) Formation conditions of mutual adaptation in human-agent collaborative interaction. Appl Intell. doi:10.1007/s10489-010-0255-y
56. Zhong N, Liu J, Yao Y (2010) Introduction to brain informatics. Cogn Syst Res 11(1):1–2

Milan Gnjatović received the Ph.D. degree in computer science from Otto-von-Guericke University Magdeburg, Germany, in 2009. Currently, he works as a postdoctoral researcher in the Department of Power, Electronics, and Communications Engineering at the University of Novi Sad, Serbia. He has also worked as a research assistant in the Department of Knowledge Processing and Language Engineering at Otto-von-Guericke University Magdeburg. His research interests include adaptive dialogue management in human-machine interaction, cognitive technical systems, affective computing, and natural language processing.



Marko Janev received his B.Sc. degree in electrical engineering from the Faculty of Technical Sciences (FTN), Novi Sad, Serbia, in 2003. He received his M.Sc. and Ph.D. degree in Applied Mathematics from the Faculty of Technical Sciences (FTN), Novi Sad, Serbia, in 2009 and 2011, respectively. Currently he is a researcher at the Mathematical institute of Serbian academy of sciences and arts. His research interests include: statistical pattern recognition, speech recognition, speech processing, image processing and fractional calculus.

Vlado Delić received his M.Sc. degree in electrical engineering from the School of Electrical Engineering in Belgrade, Serbia in 1993. He received his Ph.D. degree in electrical engineering from the Faculty of Technical Sciences (FTN) Novi Sad in 1997. Currently, he is associate professor at the Faculty of Technical Sciences (FTN) Novi Sad and the leader of the "Human Computer Interaction" research team. His research interests include statistical pattern recognition, speech recognition, speech processing and speech synthesis and acoustics.