The 6th PSU-UNS International Conference on Engineering and Technology (ICET-2013), Novi Sad, Serbia, May 15-17, 2013
University of Novi Sad, Faculty of Technical Sciences

LINGUISTIC ENCODING OF MOTION EVENTS IN ROBOTIC SYSTEM

Milan Gnjatović*, Jovica Tasevski, Dragiša Mišković, Milutin Nikolić, Branislav Borovac, Vlado Delić
University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia
*Authors to correspondence should be addressed via email: milangnjatovic@yahoo.com

Abstract: *This paper reports and discusses an implementation of a cognitively-inspired and computationally appropriate linguistic encoding of motion events in human-robot dialogue. The proposed encoding is based on a schematic system of attention in spatial language and the conceptualization of the two fundamental cognitive functions in language – i.e., the Figure and the Ground – for the bipartite and tripartite spatial scene partitioning.*

Key Words: ***Linguistic Encoding, Motion Events, Human-Robot Interaction, Spatial Scene Partitioning, Adaptive Dialogue Management***

1. INTRODUCTION

Spatial context is inherently present in spoken interaction between the user and a robotic system with an embodied conversational agent. Therefore, an important research question relates to linguistic encoding of motion events. This paper integrates and expands upon previous work on adaptive multimodal interaction with industrial robot [1,2] and linguistic encoding of motion events [3,4,5]. It reports and discusses an implementation of a cognitively-inspired and computationally appropriate linguistic encoding of motion events in human-robot dialogue.

2. RELATED WORK

Research in the field of speech-based human-robot interaction is primarily focused on scenarios in which the user utters verbal commands in order to instruct the system to perform a preset operation [6-11]. Less attention is devoted to the research question of managing interactive spoken natural language dialogue in which both the user and the system may take initiative [12]. In our previous work [1], we addressed the latter research question, and introduced a spoken natural language dialogue system that manages the interaction between the user and the anthropomorphic robotic arm. The observed spatial context consists of a set of wooden objects, placed on a table, that differ in base, height, size and color. The user is allowed to spontaneously utter instructions of different syntactic forms that relate to manipulation of the objects (e.g., "move the red thin square downwards", etc.). The task of the system is to interpret the user's commands and to perform the requested operations. When needed, the system applies an adaptive dialogue strategy and initiates a conversation in order to help the user to specify the required information. At the implementation level, three subsystems are integrated with the robotic arm. The visual subsystem recognizes the objects and determines their positions. The audio subsystem implements the functionalities of automatic speech recognition and text-to-speech synthesis for the Serbian language. The cognitive subsystem implements the functionalities of natural language processing and adaptive dialogue management. In this paper, we show that the linguistic encoding of motion events applied in the cognitive subsystem is restricted to a particular spatial scene partitioning, and report an improvement in this regard.

3. SPATIAL SCENE PARTITIONING

In general, a way to address the question of linguistic encoding of motion events is to provide an account on relations between semantic elements within the domain of meaning (e.g., the object that is moved, its path, etc.) and expressions in overt linguistic forms (e.g., the user's command) [13, p. 21]. In his well-known approach, Talmy introduces two fundamental cognitive functions in language – the Figure and the Ground – and conceptualizes them, in spatial context, as follows. *The Figure is a moving or conceptually movable entity whose path, site, or orientation is conceived as a variable, the particular value of which is the relevant issue. The Ground is a reference entity, one that has a stationary setting relative to a reference frame, with respect to which the Figure's path, site, or orientation is characterized* [14, p. 312]. In a single clause, these concepts are typically represented by nominals, as illustrated in "*the red square* [FIGURE] lay on *the table* [GROUND]".

Related to the linguistic encoding of motions events implemented in [1], the cognitive subsystem applies only a bipartite spatial scene partitioning. In other words, at the level of the user's command, a wooden object always functions as the Figure, while the table always functions as the Ground. A typical linguistic pattern of expression of a semantic motion event can be observed in command "move the red thin square downward". The noun phrase "the red thin square" relates to the object that functions as the Figure, while the Ground (i.e., the table) is not explicitely refferenced in the command. This is the expected user behavior. The table represents the working frame of the robot, so there is no need to explicitly utter a reference to it.

However, the spatial scene might have been partitioned in another way – a tripartite partitioning. It includes a Figure object, a Ground object, and a reference frame as a background [14, p. 313]. A command exemplifying this partitioning is "move *the red square* [FIGURE] under *the blue square* [GROUND]". In this example, the nominal phrase "the red square" functions as the Figure, while the nominal phrase "the blue square" functions as the Ground. More generally, in the given spatial context, any wooden object may figure both as the Figure and as the Ground (cf. "move the *blue* square under the *red* square").

Gnjatović and Delić [2] argue that the user may interchangeably adopt a bipartite and a tripartite partitioning of the spatial scene during the interaction, and that the system must be able to interpret linguistic representations containing both these encoding patterns. They introduce a computational model of a schematic system of attention in spatial language that encapsulates the dichotomy between the cognitive functions of the Figure and the Ground in both the patterns.

This work expands upon the proposed schematic system. For the purpose of the discussion, we briefly sumarize the Talmy's sketch of a basic motion event. It is a composite of four semantic elements. Besides the Figure and the Ground, two additional components of the basic motion event are the Path and the Motion. *The Path is the path followed or site occupied by the Figure object with respect to the Ground object, while the Motion refers to the presence per se of motion or locatedness in the event* [13, p. 25]. In a single clause, the Path may be represented by a preposition or an adverb, and the Motion by a verb, as illustrated in "the red square [FIGURE] *lay* [MOTION] *on* [PATH] the table [GROUND]".

4. LINGUISTIC REALIZATION OF MOTION EVENTS

To illustrate some of the important phenomena related to the linguistic realization of motion events in the course of the dialogue, let us consider the sequence of the user's commands given in Table 1. We focus on the interchangable spatial scene partitioning and frequent surface realizations of cohesive devices that create coherence in the conversation. The system should be able to cope with the following dialogue phenomena:

(1) Real-time interchangable partitioning of the spatial scene. The given sequence illustrates how the user interchangeably adopts a bipartite and a tripartite partitioning of the spatial scene during the interaction.

(2) Surface realizaton of the semantic elements. Surface realization of the observed semantic elements within the domain of meaning (i.e, Figure, Ground, Motion, Path) is realized on a part of speech level. However, the relation between the semantic elements and sufrace expressions is not necessarily one-to-one. Thus, in the given examples, the Figure and the Ground are represented by a nominal phrase and a pronoun. The Path is realized by an adverb of place in the commands illustrating the bipartite spatial scene partitioning, or by a preposition in the commands illustrating the tripartite partition. The Motion is represented by a verb[1].

(3) Anaphoric references. In the second command, the pronoun "it" functions as the Figure. It is a reference to the square specified in the first command. In contrast to this, the same pronoun in the fourth command functions as the Ground, and refers to the circle specified in the third command.

(4) Elliptical commands. In the third command, there is no explicit linguistic realization of the Figure and the Motion. The user believes that the system is aware that the square specified in the first command functions as the Figure in this command, and that the action of moving specified in the second command functions as the Motion. Therefore, this command is elliptical – the user does not include utterance constituents that refer to the Figure and the Motion.

Table 1. *The sequence of the user's commands.*

Linguistic realization	***Partitioning***
Move the square downward.	Bipartite
Move it leftward.	Bipartite
And now under the circle.	Tripartite
Move also the triangle under it.	Tripartite

In order to address these dialogue phenomena, the system must perform some sort of contextual analysis and interpretation. We address this research question in the next section.

5. MODELING CONTEXTUAL INFORMATION

For the purpose of easier representation, we restrict the spatial context shared between the user and the system to three objects – a square, a triangle, and a circle – each of which may function both as the Figure and as the Ground in the user's commands. The possible values of the Motion are determined by the two operations that can be performed by the system. It can point to an object or translate an object in the horizontal plane of the table. Semantic entities that refer to the presence of translation of an object that functins as the Figure are: *upward*, *downward*, *leftward*, and *rightward*. These directions represent the possible values of the Path in the bipartite partitioning of the spatial scene. In constrast to this, possible sites occupied by the Figure object with respect

[1] Patterns underlying the surface realization of semantic elements vary across different languages. Here, we focus primarily on lexicalization patterns in the Serbian language.

to the Ground object are: *above*, *under*, *to the left of*, and *to the right of*. These sites represent the possible values of the Path in the tripartite partitiononing of the spatial scene.

Following the linguistic encoding proposed in [3], the contextual information is represented in two hierarchical structures, collectively reffered to as *focus trees*. The first focus tree is used for processing the user's commands that instantiate the bipartite partitioning of the spatial scene. Therefore, we refer to this structure as to *bipartite focus tree.* A simplified version of this focus tree is given in Fig. 1, and the interpretation of the semantic entities contained in it are provided in Table 2. Fig. 1 also illustrates the mapping of two user's commands that instantiate the bipartite partitioning:

- "Show the square" (cf. dashed arrows marked with number 1),
- "Move the triangle leftward" (cf. dashed arrows marked with number 2).

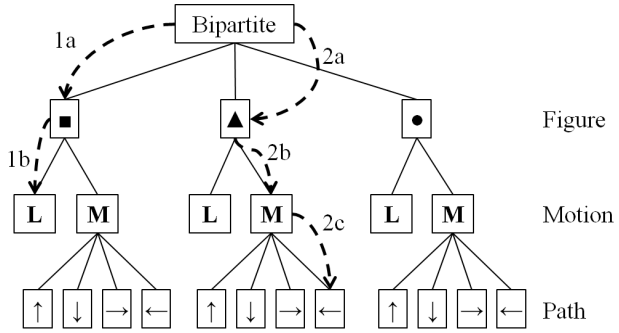

Fig. 1. *A simplified bipartite focus tree, used for processing the user's commands that instantiate the bipartite partitioning of the spatial scene.*

Table 2. *Interpretation of the semantic entities contained in the bipartite focus tree represented in Fig. 1.*

Node	***Semantic entity***	***Interpretation***
■	Figure	The square
▲	Figure	The triangle
●	Figure	The square
M	Motion	Motion of the Figure
L	Motion	Locatedness of the Figure
←	Path	Leftward translation the Figure
↑	Path	Upward translation of the Figure
→	Path	Rightward translation of the Figure
↓	Path	Downward translation of the Figure

The second focus tree is used for processing the user's commands that instantiate the tripartite partitioning of the spatial scene[2]. We refer to this structure as to *tripartite focus tree.* A simplified version of this focus tree is given in Fig. 2, and the interpretation of the semantic entities contained in it is provided in Table 3. Fig. 2 also illustrates the mapping of two elliptical user's commands that introduce the Groud and the Path components:

- "Above the square" (cf. dashed arrows marked with number 3),
- "Under the circle" (cf. dashed arrows marked with number 4).

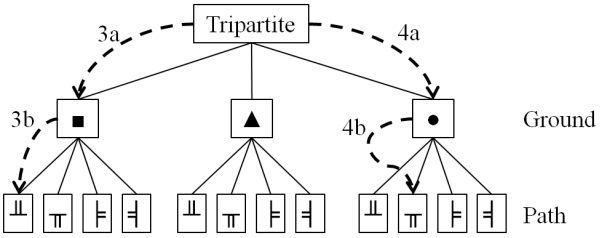

Fig. 2. *A simplified tripartite focus tree, used for processing the user's commands that instantiate the tripartite partitioning of the spatial scene.*

Table 3. *Interpretation of semantic entities in the focus tree represented in Fig. 2.*

Node	***Semantic entity***	***Interpretation***
■	Ground	The square
▲	Ground	The triangle
●	Ground	The square
╡	Path	On the left side of the Ground
╨	Path	Above the Ground
╞	Path	On the right side of the Ground
╥	Path	Under the Ground

The phrasal lexicon is organized as a collection of sets of keywords and phrases. To each node in the focus trees, a set of phrases that represent the observed node is assigned. These sets are not necessarilly disjoint or mutually different, as illustrated in Table 4.

Table 4. *Organization of the phrasal lexicon.*

Entity	***Assigned phrases***
■	"black square", "black", "square", "figure", …
▲	"black triange", "black", "triangle", "figure", …
●	"black circle", "black", "circle", "figure", …
M	"move", "translate", "shift", …
S	"show", "point to", …
←	"leftward", "to the left", …
↑	"upward", "up", …
→	"rightward", "to the right", …
↓	"downward", "down", …
╡	"on the left side", "left of", …
╨	"above", …
╞	"on the right side", "right of", …
╥	"under", "below"

6. PROCESSING OF THE USER'S COMMANDS

Processing of the user's commands is illustrated for the sequence of commands given in Table 1. The first command ("Move the square downward") instantiates the bipartite partitioning. Therefore, it is mapped onto the bipartite focus tree. The transitions of the focus of

[2] More precisely, while the system used only the bipartite focus tree to process commands that instantiate the bipartite partitioning, it uses both these focus trees to process commands that instantiate the tripartite partitioning. This is illustrated below.

attention are indicated with the dashed arrows marked with number 5 in Fig. 3. At the beginning of this dialogue fragment, the focus of attention is placed on the root of the bipartite focus tree. After the first command is processed, the current focus of attention is placed on the terminal node indicated with the arrow 5c. The current focus of attention and its ancestors specify the Figure, the Motion and the Path components of the motion event encoded in the command. It means that the system has all necessary information to perform the command.

This second command ("Move it leftward") contains two explicitly uttered keywords – *move* and *leftward* – that respectively specify the Motion and the Path. However, to map this command, it is necessary to resolve the anaphorical reference *it* that specifies the Figure. The system resolves this anaphora indirectly. Since the current focus of attention is in the subtree determined by node ■ as its root, the system infers that the square functions as the Figure in this command. The transitions of the focus of attention are indicated with the dashed arrows marked with number 6 in Fig. 3.

The third command ("And now under the circle") instantiates the tripartite partitioning of the spatial scene. At the surface level, the phrase "the circle" comes immediatelly after the utterance constituent that unambiguously relates to the Path (i.e., "under"). Therefore, the system infers that the phrase "the circle" specifies the Ground. These two semantic entities are mapped onto the tripartite focus tree, as indicated with the dashed arrows marked with number 7 in Fig. 3. On the other hand, the propositional content of this command does not include information on the Figure and the Motion. This information is extracted from the current focus of attention in the bipartite focus tree that remained unchanged. Its ancestors specify the missing components of the encoded motion event – implying that the square functions as the Figure, and the translation operation as the Motion. It should be noted that, for this command, the information on the Path is not extracted from the bipartite focus tree, because it is explicitly specified in the command. Now, the system has all necessary information to perform the command.

At the suface level of the last command ("Move also the triangle under it"), the utterance constituent that functions as the Path immediately precedes the anaphoric reference. Therefore, the system draws several conclusions:

- the command instantiates the tripartite partitioning of the spatial scene,
- the Path should be mapped onto the tripartite focus tree,
- the anaphora relates to the Ground,
- the constituent "the triangle" functions as the Figure.

The Figure and the Motion are explicitly specified in the command, and are mapped onto the bipartite focus tree (cf. dashed arrows marked with number 8 in Fig. 3). The only missing information at this moment relates to the Ground. The system resolves this anaphora in a similar manner as for the second command. Since the current focus of attention in the tripartite focus tree is placed on a node that belongs to the subtree determined by node ●, the system infers that the circle functions as the Ground. At this moment, the system has all necessary information to perform the command.

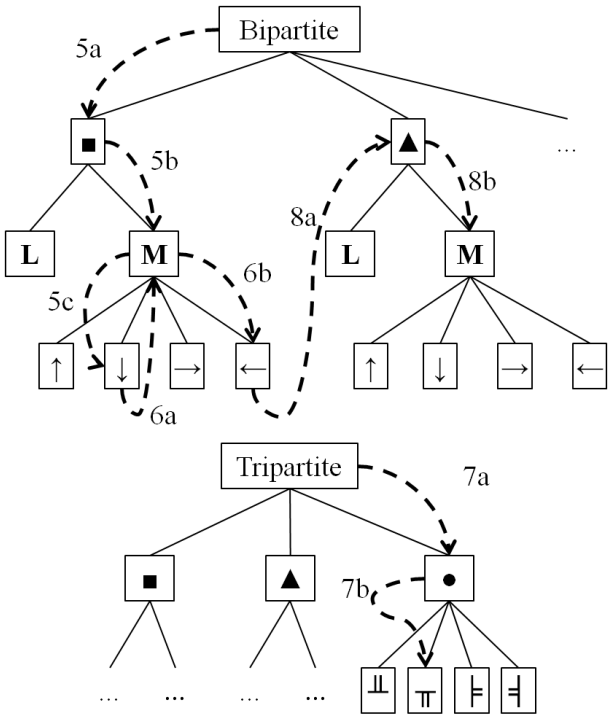

Fig. 3. *Transitions of the focus of attention. Only those parts of the focus trees that are relevant for the discussion are represented.*

7. CONCLUSION

This paper reported and discussed an implementation of a cognitively-inspired and computationally appropriate linguistic encoding of motion events in human-robot dialogue. The proposed encoding is based on a schematic system of attention in spatial language and the conceptualization of the two fundamental cognitive functions in language – i.e., the Figure and the Ground – for the bipartite and tripartite spatial scene partitioning. We discussed that the proposed encoding is appropriate to address dialogue phenomena such as real-time interchangable partitioning of the spatial scene, anaphorical references, elliptical commands, etc.

Fig. 4. *The prototype system.*

This paper also reports an improvement of the prototype system introduced in [1,2] that integrates a visual system and a dialogue system with the industrial robot ABB IRB140, an anthropomorphic robotic arm with six degrees of freedom (cf. Fig. 4). The implementation of the reported linguistic encoding within the integrated dialogue system improved primarily the natural language understanding functionality, as discussed in this paper.

Acknowledgment. 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.

5. REFERENCES

[1] M. Gnjatović, J. Tasevski, M. Nikolić, D. Mišković, B. Borovac and V. Delić, "Adaptive Multimodal Interaction with Industrial Robot", *Proc. of the IEEE 10th Jubilee International Symposium on Intelligent Systems and Informatics (SISY 2012)*, Subotica, Serbia, 2012. pp. 329–333.

[2] J. Tasevski, M. Nikolić, D. Mišković, "Integration of an Industrial Robot with the Systems for Image and Voice Recognition", *Serbian Journal of Electrical Engineering* 10(1), 2013, pp. 219-230.

[3] M. Gnjatović and V. Delić, "Attention and Linguistic Encoding of Motion Events in Human-Machine Interaction", in *Selected papers from the 3rd International Conference of Syntax, Phonology and Language Analysis*, S. Halupka-Rešetar, M. Marković, T. Milićev, and N. Milićević, (Eds), Cambridge Scholar Publishing, 2012, pp. 237-257.

[4] M. Gnjatović, M. Janev, and V. Delić, "Focus tree: Modeling attentional information in task-oriented human-machine interaction", *Applied Intelligence* 37(3), 2012, pp. 305-320.

[5] M. Gnjatović and V. Delić, "A Cognitively-Inspired Method for Meaning Representation in Dialogue Systems", *Proc. of the 3rd IEEE International Conference on Cognitive Infocommunications*, Kosice, Slovakia, 2012, pp. 383–388.

[6] J.N. Pires, "Robot-by-voice: Experiments on commanding an industrial robot using the human voice", *Industrial Robot: An International Journal*, Vol. 32, 2005.

[7] J.N. Pires, G. Veiga, R. Araujo, "Programing-by-demonstration in the coworker scenario for SMEs", *Industrial Robot: An International Journal*, Vol. 36, 2009.

[8] D. Rambabu, R. Nagaraju, B. Venkatesh, "Speech Recognition of Industrial Robot", *International Journal of Computational Mathematical Ideas*, Vol. 3, 2011.

[9] R. Hollmann, M. Hägele, "The use of voice control for industrial robots in noisy manufacturing environments", *International Symposium on Robotics*, Seoul, 2008, pp. 14-18.

[10] C. Vara-Thorbeck, V.F. Muñoz, R. Toscano, J. Gómez, J. Fernández, M. Felices, A. García-Cerezo, "A new robotic endoscope manipulator. A preliminary trial to evaluate the performance of a voice-operated industrial robot and a human assistant in several simulated and real endoscopic operations", *Surgical Endoscopy Ultrasound and Interventional Techniques*, 15(9), 2001, pp.924-927.

[11] J. Fernández-Lozano, J.M. Gómez-de-Gabriel, V.F. Muñoz, I. García-Morales, D. Melgar, C. Vara and A. García-Cerezo, "Human-Machine Interface Evaluation in a Computer Assisted Surgical System", *Proceedings of the 2004 IEEE International Conference on Robotics & Automation*, New Orleans, 2004, pp. 231-236.

[12] M. Holada, M. Pelc, "The robot voice-control system with interactive learning", *New Developments in Robotics Automation and Control*, 2008, pp. 219-228.

[13] L. Talmy, *Towards Cognitive Semantics, Volume 2, Typology and Process in Concept Structuring.* Cambridge, Mass. MIT Press, 2000.

[14] L. Talmy, *Towards Cognitive Semantics, Volume 1, Concept Structuring Systems.* Cambridge, Mass. MIT Press. 2000.