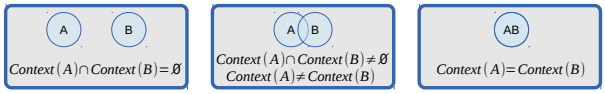
Towards an intuitive interface for instructing robots handling tasks based on verbalized physical effects

Michael Spangenberg and Dominik Henrich
Angewandte Informatik III
Robotik und Eingebettete Systeme
Universität Bayreuth, D-95440 Bayreuth, Germany
E-Mail: {michael.spangenberg | dominik.henrich}@uni-bayreuth.de

***Abstract*— A long-term goal in current robotic research is the development of intuitive interfaces for human-robot interaction. Here, one field of application are object manipulation tasks. Such tasks consist of grasping, moving, and placing objects [1]. In this work, we focus on the subtask of moving an object, which is also called the *handling* of the object. We present a method for the intuitive instruction of handling tasks through verbal commands and the execution based on verbalized physical effects. We define a set of principal physical effects and describe how a physical effect can be verbalized. Furthermore, we indicate how verbal parameters can qualitatively be transformed into robot control parameters using physics. At last, we show in a user study, that the proposed method is feasible for the intuitive instruction of handling tasks to a robot system.**

I. Introduction

One long-term goal in current robotic research is the development of intuitive interfaces for programming industrial or service robots. Whether an interface is intuitive or not, depends on many factors, which we subsume as the *context*. Thus, our basic hypothesis is that a common or shared context is fundamental for a successful communication between two agents. The three basic cases of shared context are illustrated in Figure 1. In Figure 1a, both agents have no common context, i.e. no communication is possible. Due to the existence of a common context, agents in Figure 1b and Figure 1c can communicate with each other. The difference between both situations is the way how intuitive the communication is. In Figure 1b, both agents have information which is not in the common context. This have to be named explicit and in a well-defined way, for example in a specific syntax with known parameters. In Figure 1c, both agents have the same context, so there is no need to share information between them. This describes the most intuitive communication, but is too idealistic for current human-robot applications.



(a) (b) (c)

Fig. 1. The amount of common context as criterion for the intuition of the communication between two agents A and B. The more context both agents have in common, the more intuitive they may communicate.

At this point, we have to define, which information shall be in the common context of two agents in order to obtain an intuitive communication interface for the instruction of *handling* tasks [2]. The handling of an object is the subtask of moving the object in a specific way within an object manipulation task. Such object manipulation tasks consists typically of the three steps of grasping, moving, and placing of objects [1]. We focus on the subtask of handling, because the intuitive instructions of how the motion of an object shall be executed by a robot system are more diverse than the instructions for grasping or placing of an object, which can be achieved by commands like *Grasp the object from the table!* or *Place the object on the table!*
To describe handling tasks in an intuitive way, we consider the combination of two aspects:

- natural language commands
- physical effects

We focus on these aspects, because language is a natural, intuitive interface for the communication between humans. However, robots need more formal interfaces like compliance motions [3], task frames [4], skills [5], [6] or task specifications [7] for communication. Before we can use natural language as user input, the robot needs information about the semantic of the instructed commands, especially about the used verbs, since they identify the action which shall be executed. Here we suggest to use the laws of physics, respectively physical effects.
The remainder of this paper is organized as follows: The related work is described in Section II. In Section III, we introduce our proposed concept for an intuitive interface based on verbalized physical effects. The evaluation of the concept is presented in Section IV. At last, we discuss the results and describe our future work in Section V.

II. Related work

There are many systems, which use natural language as an input modality for commanding a robot. These systems are typically built according to the 3T architecture [8], i.e. there are three layers within the system. The top layer defines the high-level interface for the user, which is typically a high-level representation, like a natural language interface. This high-level representation is then transformed into a low-level representation by the mid layer and executed at the bottom layer, the low-level robot control.

We are interested in a system, which is capable of understanding the differences between handling tasks. For example, if we command the robot to *Bump the object to the wall!*, we expect a different execution than for the command *Shove the object to the wall!* or *Touch the object!*. This abilities shall be provided to the robot by elementary abilities, or *skills* [5], which describe the transformation between well-defined states.

As mentioned before, there exists a wide range of systems, which use natural language for commanding a robot. These systems are typically used for the instruction of manipulation or navigation tasks. Since systems which use the natural language for commanding navigation tasks [9], [10], [11], do not manipulate the environment, we only focus on systems, which use natural language for commanding a robot in manipulation tasks. According to the 3T architecture, we focus on implementations of the mid (and partly the bottom) layer of such systems, i.e. we do not focus the processing of natural language.

In KANTRA, Laengle et al. [12] use natural language commands with a well-defined syntax and known parameters (object quantifiers, positions), and execute them via predefined plans of the plan execution system FATE. The spectrum of predefined plans range from high-level commands to explicit robot operations. Knoll et al. [13] assemble predefined wooden objects by skills like *pick-up*, *peg-in-hole* or *screwing*. Pires [14] commands a robot by natural language commands, which have a known syntax and known symbolic parameter values and maps them directly to robot commands. Tenorth et al. [15] take instructions from the world wide web and execute them based on learned action models. Stenmark and Nugues [16] transform natural language commands into a predicate-argument structure and execute them with the knowledge integration framework KIF [17].

We can conclude, that the former approaches transform natural language commands into two kinds of representations. On the one hand, it is the direct transformation to low-level robot commands. This representation is independent of the executed task but needs a lot of instructions and knowledge of the underlying robot control unit. On the other hand, it is the transformation into high-level skills, which is an intuitive way of bringing functionality to a robotic system. But non of the considered systems focus on bringing semantic of action verbs to a robot system

The main contribution of this paper is the definition of a set of skills for commanding a subset of object manipulation tasks, the handling of the object, to a robot system. This skill set is based on verbalized physical effects, which are used to offer semantic information about specific verbs to a robot system. This semantic information describes which symbolic parameters (words) are needed by a specific handling operation, and how this symbolic parameters have to be transformed to robot control parameters (positions/forces/torques) using sensor data and physics.

III. Concept overview

In the following sections, we describe our proposed method for the execution of natural language handling commands based on verbalized physical effects. This includes the definition and verbalization of the involved principal physical effects. The transformation of the symbolic into sub-symbolic parameters and the execution of the physical effects by a robotic manipulator will be focused on in detail in future works (see Section V).

A. Physical effects

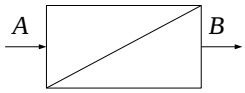
First, we describe the physical quantities, which are involved in an object manipulation task. Generally, there are seven base units defined in ISO 30-0 [18]. The manipulation of objects consists of mechanical operations. Therefore, the mechanical base units *length L*, *mass M* and *time T* are changed by handling objects.

Furthermore, derived units exist, which are combinations of the base units. This can be categorized in the following quantities [19]:

- *geometric*: length, angle, area, and volume
- *kinematic*: time, frequency, angular velocity, angular acceleration, velocity, and acceleration
- *dynamic*: mass, density, impulse, force, moment, work, power, and pressure

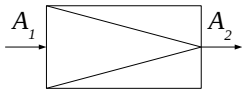
As a next step, we describe and categorize principal effects on physical quantities, which are used by Pahl et. al. [20] in the field of engineering design. We adopt these considerations and categorize the physical effects into two groups, the *elementary* and the *complex* physical effects. An elementary effect takes one input parameter and has zero or exactly one output parameter, while complex effects have multiple input and/or multiple output parameters. We define the five principal effects on physical quantities:

- *transform*



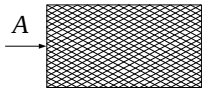
Transforms a physical quantity of type A into a physical quantity of type B

- *change*



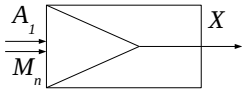
Changes the value and/or direction of a physical quantity A_1 into A_2

- *absorb*



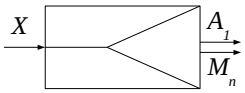
Absorbs a physical quantity A

- *merge*



Merges a set of physical quantities to a physical quantity X

- *split*



Splits a physical quantity X to a set of physical quantities

If we verbalize the physical effects at this level of abstraction, we would get terms like *transform a force into a length (displacement)*, *transform an impulse into a displacement*, *change the potential work* or *absorb a force*, which are not intuitive to verbalize for the user. Therefore, we describe in the next subsection, how we verbalize the physical effects in our approach, in order to get an intuitive interface for human-robot interaction.

B. Verbalize a physical effect

In this work, we focus on the verbalization of the elementary effects *transform*, *change* and *absorb*. Complex effects will be discussed in future works. After having defined the physical effects in the last subsection, we map an action name to each physical effect. In general, the action name should be a unique verb, which contains all the needed parameter information in symbolic form. Furthermore, it should be an intuitive expression for the user. The realization of these requirements will be discussed below.

In our approach, we use verbs as action names, because a natural description of actions can be achieved through dynamic verbs like *touch*, *lift*, and *move*. To choose suitable action names, which contain all needed information, we regard the *valency* [21] of the verbs. The valency describes the obligatory amount of complements, which are needed to define a semantically correct sentence. For example, complements can be noun phrases, which hold information about the actor of the action, or can be prepositional phrases, which hold information about local or temporal relationships. Generally, a verb can bind up to four complements. For the elementary physical effects, we need verbs with a valency of two and three, because an elementary physical effect has one actor (the robot, which is typically named implicit), one input quantity (in *transform*, *change* and *absorb* effects), and one output quantity (in *transform* and *change* effects). These complements have to be identified and mapped to the parameter of the desired physical effect. The identification of complements can be exemplarily accomplished by a statistical parser like [22] and is illustrated in Figure 2. Note that we do not focus the natural language processing, we just use the valency of the verbs to choose a valid verbalization of an physical effect.

To sum up, the usage of valency information helps us to select elements from the set of dynamic verbs and can be used as a pre-test for potential action names. But which exact verb shall be used for the description of a specific physical effect and which is intuitive for the user, can only be evaluated through the collection and analysis of empirical data, which is described in Section IV.

C. Parameter mapping

At this point, a verb is associated to a physical effect, for example the verb *to shove* is mapped to the effect *transform a force into a displacement* (see first row in Figure 3), the verb *to bump* to the effect *transform a momentum into a displacement* (see second row in Figure 3) or the verb *to touch* to the effect *absorb a force* (see fourth row in Figure 3). The input quantity of a physical effect appears at the object, which is represented in the noun phrase of the command (*the red box* in the example of Figure 2). The output quantity of the physical effect (if available) is described by the other complement of the verb, for example in a prepositional phrase (*to the wall* in the example of Figure 2).

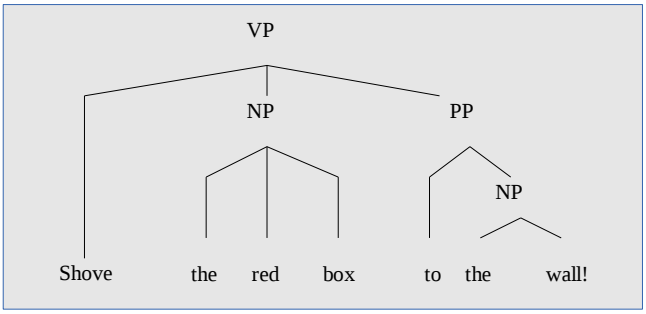


Fig. 2. Output of the used statistical parser [22] for the sample command *Shove the red box to the wall!* The *VP* denotes a *verbal phrase*, the *PP* denotes a *prepositional phrase*, and the *NP* denotes a *noun phrase*.

The next step is to map these words to robot control parameters. Here we use different sources of information, for example the laws of physics, sensor information, or an environment model. For the example in Figure 2, we use physics to calculate the value of the minimum force F_R, so that the object starts to move. The direction of the force and the displacement is calculated from the environment model. Because the object has a mass m_O and there is friction μ between the object and the ground, we can calculate the force by the equation (g denotes the acceleration of gravity)

$$F_R = \mu \cdot m_O \cdot g$$

Another example is the transformation of a momentum into a displacement. The execution of this physical effect can be instructed via the command *Bump the red box to the wall!* Here we need to calculate the momentum p_R, respectively the velocity of the robot v_R. This depends on the mass of the robot m_R, the mass of the object m_O, the friction between the object and the ground μ and the distance d. This can be calculated by using the conservation law of energies, as follows

$$v_R = \sqrt{\frac{2 \cdot d \cdot m_O \cdot \mu \cdot g}{m_R}}$$

To sum up, verbal parameters can be transformed into robot control parameters by using the laws of physics and context information like the environment model or sensor data. We show this exemplarily for the physical effect *transform a force into a displacement*, which we verbalize by the verb *to shove*, and the physical effect *transform a momentum into a displacement*, which we verbalize with the verb *to bump*. Future work will focus on this aspect in more detail.

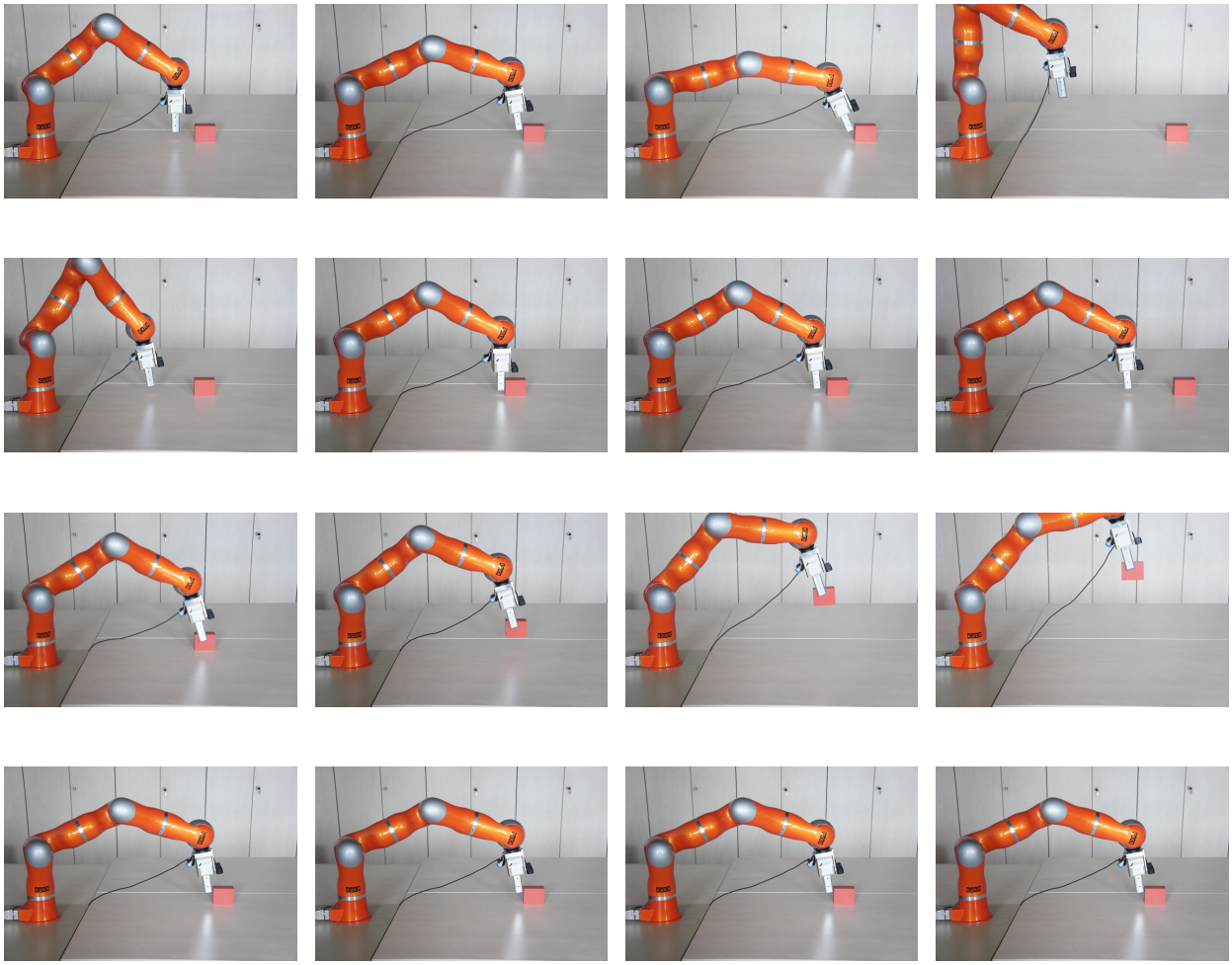


Fig. 3. Overview of the handling tasks used in the experiment. Each row describes a different handling task. For Task 1, we expect a *shove* command, which is the selected verbalization of the physical effect *transform a force into a displacement*. For Task 2, we expect a *bump* command, which is the selected verbalization of the physical effect *transform a momentum into a displacement*. For Task 3, we expect a *lift* command, which is the selected verbalization of the physical effect *change the potential energy of the object*. For Task 4, we expect a *touch* command, which is the selected verbalization of the physical effect *absorb a force*.

D. Execution

As a low-level interface to the robot control unit, we use manipulation primitives [23]. The formal definition of a manipulation primitive is

$$MP := \{HM, \tau, \lambda\}$$

where HM denotes a *hybrid motion* of the robot. It is called hybrid motion, because every degree of freedom can be controlled by a different control strategy, for example position control, force control, or distance control. These hybrid motions are based on the concept of compliance frames [3]. Therefore, the hybrid motion has two components

$$HM := \{TF, D\}$$

in which TF denotes the *task frame*, which is relative to a *base frame*. The D describes the reference variable and control strategy for each degree of freedom. The τ describes a set of tool commands, which can be executed during a manipulation primitive, for example commands like $\{Gripper, Open\}$ or $\{Camera, TakePhoto\}$. At last, the λ describes a set of termination criteria. This can be either the reference variables in D or any sensor signal.

IV. Evaluation

In this section, we investigate how well our approach is suitable for human-robot interaction. In order to analyze this, we set up a user study, in which test persons shall command a robot for the set of handling tasks in Figure 3. We demonstrated the test persons the execution of the handling tasks and asked them how they would instruct a robot using natural language commands. All commands were instructed in German, although in this paper their English translations are used [1].

[1]The German verb *schieben* was translated by *to shove*, *stoßen* by *to bump*, *heben* by *to lift*, *drehen* by *to rotate*, *rollen* by *to roll*, *berühren* by *to touch*, *schießen* by *to shoot*, *positionieren* by *to position* and *drücken* by *to push*.

TABLE I

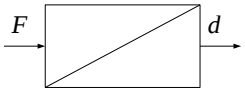
RESULTS OF THE USER EXPERIMENTS

	Group A		Group B		Group C		Group D		Total
	Verb	Ratio r_A	Verb	Ratio r_B	Verb	Ratio r_C	Verb	Ratio r_D	Ratio r_T
Task 1									
top-down	*Shove*	67%	*Shove*	100%	*Shove*	100%	*Shove*	67%	83%
	Push	33%					*Position*	33%	17%
bottom-up	*Shove*	100%	*Shove*	100%	*Shove*	100%	*Shove*	100%	100%
Task 2									
top-down	*Bump*	67%	*Bump*	100%	*Bump*	100%	*Bump*	67%	83%
	Push	33%					*Shoot*	33%	17%
bottom-up	*Bump*	100%	*Bump*	100%	*Bump*	100%	*Bump*	100%	100%
Task 3									
top-down	*Lift*	100%	*Lift*	100%	*Lift*	100%	*Lift*	100%	100%
bottom-up	*Lift*	100%	*Lift*	100%	*Lift*	100%	*Lift*	100%	100%
Task 4									
top-down	*Touch*	67%	*Touch*	67%	*Touch*	100%	*Touch*	67%	75%
	Push	33%	*Push*	33%			*Push*	33%	25%
bottom-up	*Touch*	100%	*Touch*	100%	*Touch*	100%	*Touch*	100%	100%

The user study was performed with twelve test subjects, which were divided into four groups with three persons in each. In Group A were persons with general education, representing a standard user without special knowledge in computer science. Persons with knowledge in natural science or mathematics were arranged in Group B. In Group C were persons with skills in computer science, especially in programming. The last Group D consisted of persons, which have expertise in robotics and in programming of robot systems.

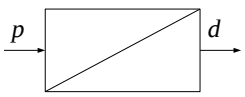
The user study was divided into two parts, the top-down and bottom-up experiment. First, each test person executed the top-down experiment. In this part, there were made no restrictions to the instructed commands, because we are interested in the verbs used without any influence by a given context or pre-defined commands. On the contrary, in the second part, we defined a set of instructions, which are available for the commanding of handling tasks. We used the following subset of verbalized physical effects as pre-defined commands (the input of the physical effect is mapped and applied to the *Object*, the output of the physical effect is mapped to the *Position*).

- SHOVE an *Object Preposition Position*



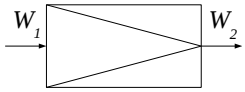
Transform the force F into a displacement d

- BUMP an *Object Preposition Position*



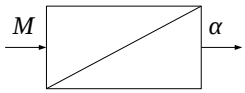
Transform the momentum p into a displacement d

- LIFT an *Object Preposition Position*



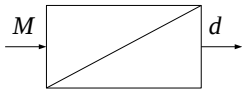
Change the potential energy W_1 into W_2, with $W_1 < W_2$

- ROTATE an *Object Preposition Position*



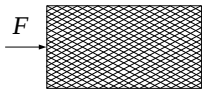
Transform the torque M into an angle α

- ROLL an *Object Preposition Position*



Transform the torque M into a displacement d

- TOUCH an *Object*



Absorb the force F

The results of the experiments are shown in Table I. For each task (1-4), user group (A-D), and approach (top-down/bottom-up), we illustrate the given action verbs. The Ratio r_i with $i \in \{A, B, C, D, T\}$ denotes the usage of a specific verb in each experiment for a user group, respectively for all user groups.

First, we discuss the results of the top-down experiment. In total, the test groups characterized each task in the top-down experiment by a dominant verb, whose total ratio r_T ranges from 75% up to 100 %. The main deviation can be found in the instructions for Task 4, where 25% of the test persons chose the verb *push*. This verb is also used once for

Task 1, therefore the meaning of this verb is apparently not unique. We can explain this by using two examples. If we push through a door, the door typically starts moving, so we transform a force or momentum into a motion. However, if we push an object with a big mass, like a car or a house, these objects absorb the force and there is no motion. Since this verb can stand for more than one physical effect, it does not meet the requirements for the verbalization of a physical effect (see Section III-B).

Next, we discuss the results of the bottom-up experiment. Here we can point out two observations. First, the test subjects of all test groups can successfully instruct the handling tasks using the predefined command set. This applies especially to the persons with no experience in programming (Group A and B). Second, all test persons use the expected verbalization for the demonstrated physical effects. This shows, that the users intuitively understand the differences between the selected verbs.

At last, we compare the results of the top-down and bottom-up experiment. Our goal is to achieve a high level of coincidence between the results of both experiments. We note that for each task, our proposed verbalization in the bottom-up experiment matches quite good the dominant expression of the top-down experiment. Furthermore, the users who used the ambiguous expression *push*, can successfully specify the handling task with our predefined verbalization. This early user experiment indicates, that the proposed method is suitable as an intuitive interface for human-robot interaction.

V. Conclusion and future work

We have described a new approach for bringing semantic information to a robot system. We have defined a skill set for the intuitive instruction of handling tasks, which is based on verbalized physical effects. This skill set consists of elementary and complex physical effects and has the goal to provide one skill for each physical effect. This skill set can then be the base for more complex tasks.

We have focused on the verbalization of elementary physical effects. To verbalize this effects by an intuitive expression, we have pointed out, that an intuitive verbalization can only be determined by collection and analysis of empirical data. For demonstration, we described the verbalization of six elementary physical effects and evaluate this verbalization through user experiments. The results of the user experiments show, that the proposed method is suitable as an intuitive interface for human-robot interaction.

Future work may include the verbalization of complex physical effects, the transformation of the high-level parameters into low-level robot control parameters, and the execution of the robot motions based on manipulation primitives.

References

[1] J. R. Flanagan, M. C. Bowman, and R. S. Johansson, "Control strategies in object manipulation tasks," *Current opinion in neurobiology*, vol. 16, no. 6, pp. 650–659, 2006.

[2] The Association of German Engineers (VDI), "VDI 2860 Assembly and handling; handling functions, handling units; terminology, definitions and symbols," *Beuth*, 1990.

[3] M. T. Mason, "Compliance and force control for computer controlled manipulators," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 11, no. 6, pp. 418–432, 1981.

[4] J. De Schutter and H. Van Brussel, "Compliant robot motion i. a formalism for specifying compliant motion tasks," *The International Journal of Robotics Research*, vol. 7, no. 4, pp. 3–17, 1988.

[5] T. Hasegawa, T. Suehiro, and K. Takase, "A model-based manipulation system with skill-based execution," *IEEE Transactions on Robotics and Automation*, vol. 8, no. 5, pp. 535–544, 1992.

[6] H. Mosemann and F. M. Wahl, "Automatic decomposition of planned assembly sequences into skill primitives," *IEEE Transactions on Robotics and Automation*, vol. 17, no. 5, pp. 709–718, 2001.

[7] R. Smits, T. De Laet, K. Claes, H. Bruyninckx, and J. De Schutter, "iTASC: a tool for multi-sensor integration in robot manipulation," in *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2008, pp. 426–433.

[8] R. Peter Bonasso, R. James Firby, E. Gat, D. Kortenkamp, D. P. Miller, and M. G. Slack, "Experiences with an architecture for intelligent, reactive agents," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 9, no. 2-3, pp. 237–256, 1997.

[9] T. Kollar, S. Tellex, D. Roy, and N. Roy, "Toward understanding natural language directions," in *IEEE International Conference on Human-Robot Interaction*, 2010, pp. 259–266.

[10] S. Lauria, G. Bugmann, T. Kyriacou, and E. Klein, "Mobile robot programming using natural language," *Robotics and Autonomous Systems*, vol. 38, no. 3, pp. 171–181, 2002.

[11] C. Matuszek, E. Herbst, L. Zettlemoyer, and D. Fox, "Learning to parse natural language commands to a robot control system," in *Experimental Robotics*. Springer, 2013, pp. 403–415.

[12] T. Laengle, T. C. Lueth, E. Stopp, G. Herzog, and G. Kamstrup, "KANTRA - a natural language interface for intelligent robots," in *Intelligent Autonomous Systems*, 1995, pp. 357–364.

[13] A. Knoll, B. Hildenbrandt, and J. Zhang, "Instructing cooperating assembly robots through situated dialogues in natural language," in *IEEE International Conference on Robotics and Automation*, vol. 1. IEEE, 1997, pp. 888–894.

[14] J. N. Pires, "Robot-by-voice: experiments on commanding an industrial robot using the human voice," *Industrial Robot: An International Journal*, vol. 32, no. 6, pp. 505–511, 2005.

[15] M. Tenorth, D. Nyga, and M. Beetz, "Understanding and executing instructions for everyday manipulation tasks from the world wide web," in *IEEE International Conference on Robotics and Automation*, 2010, pp. 1486–1491.

[16] M. Stenmark and P. Nugues, "Natural language programming of industrial robots," in *44th International Symposium on Robotics*, 2013.

[17] M. Stenmark and J. Malec, "Knowledge-based industrial robotics," in *Twelfth Scandinavian Conference on Artificial Intelligence*, 2013.

[18] International Organization of Standardization, *ISO Standards Handbook: Quantifies and Units*, 1993.

[19] J. Awrejcewicz, *Classical Mechanics: Kinematics and Statics*, ser. Advances in Mechanics and Mathematics. Springer, 2012.

[20] G. Pahl, W. Beitz, J. Feldhusen, and K. Grote, *Engineering Design: A Systematic Approach*. Springer, 2007.

[21] T. Herbst, *A Valency Dictionary of English: A Corpus-based Analysis of the Complementation Patterns of English Verbs, Nouns and Adjectives*. Gruyter, 2004.

[22] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2003, pp. 423–430.

[23] T. Kröger and B. Finkemeyer, "Robot motion control during abrupt switchings between manipulation primitives," in *Workshop on Mobile Manipulation at the IEEE International Conference on Robotics and Automation*, 2011.