SEVENTH FRAMEWORK PROGRAMME

Deliverable number:	D5.2
Deliverable Title:	Benchmark and performance index definition
Type (Internal, Restricted, Public):	PU
Authors:	C. Schou, O. Madsen, J. Rytz, D.Vitkute-Adzgauskiene H. Langer, D. Nyga, M.Tamosiunaite, F. Wörgötter
Contributing Partners:	AAU, UoB, SDU, VMU, UGOE

ACAT

Project acronym:	ACAT
Project Type:	STREP
Project Title:	Learning and Execution of Action Categories
Contract Number:	600578
Starting Date:	01-03-2012
Ending Date:	28-02-2015

Contractual Date of Delivery to the EC:	31-08-2013
Actual Date of Delivery to the EC:	30-08-2013

Content

1. Executive summary

This document presents benchmarks and key performance indicators for the ACAT system. First we give short description of demonstrator scenarios along the lines of which the ACAT system performance is to be measured. Then we define four categories of key performance indicators: "Overall System", "Process Memory Formation", "Compilation and Action Detailing" as well as "Knowledge and information content". We provide descriptions of the key performance indicators, as well as briefly define a procedure how each indicator will be measured.

2. Introduction

This document presents benchmarks and key performance indicators for the ACAT system. The goal of this document is to identify and specify key performance indicators both for the ACAT system as a whole and for most important subsystems, focusing on the core scientific questions of this project. Hence many of these benchmark indicators are somewhat unconventional. In general, these key performance indicators will help to assess the success of the ACAT approach. For each of these indicators the method of measuring it will also be described. Also, given the basic-research oriented scope of the ACAT project it may happen that some of these indicators cannot be used in the end and have to be discarded and/or replaced by others.

A description of two demonstrator scenarios and related instruction sheets is presented in D5.1. These scenarios form the main benchmarks for the entire ACAT system. As these scenarios are thoroughly described in D5.1 only a brief overview will be presented in this document. The performance indicators are listed in tables and are numerated according to subsystem.

3. Main Benchmark Scenarios

In ACAT two main demonstrators will form the main benchmarks of the ACAT system. These two scenarios, IASSES and CHEMLAB, are thoroughly described in D5.1, but will be briefly presented in this section, too.

3.1. IASSES Scenario

The IASSES scenario will focus on manufacturing tasks from the production of rotors for submersible pumps at the SQ-factory at the Danish company Grundfos. The production environment from Grundfos and thus the selected tasks will be replicated at Aalborg University. At the moment two benchmarks have been planned in relation to the IASSES scenario.

3.1.1. Rotor Cap Collection Benchmark

The goal of this benchmark is to pick a cylindrical component called a rotor cap from a conveyor belt and place it in a fixture on the robot, see Figure 1.

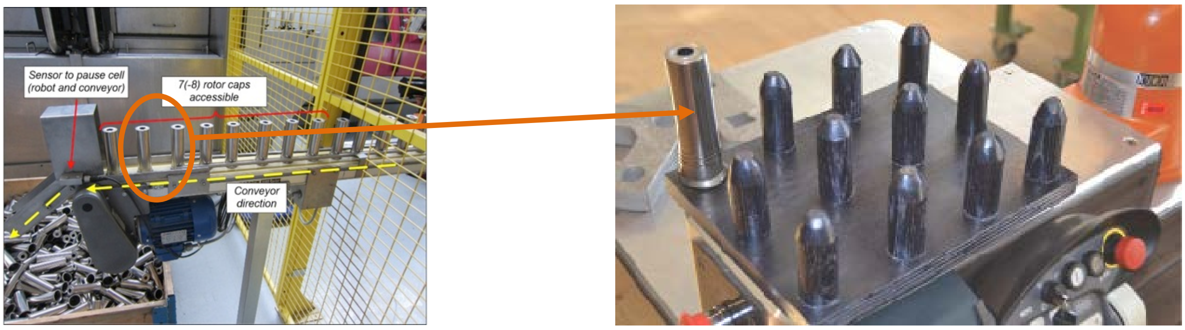
Sensor to pause cell (robot and conveyor)
7(-8) rotor caps accessible
Conveyor direction

Figure 1*: Rotor caps are initially located on a conveyor that moves one step approx. each minute. The rotor caps have to be moved from the conveyor before they fall into a bin, and afterwards they have to be placed in a fixture.*

In order to pick the rotor caps, the robot must first turn off the conveyor belt using a small switch on the conveyor.

3.1.2. Rotor Assembly Benchmark

In the rotor assembly benchmark the rotor for the electrical motor of a SQFlex submersible pump is assembled from the components shown in Figure 2.

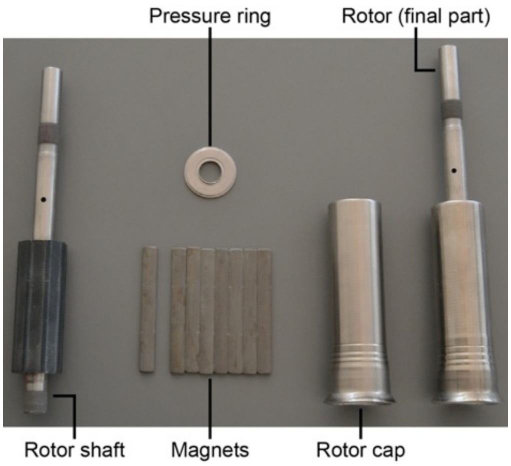

Figure 2*: Overview of the components used in the assembly of the SQFlex rotor. 1x rotor shaft, 1x pressure ring, 8x magnets, and 1x rotor cap are assembled into the SQFlex rotor.*

The task is carried out at a workstation containing a hydraulic press, see Figure 3. This workstation is already available as a replica at Aalborg University, see Figure 3.

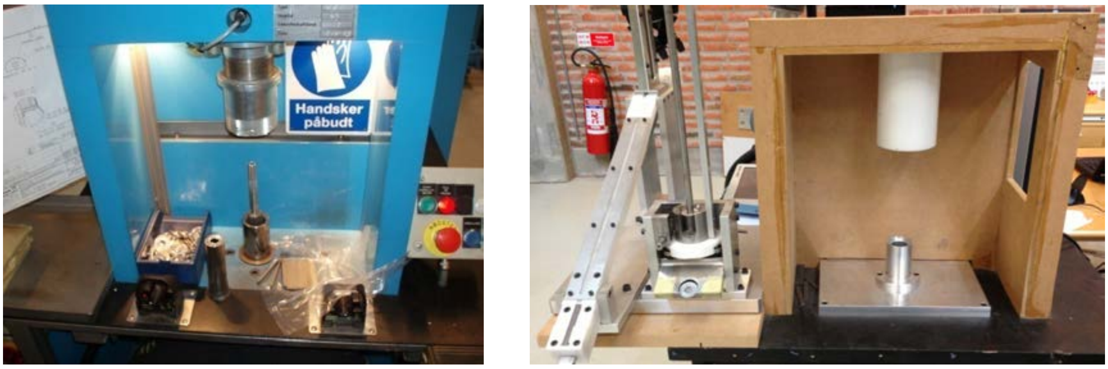

Figure 3*: Left: Workstation containing a hydraulic press at which the SQFlex rotor is assembled at Grundfos. Right: Replication of this workstation at Aalborg University.*

The manual task of assembling the rotor at Grundfos is carried out as follows:

1. The pressure ring is mounted onto the rotor shaft before it is placed into the press' fixture.
2. Eight magnets are placed on the sides of the rotor shaft. These magnets must be correctly oriented and aligned with the octagonal shape of the core (part of the rotor shaft) in order to fit into the fixture.
3. The rotor cap is placed on top of the fixture with the rotor axle sticking through the center hole. Due to limited clearance above the rotor axle the rotor cap must be tilted when placed on the rotor axle.

4. The press is then activated.
5. Afterwards the pressed rotor is removed from the press. Again the clearance above the pressed rotor is limited, and the rotor must be tilted in order to be removed.
6. The pressed rotor is placed in a moveable fixture plate which holds a number of units.

3.2. CHEMLAB Scenario

The selected scenario is the process of DNA extraction from a sample. The process involves the handling of liquids (pouring, decanting, etc.) and usage of standard laboratory equipment such as jars of different size and shape, filter cartridges, and a centrifuge. In order to be successful the process has to be executed under the required constraints (temperature, time schedule, etc.) stated in the respective lab protocol.

Success can be validated easily: the result of the successful process is a visible DNA pellet. All sub-processes involved either have an intermediate result which can be defined precisely or it can be observed directly if the sequence of actions is executed appropriately (e.g., with the required amounts of substances and according to the time constraints).

Our research focus, however, is not on the success of physical process itself. Our focus is rather on the planning, reasoning, and knowledge representation problems that have to be solved in order to enable the robot to master this particular task as well as other related ones. Therefore, it is not sufficient to define the benchmark in terms of the question if the robot finally happens to extract some DNA, or not. The benchmark should explicitly reflect if the robot succeeds because of its ability to *understand the content* of the instructions given in the lab protocol, to combine these typically underspecified and vague information with appropriate *background knowledge* (both domain-specific and commonsense), and to *reason* on the basis of this integrated knowledge.

The CHEMLAB scenario will be implemented at University of Bremen on a PR2 robot, see Figure 4. The implementation on the robot platform will demonstrate if the plans derived from the vaguely formulated, underspecified instructions are sufficient for successful execution in a real-world environment, and how strong the impact of the level of abstractness in the plans on the robot's performance is.

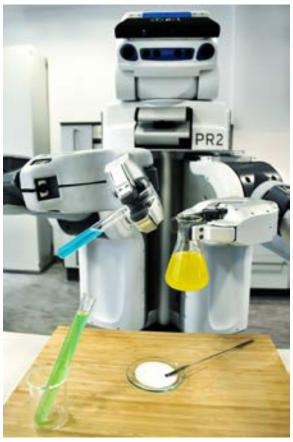

Figure 4*: The PR2 robot at University of Bremen that will be used for the CHEMLAB scenario.*

4. Key Performance Indicators

This section presents the key performance indicators for the ACAT system. These are formulated in four main subsections: Overall system benchmark (section 4.1), process memory formation benchmark (4.2), compilation benchmark (4.3) and benchmark on knowledge and information contents (4.4).

4.1. Overall System

The following key performance indicators show the performance of the system as a whole.

Key Performance Indicators:

1.1 ***Name*** **Setup time for a new task**

Description Total setup time on a new task (instruction sheet)

Measurement Feed the ACAT system a new (unknown) instruction sheet and measure the setup time until task is ready for execution (measured in seconds)

1.2 ***Name*** **Robustness during setup**

Description Robustness of the system when processing a new instruction sheet

Measurement Feed the ACAT system multiple new (unknown) instruction sheets and measure the percentage of successful established task sequences. A task sequence is successful if it achieves the specified goal of the task (end state)

1.3 ***Name*** **Robustness during execution**

Description Robustness of the execution phase

Measurement During multiple executions of a task instantiated from an instruction sheet, measure the successful completions of the task.

1.4 ***Name*** **Cycle time during execution**

Description The cycle time of the task during execution

Measurement During multiple executions of a task instantiated from an instruction sheet, measure the mean cycle time. This could be compared to other task instantiation methods

4.2. Process Memory Formation

Process memory formation is a data-driven process where topic related text and image material is analyzed in order to extract action verbs and verb-associated objects thus forming a back-bone for Action Category formation. The memory is grounded by robot control data used for robot execution, where robot control data is available only for a sub-set of textual entries. How to generalize execution over the textual entries where controls are not directly known is the research question of the ACAT project.

Topic-related data for action verb and verb-associated object extraction includes domain-related texts and image databases. Texts are compiled as domain-specific corpora, covering both domain-specific texts crawled from internet, as well as other available domain specific material – instruction sheets, process documentation, etc. Wherever possible, texts are supplemented with adequate metadata, aiming at more efficient analysis process. Such metadata may include linguistic annotations (e.g. morphosyntactic features for text), information on text and video alignment, etc.

Both verbs and verb-associated objects are extracted from domain-specific data collections using semi-automatic procedures. Automation of the extraction process is based on the application of different natural language technologies (NLP). Additional semantic information in the form of supplementary databases and ontologies is employed for building a domain-oriented ontology of action verbs and associated objects, suitable for automated completion of instruction sheets in the following steps. Both action verbs and terms denoting verb-associated objects are grouped into synsets, i.e. groups of terms with the same meaning.

The overall action verb and action-related object extraction process, i.e. the domain-specific action ontology building process if presented in Fig. 5.

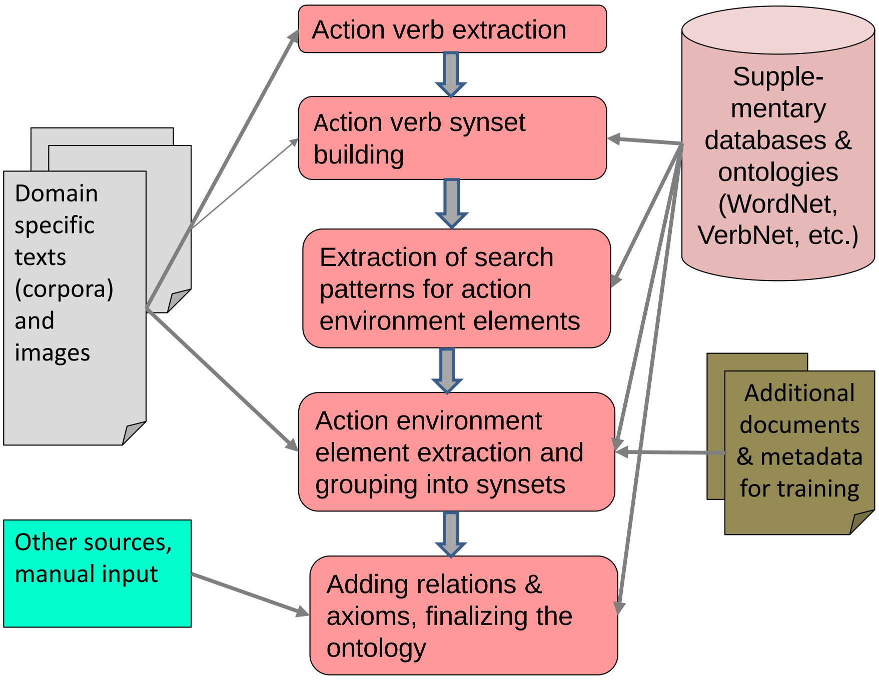
Action verb extraction
Action verb synset building
Extraction of search patterns for action environment elements
Action environment element extraction and grouping into synsets
Adding relations & axioms, finalizing the ontology
Domain specific texts (corpora) and images
Other sources, manual input
Supple-mentary databases & ontologies (WordNet, VerbNet, etc.)
Additional documents & metadata for training

Fig.5. The action verb and action-related object extraction process

The robot control data is collected from actual robot execution and stored in association with corresponding linguistic data.

Key Performance Indicators:

2.1 ***Name*** **Linguistic action ontology**

Description Number of action verbs in the ontology and number of synsets.

Measurement Determined by the number of action verbs and synsets available in the process memory by the end of the project.

2.2 ***Name*** **Object categories**

Description Number of object categories saved in the process memory

Determined by the number of object categories in the linguistic object ontology and the number of associated object images/models available to aid recognition and robotic manipulation of those objects.

2.3 *Name* **Number of action grounding instances**

Description Number of robot execution/control instances stored in the process memory

Measurement Determined by the number of robot execution/control instances stored in the process memory by the end of the project.

2.4 *Name* **Action categories**

Description Number of action categories saved in the process memory

Measurement Determined by the number of action categories available in the process memory by the end of the project.

4.3. Compilation and Action Detailing

Compilation and action detailing is the process of creating a Formal Instruction Representation that provides full information required for the planning and execution process. The main steps in this process are: Take an instruction from an instruction sheet, fill in missing information, extract the corresponding Action-Category from the library and add all the information that is needed to actually execute the action.

Key Performance Indicators:

3.1 *Name* **Setup time**

Description Time in seconds of human intervention to set-up a system for instruction sheet compilation.

Measurement Measure the time it takes a human to setup the system for compilation of a new instruction sheet.

3.2 *Name* **Instruction sheet compilation time**

Description The time used for compilation of a new instruction sheet into action categories.

Measurement Measure the time it takes to compile multiple new (unknown) instruction sheets into action categories.

3.3 *Name* **Human intervention**

Description The level of human intervention needed in the compilation phase.

Measurement Measured by the time (seconds) of human intervention need during the compilation of a new instruction sheet.

3.4 *Name* **Decomposing complex verbs**

Description Assessment of how well the text compiler decomposes complex verbs into simple verb sequences.

Measurement Measured by number of un-recognized, un-progressed or failed verbs of a new instruction sheet.

3.5 *Name* **Number of instruction sheets compiled**

Description The total number of instruction sheets compiled throughout the ACAT project.

Measurement Measured by number of different or variations of instruction sheets compiled.

4.4. Knowledge and Information Content

This section lists key performance indicators for the knowledge and information content of cognition-enabled robot control systems.

Benchmarking the performance and, in particular, the cognitive capabilities of a cognition-enabled robot directly is difficult. This is because many measurable performance factors (time needed to complete a task, success rates for tasks, etc.) also depend on other factors such as the quality of sensors and actuators, which are beyond the scope of the ACAT project. A more objective performance measure could therefore be the *knowledge and information content* of cognition-enabled robot control systems.

The information content of robot control systems can be assessed through *query-based benchmarking*. To this end, we will develop structured libraries of queries in a formalized representation language that require different kinds of cognitive capabilities. The proposed library of queries for benchmarking will include queries that test the following capabilities:

1. The competent interpretation of vaguely, ambiguously, and incompletely formulated tasks (e.g., lab protocols).

2. The successful answering of queries regarding what the robot has done, how, and why, and what it is capable of accomplishing.

3. The robot’s ability to answer queries about the lab environment it is to operate in and the equipment it uses.

4. The ability to understand scenes and the ability to form memories and process models of the environment.

5. The ability to answer queries about the expected consequences of actions depending on the action parameterizations and the contexts they are executed in.

Hence, the benchmarks can be specified by a set of queries the robot can answer. The queries will be formulated in a prolog-based query language, but are given here in plain English for better readability. The "Knowledge and Information Content" benchmarking procedures will be implemented only for the CHEMLAB scenario.

Key Performance Indicators:

4.1	***Name***	**Causal relations**
	Description	A set of relevant causal relations in instructions correctly understood by a robot, even if these are implicit
	Measurement	The list of extracted causal relation together with a short statistical summary on the list entries will be provided. For example, the implicit causal relation between the application of vacuum and the drain of the liquid in the instruction: "Apply vacuum until all liquid has drained" should be detected. Hence, the robot should be able to answer queries such as: - What is the effect of applying vacuum? - What caused the drain of the liquid? - What happened if we didn't apply vacuum?

- What is the difference between the state before applying vacuum and the state after having it applied?

4.2	***Name***	**Vague quantities**
	Description	A set of correct instances when inferring vaguely formulated quantities
	Measurement	A list of inferred instances of vaguely formulated quantities together with a short statistical summary on the list entries will be provided. Example: "Mix gently by inverting 4-6 times" - The robot should know that inverting 4, 5, and 6 times a proper executions of the task description.

4.3	***Name***	**Missing objects**
	Description	A set of examples with inferred missing objects and roles
	Measurement	A list of inferred missing object together with a short statistical summary on the list entries will be provided. Example: "Mix gently by inverting 4-6 times" - The robot should infer that the missing direct object of the action verb "invert" refers to the sample jar which it is manipulating. Example: “Add 4 drops of hydrochloric acid” - The robot should infer that the role *instrument* is not specified and that a pipette can serve as an appropriate object for filling this action parameter

4.4	***Name***	**Disambiguation**
	Description	A set of instances in inferring correct meanings of ambiguous words
	Measurement	A list of inferred meaning of ambiguous words together with a short statistical summary on the list entries will be provided. Example: "Add a scoop of X to Y" - The robot should infer that “scoop” refers to the amount of substance X to be added to Y and not the physical object of a scoop, which might serve as an instrument for accomplishing this task.

5. Conclusions

Four groups of performance indicators, important for the evaluation of the ACAT project, have been defined. We have specified the key performance indicators, as well their measurement procedures. As exact measurement procedure (including all contextual details) cannot be defined at this stage of the project, only general guidelines how to approach parameter measurement are given. E.g. some of the performance indicators are given in the case of a new scenario (e.g. indicator 1.1), and the values will depend on how much the new scenario differs from the one that was already implemented before. Or e.g. questions like in indicator 4.1 can be formulated only knowing the full coverage of the process memory of the ACAT system. In highly experimental project as ACAT such details are not known in advance thus only the general guidelines for measuring the key performance indicators have been provided here. Hence some of this may change in the course of the project.