Deliverable number:	D1.4
Deliverable Title:	Text corpora and image databases – Update of D1.1
Type (Internal, Restricted, Public):	PU
Authors:	Daiva Vitkute-Adzgauskiene, Irena Markievicz, Tomas Krilavicius, Minija Tamosiunaite
Contributing Partners:	VDU, UGOE

Project acronym:	ACAT
Project Type:	STREP
Project Title:	Learning and Execution of Action Categories
Contract Number:	600578
Starting Date:	01-03-2013
Ending Date:	29-02-2016

Contractual Date of Delivery to the EC:	29-02-2016
Actual Date of Delivery to the EC:	29-02-2016

Content

1. Executive summary

This document presents a summary of ACAT task-specific corpora development, since the situation described in the deliverable D1.1. We give a description of the procedures used for additional corpus data acquisition, processing and storing, as well as the description of data sources used. Finally, we describe the structure of the updated corpora and the results of corpus content analysis.

ACAT corpora texts are used in order to extract action verbs and verb-associated objects thus forming a backbone for Action Category formation. Action verbs and associated object names are stored in the ACAT ontology, and, further, are matched against semantic roles in robot instruction texts, thus allowing extracting action-related details in the form of Action Data Tables (ADTs).

2. Introduction

This document presents a summary of ACAT task-specific corpora development, since the situation described in the deliverable D1.1.

The ACAT corpora data are accumulated for two main scenarios – CHEMLAB and IASSES. Procedures, used for data accumulation are similar for both scenarios, while the main difference lies in the sources used. CHEMLAB and IASSES scenarios are thoroughly described in D5.1.

The primary CHEMLAB 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. However, information on performing general chemical experiments is also used in order to have better coverage of necessary actions in performing chemical

The IASSES scenario focuses on manufacturing tasks from the production of rotors for submersible pumps at the SQ-factory at the Danish company Grundfos.

Both the accumulated corpora and image databases are stored on a subversion (SVN) server, dedicated to the ACAT project (URL http://kleinas.vdu.lt/svn/ACAT-416859).

The goal of this document is to present the final status of corpora accumulated for two ACAT project scenarios – CHEMLAB and IASSES. Structure and basic statistical data for the accumulated ACAT corpora is provided, in line with the description of the procedures and sources used for data acquisition and preparation.

3. Corpus status summary for D.1.1

3.1. ACAT corpus formation: procedures

In the first stage of ACAT corpus development, both crawling of freely available Internet resources and use of specific domain-focused document databases were employed for corpus data acquisition. Crawling was executed by applying a focused crawler, using domain-specific keyword lists, accumulated by applying pre-analysis of domain specific texts and expert input, and an URL list at its input.

Domain-specific document databases, used for corpus text collection, included available industrial databases with task specific manuals, training materials and scientific papers.

In the first processing layer, corpus data was cleaned using boilerplate removal schemes, i.e. detecting and removing the surplus "clutter" (HTML tags, templates) around the main textual content. PDF

documents were converted to the plain-text format, using *PDFBox* tools (http://pdfbox.apache.org/). *PDFBox* is an open source Java PDF library for working with PDF documents.

Plain-text, obtained as the result of the first processing layer, was supplied to the second processing layer, dedicated to additional text filtering. In this layer, keyword lists and stop-word lists are used to filter out texts, which are irrelevant or weakly linked to the domain.

Finally, morphological annotation of the corpus texts was accomplished using Stanford tools for morphological analysis (http://nlp.stanford.edu/software/). This annotation level is obligatory in order to be able to identify action verbs and action object categories in the text. The final result of this processing level is an XML document for each corpus text.

For the CHEMLAB scenario, Internet crawling was used as the main source of information. The process was executed in iterations, with the resulting lists of extracted keywords from one iteration used as focusing info for the next iteration. Also, documents with tutorials and training material on chemical and biotechnological experiments were used, as well as scientific documents from the PUBMED database (http://www.ncbi.nlm.nih.gov/pubmed).

For the IASSES scenario, the main sources of information are collections of different manual documents in PDF format.

3.2. ACAT Corpus structure

The size of the accumulated **CHEMLAB** corpus at the time of D1.1 was **8919087** running words. It was structured in the following way:

- 33,84% (3018220 running words) - general chemistry texts, crawled from internet, mainly tutorials for chemical experiments;
- 41.51% (3702313 running words) - biochemistry and biotechnology texts, crawled from the Internet;
- 24.65% (2198554 running words) - biochemistry and biotechnology texts from PUBMED electronic library.

The size of the accumulated **IASSES** corpus at the time of D1.1 was **3563775** running words. It consists of manuals, assembly instructions and descriptions, crawled from the Internet and obtained from project partner document libraries.

Both CHEMLAB and IASSES corpora are available in two formats:

1) Plain text format;
2) XML format with morphological tags added.

The morphological tags for CHEMLAB and IASSES corpora were formed, using Stanford annotation tools and POS (Part-Of-Speech Tagging Guidelines), designed for the Pen Treebank Tagging Project (http://repository.upenn.edu/cis reports/570).

The list of tags used for morphological annotation is presented in Table 1.

Table 1. *Tags used for morphological annotation of ACAT corpus data (source:* http://repository.upenn.edu/cis_reports/570)

Coordinating conjunction	CC
Cardinal number	CD
Determiner	DT
Existential there	EX
Foreign word	FW
Preposition or subordinating conjunction	IN
Adjective	JJ
Adjective, comparative	JJR
Adjective, superlative	JJS
List item marker	LS
Modal	MD
Noun, singular or mass	NN
Noun, plural	NNS
Proper noun, singular	NP
Proper noun, plural	NPS
Predeterminer	PDT
Possessive ending	POS
Personal pronoun	PP
Possessive pronoun	PP$
Adverb	RB
Adverb, comparative	RBR
Adverb, superlative	RBS
Particle	RP
Symbol	SYM
to	TO
Interjection	UH
Verb, base form	VB
Verb, past tense	VBD
Verb, gerund or present participle	VBG
Verb, past ~articiple	VBN
Verb, non-3rd person singular present	VBP
Verb, 3rd person singular present	VBZ
Wh-determiner	WDT
Wh-pronoun	WP
Possessive wh-pronoun	WP$
Wh-adverb	WRB

4. Corpus development since D.1.1

Further corpus development was aimed at both expanding the corpus with texts relevant for IASSESS and CHEMLAB scenarios, and focusing of the previously collected corpus texts to achieve higher degree of relevance to ACAT experiment field.

4.1. Expanding corpora with transcribed experiment data

At this stage, video data on CHEMLAB and IASSESS scenario documents was used for updating the ACAT corpora. Videos were transcribed in order to use action and action object name extraction procedures.

The following videos were processed for the CHEMLAB domain (11):

- Plasmid DNA Extraction (Miniprep)
- Plasmid DNA Extraction (Traditional Alkaline Lysis method)
- Plasmid DNA Extraction (Midiprep)
- Plasmid DNA Extraction (Megaprep)
- DNA Extraction by Phenol/Chloroform
- Plasmid DNA Extraction (CsCI Isolation)
- DNA Extraction from Mouse Tail (Manual)
- DNA Miniprep (BioBridge Labs)
- E.coli miniprep
- Lab Report DNA Isolation from a human blood sample
- Isolation of DNA from Human Cheek Cells

Further, by applying automated action word and action object word extraction procedures, 75 distinct action words and 49 main action-related object names were obtained from the added corpora texts for the inclusion to the ACAT CHEMLAB ontology.

The following videos were processed for the IASSES domain (14):

- Rotor Shaft Assembly using the KUKA LWR
- Rotor Assembly in KRL using the KUKA LWR
- Rotor Assembly using the KUKA LWR
- Mobile robots working in cooperation in an industrial environment at Automatica 2014
- TAPAS M39: Testing and validation of RTD work at Grundfos A/S pump production
- TAPAS M24-27 demo at Grundfos, January 2013
- TAPAS - Month 24 demonstration: complex assembly, machine tending and quality inspection
- 7 skill descriptions from Skill Library (PegInHole, Rotate, Place, PlaceOnto, PlaceInto, Pick, MoveTo)

Further, by applying automated action word and action object word extraction procedures, 35 distinct action words and 38 main action-related object names were obtained from the added corpora texts for the inclusion to the ACAT IASSES ontology.

4.2. Focused corpora

In order to achieve higher degree of relevance to ACAT experiments, focusing of the accumulated ACAT corpus texts was done. A set of action words and action object names acquired from the transcribed videos was used for constructing keys for focusing.

All focused data were used in the ontology learning process, to add additional action, main object, primary object and secondary object information to the ontology.

The algorithm for ACAT corpora focusing is presented in Fig.1.

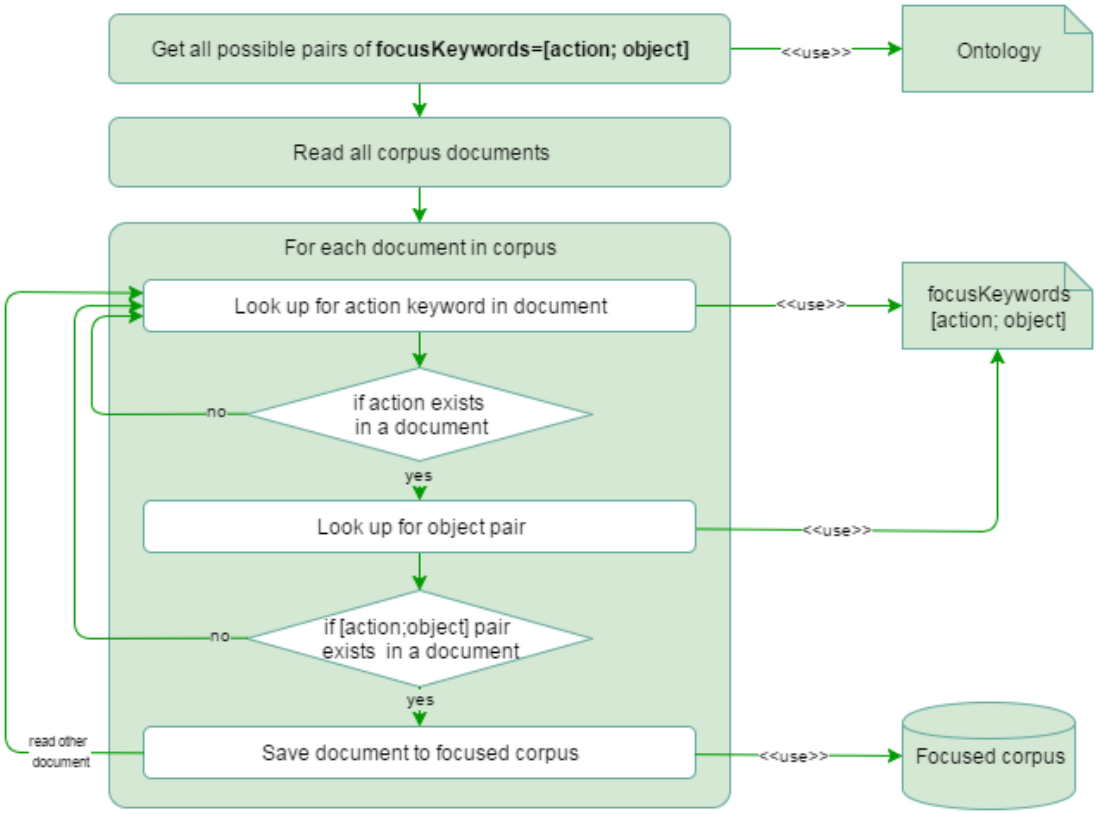

Fig. 1. *Algorithm for ACAT corpora focusing*

The main steps for the corpora focusing algorithm are:

1) A so-called core ontology was built from action words and action object names acquired from the transcribed videos, also adding synonym names from the WordNet ontology.
2) Possible combinations of action words and action object names from the core ontology were constructed. At least 1 action word and 1 action object word was necessary. The selected combinations were considered to be focusing keys.
3) The Internet-crawled texts of the corpora in D1.1 were filtered, leaving only texts containing at least one of the focusing keys.

The size of the **focused CHEMLAB** corpus is **3,074,259** running words. The size of the **focused IASSES** corpus is **240,892** running words (Table 2).

Table 2. *Focused CHEMLAB corpus and focused IASSES corpus statistics*

Focused CHEMLAB corpus			**Focused IASSES corpus**	
Unique words	44,161		Unique words	26,580
Unique lemmas	34,093		Unique lemmas	20,210
Running words	3,074,259		Running words	240,892

In comparison with data crawled from the Internet, focused corpora decreased: CHEMLAB – to **34.47%** of previous texts, IASSES – **6,76%** of previous data.

Below you can find the most common noun-keywords (Table 3) and verb-keywords (Table 4) from both CHEMLAB and IASSES focused corpora.

Table 3. *The most common noun-keywords from focused IASSES and CHEMLAB corpora*

Focused IASSES		**Focused CHEMLAB**	
Lemma	**Frequency**	**Lemma**	**Frequency**
rotor	885	Cell	7,209
cap	865	Dna	6,511
screw	736	Gene	5,298
lock	383	Protein	4,194
stator	445	Water	3,902
hex	136	Method	3,142
assembly	79	Process	3,057
bearing	78	Chemical	2,979
exciter	76	Material	2,957
generator	69	Molecule	2,729
engine	64	Research	2,614
diode	63	Sequence	2,325
flange	62	Element	2,318
blade	58	Structure	2,15
exciter	56	Level	2,149
nut	54	Acid	2,149
voltage	54	Chemistry	2,122
washer	45	Result	2,104
measure	44	Sample	2,053
lead	44	Plant	2,04

Table 4. *The most common verb-keywords from focused IASSES and CHEMLAB corpora*

CHEMLAB		IASSES	
Lemma	**Frequency**	**Lemma**	**Frequency**
Have	25,048	Remove	789
Use	11,394	Replace	735
Do	9,664	Install	522
Make	6,611	Screw	420
Include	4,866	Use	208
Find	4,482	Apply	84
Provide	4,391	Lose	53
Take	3,934	Attach	21
Get	2,872	Ground	19
Give	2,745	Connect	18
Produce	2,263	Do	17
Add	2,249	Lock	14
Contain	2,014	Require	14
Leave	1,813	Rotate	14
require	1,668	Push	13
allow	1,538	Include	12
form	1,521	Disconnect	12
create	1,507	Turn	11
hold	1,494	Consist	10
increase	1,357	Make	10
bring	1,35	Hold	9
start	1,289	Insert	9

5. Conclusions

The corpora compiled for the CHEMLAB and IASSES scenarios, serve as the basis for symbolic background knowledge formation. Further development of ACAT corpora since the situation described in D1.1, allowed us to build corpora focused on CHEMLAB and IASSES scenario specifics.

The texts in the domain-specific ACAT corpora make it possible to develop the ACAT ontology, which is used for storing keywords for actions and action related objects. The ACAT ontology data is employed by the textual instruction compiler by matching this data against semantic roles identified in robot instruction texts in order to extract action-related details in the form of Action Data Tables (ADTs).

Corpora can be continuously updated by adding additional CHEMLAB and IASSES related texts whenever they become available.

6. References

1. Markievicz, Irena; Kapočiūtė-Dzikienė, Jurgita; Tamošiūnaitė, Minija; Vitkutė-Adžgauskienė, Daiva. Action classification in action ontology building using robot-specific texts // Information technology and control. Kaunas: Technologija. ISSN 1392-124X. 2015, t. 44, nr. 2, p. 155-164. Internet access: <http://www.itc.ktu.lt/index.php/ITC/article/view/7322/6881>. [Databases: Science Citation Index Expanded (Web of Science); INSPEC; VINITI; Scopus]; [Citation index: 0,623(F) (2014)].

2. Markievicz, Irena, Daiva Vitkute-Adzgauskiene, and Minija Tamosiunaite. "Ontology learning in practice – using semantics for knowledge grounding". *"E-learning as a Socio-Cultural System: A Multidimensional Analysis"*, IGI Global book series Advances in Educational Technologies and Instructional Design", ISSN: 2326-8905. 2014, pp.158-171.

3. Markievicz, Irena, Daiva Vitkute-Adzgauskiene, and Minija Tamosiunaite. "Semi-supervised Learning of Action Ontology from Domain-Specific Corpora."*Information and Software Technologies*. Springer Berlin Heidelberg, 2013. 173-185.