PROJECT PERIODIC REPORT	
Czech-Norwegian Research Programme (CZ09)	
Norwegian Financial Mechanism 2009-2014	
Programme area	**Bilateral Research Cooperation**
Periodic report	**2/2015**
Period covered	**1 January to 31 December 2015**
Project ID number	7F14047
Acronym	HaBiT
Project title in English	Harvesting big text data for under-resourced languages
Project title in Czech	Získávání velkých textových dat pro jazyky s nedostatečným množstvím jazykových zdrojů
Project Promoter (name, full address)	Masarykova univerzita Žerotínovo nám. 617/9, 601 77 Brno Czech Republic
Project Partner(s) (name, full address)	Norges teknisk-naturvitenskapelige universitet Høgskoleringen 1, 7491 Trondheim Norway
Full Name of Principal Investigator	doc. PhDr. Karel Pala, CSc.
Signature of Principal Investigator	
Statement	*I hereby declare that the information I state in the project periodic report is accurate, true and complete. I am aware that if the information has been reversed in the opposite, I will face sanctions from the Programme Operator.*
Done in	Brno, Czech Republic
Date	10/01/2016

On behalf of Project Promoter			
Stamp of Project Promoter			
Statutory authority of Project Promoter	**Name(s):**		
	Signature(s):		
	Position:		

1. GENERAL INFORMATION ABOUT PROJECT

1.1 Activity of research and development in project

Basic research	Applied research	Experimental development
☒ 0-100 %	☐ 0-100 %	☐ 0-100 %

Project Promoter: Masaryk university (MU)

Basic research	Applied research	Experimental development
☒ 0-100 %	☐ 0-100 %	☐ 0-100 %

Project partner: Norwegian University of Science and Technology (NTNU)

Basic research	Applied research	Experimental development
☒ 0-100 %	☐ 0-100 %	☐ 0-100 %

1.2 Official project starting date reported to Programme Operator (dd/mm/yyyy)

01/10/2014

1.3 Project duration in months in total (number, e.g. 30)

31

1.4 Total project costs in Project contract (in CZK)

24,468,000

1.4.1 Total grant in Project contract (in CZK)

24,468,000

1.4.1.1 Grant for reporting period approved by Programme Operator (in CZK)

9,430,000

1.5 Number of partners (including Project Promoter, e.g. 2)

2

1.6 Ethical issues (max ¼ page A4)

NO

2. SCIENTIFIC AND MANAGEMENT PART

2.1 Publishable summary in English (max. 3/4 page A4)

The main objectives of the HaBiT project are to gather large-scale text data (corpora) from the Web for under-resourced languages, involving Norwegian, partly Czech and the major languages of Ethiopia — Amharic, Afaan Oromo, Tigrinya, Somali — and to build shallow processing applications. The gathered data will be processed to make it usable in various language applications, such as information extraction or machine translation. Furthermore, in the process of collecting corpora data, existing tools for building web text resources will be further developed and improved since the Ethiopian languages are quite different from most European languages. Applications for the given languages will be built to allow for the separation and disambiguation of multiple senses of words.

One of the project's aims is to build a multi-billion word Norwegian corpus using the tools co-developed by the FI MU NLP Centre, Masaryk University and utilized in a previous joint EU-funded project with NTNU ("PRESEMT: Pattern REcognition-based Statistically Enhanced MT", 2010-2012). Second, NTNU collaborates with University of Oslo and the universities of Addis Ababa and Hawassa, Ethiopia, in a project to support linguistic resource building in Ethiopia funded by Norad ("Linguistic Capacity Building — tools for the inclusive development of Ethiopia", NORHED 2013-2018). It is natural to link these activities and to include processing of the four major languages in Ethiopia in the present project: the HaBiT project is able to feed into and leverage on the NORHED project, thoroughly testing the technologies and thus addressing also the call topics on technology assessment, verification and testing, as well as on ICT meeting societal challenges, hence obtaining a relevant added value also in the political respect through cooperation with a less-developed country.

After the specification of the corpora and corpus building modules and the requirements for word-sketch grammars together with the particular tools, that are used for collecting data, the main aims in the reported period have been devoted to the methods and techniques of building the starting versions of the HaBiT system components, which are currently presented in the first prototype of the whole HaBiT system. The components and respective output data concentrate on efficient processing of under-resourced language texts and building new linguistic modules and resources for these languages. The results of the project development are presented further in the periodic report.

Up-to-date information can be found at the project website www.habit-project.eu.

2.1.1 Publishable summary in Czech (max. 3/4 page A4)

Hlavním cílem projektu Habit je jednak shromáždit velká textová data (korpusy) z Webu pro jazyky s nedostatečnými zdroji včetně norštiny a částečně češtiny a také hlavní etiopské jazyky - amharštinu, afaan oromštinu, tingrinštinu, somálštinu a jednak vytvořit aplikace pro jejich mělké zpracování. Získaná data budou zpracována tak, aby mohla být použita v řadě jazykových aplikací, jako je extrakce informací a strojový překlad. V průběhu budování korpusových dat budou dále vyvíjeny a zlepšovány existující nástroje pro získávání webových textových zdrojů, protože etiopské jazyky se výrazně liší od od většiny evropských jazyků. Aplikace pro zmíněné jazyky budou budovány tak, aby dovolily zkoumat a vyčleňovat víceznačnost slovních významů.

Jedním z cílů projektu bude vytvořit více miliardový korpus norštiny, a to s použitím nástrojů spoluvyvinutých Centrem ZPJ FI Masarykovy univerzity a využitých v předchozím společném EU projektu realizovaném spolu s NTNU ("PRESEMT: Pattern REcognition-based Statistically Enhanced MT", 2010-2012). Za druhé, NTNU spolupracuje s Univerzitou v Oslo a dvěma etiopskými univerzitami v projektu podporujícím budování jazykových zdrojů v Etiopii a finančně zajištěným organizací Norad ("Linguistic Capacity Building – tools for the inclusive development of Ethiopia", NORHED 2013-2018).

Je přirozené propojit tyto aktivity a zahrnout zpracování čtyř hlavních etiopských jazyků do současného projektu: HaBiT project tím umožní datové vazby na NORHED projekt a důkladné testování technologií. Tak bude možno adresovat rovněž témata týkající se hodnocení technologií, jejich verifikování a testování včetně toho, jak informační technologie (ICT) splňují sociální výzvy. To umožní získat relevantní přidanou hodnotu také v politickém ohledu díky spoluprácí s méně rozvinutou zemí.

Po specifikaci korpusů a modulů pro budování korpusů a požadavků na *word-sketch* gramatiku spolu s nástroji používanými pro získávání dat byl výzkum a vývoj ve sledovaném období věnován metodám a technikám budování první verze systémových komponent, které jsou aktuálně k dispozici v prvním prototypu systému HaBiT. Komponenty a výstupní data se soustřeďují na efektivní zpracování textů jazyků s nedostatečným množstvím jazykových zdrojů a na budování nových lingvistických modulů a zdrojů pro tyto jazyky. Jednotlivé výsledky výzkumu realizovaného aktuálně v projektu jsou podrobně prezentovány v dílčí zprávě.

Aktuální informace lze najít na webové stránce projektu: www.habit-project.eu.

2.2 Project objectives for reporting period (max. ½ page A4)

In the reported period the main project objectives fall into the following work packages and the respective tasks:

- WP1: System integration
- WP2: Multi-billion word corpus building
- WP3: Corpora for under-resourced languages
- WP4: Shallow processing grammars and tools
- WP5: Multi-sense and multi-lingual word spaces
- WP6: Requirements and evaluation

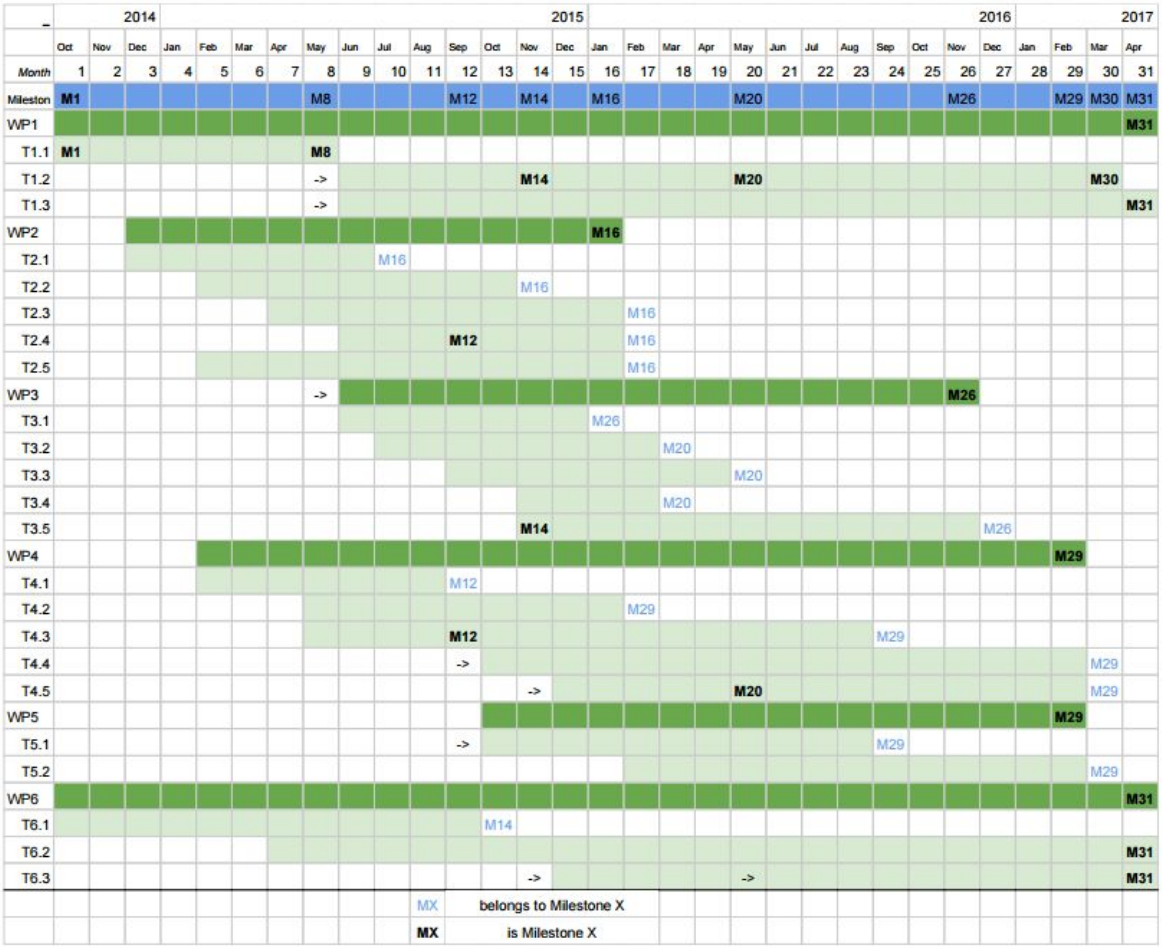
The interim results of the project are presented in the following deliverables as planned for the reported year 2015:

- D1.1.1 System specifications: Overall system design definitions
- D1.1.2 Specification of corpora and the corpus building module
- D1.1.3 Specification of word-sketch grammars and tools
- D1.1.4 Specification of the semantic content matching and wordspace module
- D1.2.1 The HaBiT system v1: First integrated system prototype
- D4.1: Methodology of Sketch Grammar evaluation
- D6.1 Project evaluation plan

The outputs of the project were presented at scientific conferences, a comprehensive list with detailed descriptions is stated further in the section 2.4.5 Project output(s).

The texts of deliverables are available at the project website
http://habit-project.eu/wiki/InterimResults

2.3 Work progress and achievements during reporting period (max. 3 pages A4)



The overall workflow of the respective tasks conducted during the reported period is graphically displayed in the project Gantt chart above. The main efforts were devoted to analysis and implementation of new techniques and methods for gathering and shallow linguistic processing of new large corpora of the respective languages, i.e., the four largest Ethiopian languages (Amharic, Oromo, Tigrinya and Somali), Czech and Norwegian. The current achievements in these tasks are detailed further in this section, split to individual work packages.

2.3.1 Description of project implementation (max. 2 pages A4)

Work Package 1

The task T.1.1 followed the Kick-Off Meeting and we have defined the crucial specifications of individual system parts. The most important are D1.1.1 System specifications containing the overall system design and D1.1.2 Specification of corpora and the corpus building module. There are two additional specifications D1.1.3 and D1.1.4 devoted to specialized area of the project. According to the D1.1.1 the first prototype was built (D1.2.1). During the work on the first prototype, we have slightly refined the specifications D1.1.1 and D1.1.2. We suppose that there could be more revisions of the specifications.

Work Package 2

We have used the SpiderLing web crawler to build Czech and Norwegian corpora. To create bigger corpora than before we have made several modifications to the SpiderLing tool, for example relaxing the original one domain restriction. We have also added new language models for Norwegian, because we wanted to distinguish Bokmål and Nynorsk Norwegian varieties. The first version of created corpora are incorporated into the first HaBiT prototype.

Work Package 3

We are using the same SpiderLing tool for creating corpora in four Ethiopian languages. There are the first versions of the Amharic, Oromo, Somali and Tigrinya corpora already in the system. We have decided to use a different strategy to build corpora for some languages. The SpiderLing is going to crawl the Ethiopian Web without the limit to one language (which is the typical usage otherwise) and filter text for each language in the next stage. We have also investigated methods to automatically annotate multi-lingual and code-mixed corpora.

Work Package 4

The work on this packages started with investigating the evaluation strategies for Sketch Grammars, because for some languages there are several versions of Sketch Grammars and it is not clear which one is better. The result is D4.1 Methodology of Sketch Grammar evaluation. We are now working on evaluation of existing Sketch Grammars and also on tools for helping development of new and revision of existing Sketch Grammars.

Work Package 5

Within this WP we have investigated the use of unlabeled texts to create auxiliary language models, that e.g. can be used in domain adaptation. Late in 2015, we started work on developing matching logic for mono- and multi-lingual word space models.

Work Package 6

The output of this work package is D6.1 Project evaluation plan. The work continues on usabilities studies.

2.3.2 Milestones achievement

No.	Milestone title	WP no.	Lead partner (abbreviation)	Planned achievement date dd/mm/yyyy	Actual/Forecast achievement date dd/mm/yyyy
1	MS1 Kick-Off Meeting [M1]		MU	31/10/2014	21/11/2014
2	MS2 System specification [M6]	1	MU	31/03/2015	28/02/2015
3	MS8 system specification	1	MU	31/05/2015	31/05/2015
4	MS12 first versions of the software modules	2/4	MU	30/09/2015	30/09/2015
6	MS14 evaluation strategy (testing of prototype)	1/3/6	MU	30/11/2015	30/11/2015

Note: This table is cumulative. It should show all milestones in the whole project period.

2.3.3 Deliverables achievement

No.	Deliverable title	WP no.	Lead partner (abbreviation)	Planned delivery date dd/mm/yyyy	Actual/Forecast delivery date dd/mm/yyyy
1	D1.1.2 Specification of corpora and the corpus building module	1	MU	31/03/2015	28/02/2015
2	D1.1.3 Specification of word-sketch grammars and tools	1	MU	31/03/2015	28/02/2015
3	D1.1.4 Specification of the semantic content matching and wordspace module	1	NTNU	31/05/2015	31/03/2016
4	D4.1: Methodology of Sketch Grammar evaluation	4	MU	30/09/2015	30/09/2015
5	D1.2.1 The HaBiT system v1: First integrated system prototype	1	MU	30/11/2015	30/11/2015
6	D6.1 Project evaluation plan	6	NTNU	30/11/2015	31/01/2016

Note: This table is cumulative. It should show all deliverables in the whole project period.

2.4 Work Packages (WPs) (max. 6 A4 pages)

2.4.1 Project work packages (WP)

WP number	Title	Planned date of start (dd/mm/yyyy)	Actual/Forecast date of end (dd/mm/yyyy)
WP1	System integration	06/2014	04/2017
WP2	Multi-billion word corpus building	08/2014	01/2016
WP3	Corpora for under-resourced languages	01/2015	11/2016
WP4	Shallow processing grammars and tools	10/2014	02/2017
WP5	Multi-sense and multi-lingual word spaces	06/2015	02/2017
WP6	Requirements and evaluation	06/2014	04/2017

2.4.1.1 WP number

WP1

2.4.1.2 WP title

System integration

2.4.1.3 WP leader

Karel Pala

2.4.1.4 WP start date

01/06/2014

2.4.1.5 WP end date

30/04/2017

2.4.1.6 WP objectives

To define the overall system specifications, to integrate and test all software modules produced, to create a demo website

2.4.1.7 WP task

Task T1.1: System and modules requirements specification was finished delivering D1.1.1 Task T1.2: Module integration and testing and Task T1.3: Coordination of the platform development were in progress during preparation of the first HaBiT system prototype.

2.4.1.8 WP deliverable

- D1.1.1 System specifications [M8]: Overall system design definitions

- D1.1.2 Specification of corpora and the corpus building module [M12]
- D1.1.3 Specification of word-sketch grammars and tools [M12]
- D1.1.4 Specification of the semantic content matching and wordspace module [M12]
- D1.2.1 The HaBiT system v1 [M14]: First integrated system prototype

2.4.1.9 WP milestone

MS8 System specification: Overall system specification completed
MS14 HaBiT system v1: First integrated system prototype, including first versions of the software modules, ready for performance and quality validation

2.4.1.10 WP Human resources

Qualification level: 9 - Karel Pala (docent), Aleš Horák (docent), Pavel Rychlý (docent), Miloš Jakubíček (RNDr., PhD student), Vojtěch Kovář (PhD, postdoc), Björn Gambäck (professor), Janne Bondi Johannessen (professor), Lars Bungum (researcher), Hans Moen (researcher)
Person-months: 3

2.4.2.1 WP number

WP2

2.4.2.2 WP title

Multi-billion word corpus building

2.4.2.3 WP leader

Pavel Rychlý

2.4.2.4 WP start date

01/08/2014

2.4.2.5 WP end date

31/01/2016

2.4.2.6 WP objective

To mine the internet to gather multi billion word corpora in Norwegian and Czech. A web crawler SpiderLing will be used. The corpora will be morphologically annotated and indexed for fast searching.

2.4.2.7 WP task

Task T2.1: We have built the language dependent models required by the next step, for both languages.
Task T2.2: The SpiderLing crawler was run to mine text from the internet. The crawler was improved to filter out web pages without fluent sentences (e.g., spam) based on Trust Rank (Gyöngyi et al. 2004).
Task T2.3: The gathered data were deduplicated using the Onion tool, tokenized, split to sentences and transformed in the XML format with separate structures for document, paragraph, sentence.
Task T2.4: We are currently morphologically annotating the processed corpus data to allow further linguistic analysis.
Task T2.5: After wide investigation we have chosen OpenSubtitles for building parallel Czech-Norwegian corpus. Data for both languages are being processed now.

2.4.2.8 WP deliverable

D2.1: The new version of the web crawler SpiderLing was built [M8]
D2.2: A Norwegian corpus, sized 2 billion tokens, cleaned (without boilerplate, de-duplicated), morphologically annotated, indexed for fast searching. [M13] In this version of the corpus we aimed at the very clean content. We are tuning several parameters to create bigger Norwegian corpus.

2.4.2.9 WP milestone

MS12 Web Corpus building pipeline: First version of the pipeline for creating text corpora from web. The first version of the Norwegian corpus build using this pipeline. [M12]

2.4.2.10 WP Human resources

Qualification level: 10 - Karel Pala (docent), Aleš Horák (docent), Pavel Rychlý (docent), Vít Suchomel (Mgr., PhD student), Vít Baisa (Mgr., PhD student), Miloš Jakubíček (RNDr., PhD student), Vojtěch Kovář (PhD, postdoc), Adam Rambousek (PhD, postdoc), Björn Gambäck (professor)
Person-months: 12

2.4.3.1 WP number

WP3

2.4.3.2 WP title

Corpora for under-resourced languages

2.4.3.3 WP leader

Pavel Rychlý

2.4.3.4 WP start date

01/01/2015

2.4.3.5 WP end date

30/11/2016

2.4.3.6 WP objective

To mine the internet to gather multi million word corpora in Amharic, Afaan Oromo, Tigrinya and Somali. A web crawler SpiderLing will be used. The corpora will be morphologically annotated and indexed for fast searching.

2.4.3.7 WP task

Task T3.1: We have collected texts, wordlists and tools for all target languages. Task T3.2: We have used SpiderLing and other tools to built the first version of the Amharic corpus. Corpora for other Ethiopian languages are in the process. Task T3.3: Custom scripts for harvesting specific web domains are in preparation. Task T3.5: The Amharic web corpus data have been morphologically annotated to allow further linguistic analysis.

2.4.3.8 WP deliverable

D3.1a + D3.1b The Amharic web corpus, sized 17.7 million tokens, morphologically annotated, is ready for searching [M14]

2.4.3.9 WP milestone

Milestone MS14: delivery of D3.1b in month 14. An annotated corpus for Amharic

2.4.3.10 WP Human resources

Qualification level: 11 - Karel Pala (docent), Aleš Horák (docent), Pavel Rychlý (docent), Vít Suchomel (Mgr., PhD student), Vít Baisa (Mgr., PhD student), Zuzana Nevěřilová (PhD, postdoc), Björn Gambäck (professor), Janne Bondi Johannessen (professor) Person-months: 13

2.4.4.1 WP number

WP4

2.4.4.2 WP title

Shallow processing grammars and tools

2.4.4.3 WP leader

Aleš Horák

2.4.4.4 WP start date

01/10/2014

2.4.4.5 WP end date

28/02/2017

2.4.4.6 WP objective

A Word Sketch table displays one page collocational behaviour of a word or phrase. Word sketches are automatically generated from a Word Sketch Grammar which has to be defined for each language. The objectives of this WP are to develop methods and tools for easy setup of the basic Sketch Grammar for a new language and tools and methods for enhancing, evaluating and debugging existing Sketch Grammars.

2.4.4.7 WP task

Task T4.1: We have defined a procedure for evaluating a sketch grammar. (D4.1)
Task T4.2: Work on tools for Sketch Grammar maintenance is in progress.

2.4.4.8 WP deliverable

D4.1: Methodology of Sketch Grammar evaluation. Month of delivery: [M12]

2.4.4.9 WP milestone

MS12 delivery of D4.1

2.4.4.10 WP Human resources

Qualification level: 11 - Karel Pala (docent), Aleš Horák (docent), Pavel Rychlý (docent), Vít Suchomel (Mgr., PhD student), Vít Baisa (Mgr., PhD student), Miloš Jakubíček (RNDr., PhD student), Vojtěch Kovář (PhD, postdoc), Adam Rambousek (PhD, postdoc), Zuzana Nevěřilová (PhD, postdoc), Björn Gambäck (professor)
Person-months: 11

2.4.5.1 WP number

WP5

2.4.5.2 WP title

Multi-sense and multi-lingual word spaces

2.4.5.3 WP leader

Björn Gambäck

2.4.5.4 WP start date

01/06/2015

2.4.5.5 WP end date

28/02/2017

2.4.5.6 WP objectives

- To investigate word-level semantic matching and disambiguation
- To increase usability through semantic search options
- To create parallel, multi-lingual word spaces
- To investigate multi-sense word space methods

2.4.5.7 WP task

Task T5.1: Semantic search and disambiguation
- Development of semantic search models
- Development of matching logic for mono- and multi-lingual word space models

2.4.5.8 WP deliverable

D5.1 Semantic search interface v1 [M12]: Search interface with rudimentary functionality

2.4.5.9 WP milestone

--

2.4.5.10 WP Human resources

Qualification level: 9 - Karel Pala (docent), Aleš Horák (docent), Björn Gambäck (professor), Lars Bungum (researcher), Hans Moen (researcher)
Person-months: 8

2.4.6.1 WP number

WP6

2.4.6.2 WP title

Requirements and evaluation

2.4.6.3 WP leader

Karel Pala

2.4.6.4 WP start date

01/06/2014

2.4.6.5 WP end date

30/04/2017

2.4.6.6 WP objectives

To ensure that the system meets user demands, the usability of the system; to ensure availability of support documentation

2.4.6.7 WP task

Task T6.1: Project evaluation strategy is summarised in the D6.1 Project evaluation plan
Task T6.2: Usability studies -- we have collected use cases to be included in the final testing of the HaBiT system

2.4.6.8 WP deliverable

D6.1 Project evaluation plan [M8]: Definition of evaluation criteria and methods

2.4.6.9 WP milestone

--

2.4.6.10 WP Human resources

Qualification level: 11 - Karel Pala (docent), Aleš Horák (docent), Pavel Rychlý (docent), Vojtěch Kovář (PhD, postdoc), Zuzana Nevěřilová (PhD, postdoc), Björn Gambäck (professor), Janne Bondi Johannessen (professor), Lars Bungum (researcher)
Person-months: 5

2.4.7 Project output(s)

Type of output	Title	Date of accomplishment (mm/yyyy)	Date of realization (mm/yyyy)
D	Annotation of Multi-Word Expressions in Czech Texts	10/2015	12/2015
D	DEBWrite: Free Customizable Web-based Dictionary Writing System	04/2015	08/2015
D	Increasing Coverage of Translation Memories with Linguistically Motivated Segment Combination Methods	06/2015	09/2015
D	Interactive Visualizations of Corpus Data in Sketch Engine	01/2015	05/2015
D	Longest-commonest Match	04/2015	08/2015
D	SemEval-2015 Task 15: A CPA dictionary-entry-building task	08/2014	06/2015
D	Towards Automatic Finding of Word Sense Changes in Time	10/2015	12/2015

Remark: *the outputs indicated in 2.4.7 represent a combination of the publications and software tools which are referred to in them.*

2.5 Project management during reporting period (max. 2 pages A4)

The project has been managed during the reporting period in a standard way. The following tasks have been solved:
- monitoring and auditing of project, i.e. monitoring of the technical content and progress is run by the Project Management Board, which consists of the team leaders of each party. The process has been going on in a standard way and without any problems.
- intellectual property rights - in this respect the Consortium follows the main aim - to make project outputs — resources and tools - to be openly accessible.
- risk management and quality assurance plan - risk assessment has been controlled by the project promoter, who ensures that this process will be continuously observed throughout the whole project duration. Quality of each project component and risk management has been monitored and evaluated in agreement with the Quality Assurance Plan (see 2.6.1)
- technology management plan has been followed to keep the individual project components on the desired quality level - this can be seen in the sections describing the particular project results.

We would like to add that the mentioned plans have been described separately in the project documentation (in TracWiki), so we do not feel the need to repeat what has already been said.

Personal changes that took place in the Czech and Norwegian teams are described and justified in the sect. 2.5.

The regular project meeting took place in September 5-6 2015 in Oslo at the University of Oslo campus at Blindern.
The participants:

- Masaryk University (MU) - Aleš Horák, Pavel Rychlý, Vít Suchomel
- NTNU Trondheim - Björn Gambäck, Lars Bungum
- University of Oslo (UiO) - Janne B. Johannessen, Lutz Eberhard Edzard (Saturday), Kristin Hagen, Anders Nøklestad, Joel Priestley
- Addis Ababa University (AAU), Ethiopia - Derib Ado, Feda Negesse (Binyam Ephrem, e-mail contact, he did not attend)

Janne Bondi Johannessen, Björn Gambäck and Lars Bungum also met with the colleagues from the universities in Oslo, Hawassa and Addis Ababa at Rondane, Norway on Sep 1-4.

Björn Gambäck attended a meeting with representatives of the Norwegian Research Council and the Czech Ministry of Education, Youth and Sports in Trondheim on October 23, 2015.

Project planning is going on in a standard way. The status did not undergo any change.
No deviations from the planned milestones have occurred, and all deliverables but one have been prepared on time, with the exception being Deliverable 1.1.4 which have been postponed awaiting the employment of a further postdoc researcher at NTNU.
No changes to the legal status of any of the Project Promoter/partners, in research organisations and SMEs took place.

Development of the project website has been going in the accordance with the plan (the website is referenced in sect 2.9.1.)

2.5.1 List of Project Promoter's staff working on project during reporting period (all members connected to personal costs) and changes at the team

Full name	Position in project	Full-/ part-time	Work load (1.0-0.0)	Hired on (dd/mm/yyyy)	Quit on (dd/mm/yyyy)
doc.PhDr. Karel Pala, CSc.	project team leader	part-time	0.20	01/10/2014	--
doc.RNDr. Aleš Horák, Ph.D.	researcher	part-time	0.20	01/10/2014	--
Mgr. Pavel Rychlý, Ph.D.	researcher	part-time	0.20	01/10/2014	--
Mgr. Vít Suchomel	researcher, PhD student	part-time	0.40	01/10/2014	--
Mgr. Vít Baisa	researcher, PhD student	part-time	0.25	01/10/2014	--
Mgr. Miloš Jakubíček	researcher, PhD student	part-time	0.25	01/10/2014	--
RNDr. Vojtěch Kovář, Ph.D.	researcher	part-time	0.25	01/10/2014	--
Mgr. Marek Medveď	project administrator	part-time	0.30	01/08/2015	31/12/2015
Staff changes in the reporting period					
Mgr. Lucia Kocincová	project administrator	part-time	0.30	01/10/2014	01/08/2015
RNDr. Adam Rambousek, Ph.D.	researcher	part-time	0.20	01/01/2015	--
RNDr. Zuzana Nevěřilová, Ph.D.	researcher	part-time	0.10	01/01/2015	--

2.5.2 List of Project partner's staff working on project during reporting period (all members connected to personal costs) and changes at the team

Full name	Position in project	Full-/ part-time	Work load (1.0-0.0)	Hired on (dd/mm/yyyy)	Quit on (dd/mm/yyyy)
Prof. Björn Gambäck, PhD	partner team leader	part-time	0.20	01/11/2014	--
Prof. Janne Bondi Johannessen, PhD	researcher	part-time	0.10	01/11/2014	--
Staff changes in the reporting period					
Lars Bungum	researcher	full-time	1.0	01/06/2015	30/04/2017
Hans Moen	temporary researcher	part-time	0.6	17/04/2015	30/06/2015

2.5.3 Personnel changes justification (max. 1 page A4)

Personnel changes did not affect the project budget. Hiring two extra researchers (A.Rambousek and Z.Nevěřilová) in the team of MU was needed to cope with the shortened project length caused by delayed start of the project and was necessary because of the high demand on work. A.Rambousek works on analysis and implementation of efficient processing and presentation of the lexicon of given languages. Z.Nevěřilová is devoted to the analysis of the collocational properties of given languages. Another administrative changes in the MU team was a 4 months substitution of L.Kocincová, the project administrator, by M. Medveď. This change was caused by the fact that L.Kocincová was on a 4-months research stay abroad (in Bergen, Norway). She returned to the Czech Republic in the end of 2015 and is working back in the team since January 2016.

According to the original plan, two postdoctoral researchers should have been employed at NTNU in the period, and Lars Bungum was employed as one of those. However, due to problems with filling the second position on a longer term, Hans Moen was temporarily engaged in HaBiT during the summer of 2015. A postdoc from India is going to join the NTNU group to work on the semantic content matching and word sense disambiguation from the Spring of 2016 and until the end of the project.

2.6 Monitoring and auditing of project (max. 2 pages A4)

Monitoring of the technical content and progress is run by the Project Management Board, which consists of the team leaders of each party. The Board is responsible for handling most of the scientific issues, planning of the project, and changes of the work plan. In case of any important issues or problems, the Project Management Board has the responsibility to resolve them. Moreover, the Board reviews all publications and project deliverables. Administrative issues, including day-to-day operations, and regular project management are handled by the Project Promoter.
The project has not been audited from any institutions. Each of project partners proposes an independent expert, who will evaluate the project report and its essentials. On the MU side the expert is Prof. Václav Matoušek from the West Bohemia University in Pilsen. On the NTNU side the expert is Prof. Lars Borin from the University of Gothenburg, Sweden.

2.6.1 Risk management and quality assurance (max. 2 page A4)

Quality of each project component and risk management is monitored and evaluated according to a Quality Assurance Plan. This document was prepared by both project partners and approved by the Project Management Board.
Risk assessment is controlled by the project promoter, who ensures that this process will be continuous throughout the whole project duration. The project promoter also discusses relevant risk issues with the Board at regular meetings to secure that each potential risk is taken seriously. The Risk Management Plan assesses any potential probabilities of risks and also defines measures that are needed to be taken in order to minimize any impact of the risks. This plan is being reviewed and improved regularly to ensure that any new risks are taken into account.

2.6.2 Irregularities (max. 1 page A4)

No irregularities have occurred.

2.7 Intellectual property rights management (max. 1 page A4)

The main aim of the consortium is to make any project output — resources and tools —open access. Access rights are given to all the project partners including the Ethiopian collaborators. The Project Management Board controls the distribution of intellectual property rights in accordance with the Partnership Agreement. All decisions regarding intellectual property rights are summed up in the Technology Management Plan, which also contains plans for the exploitation of results and licensing strategies.

2.8 Scientific publications and dissemination of project in reporting period (max. 2 pages A4)

The progress of the project has been published and presented at scientific conferences related to the topics of the HaBiT project. The list of conferences includes:

- Electronic lexicography in the 21st century: linking lexical data in the digital age, eLex 2015, Herstmonceux Castle in Sussex, United Kingdom.
- the Workshop on Innovative Corpus Query and Visualization Tools at NODALIDA 2015, Vilnius, Lithuania.
- the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, Colorado, USA.
- the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA) at EMNLP 2015, Lisbon, Portugal.
- the 10th Conference on Recent Advances in Natural Language Processing (RANLP), Hissar, Bulgaria.
- the Workshop on Natural Language Processing for Translation Memories (NLP4TM), Hissar, Bulgaria.
- the 9th International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT), Larnaca, Cyprus.
- the 7th meeting of Forum for Information Retrieval Evaluation (FIRE), Gandhinagar, Gujarat, India.
- the 12th International Conference on Natural Language Processing (ICON), Trivandrum, Kerala, India.
- the 9th Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2015), Karlova Studánka, Czech Republic. This workshop was partly organized as a HaBiT information event, see the information on the project web pages for the details.

More information about published papers directly related to the HaBiT project can be found below.

2.8.1 Reporting on scientific publications (length according your need)

Towards Automatic Finding of Word Sense Changes in Time - *Vít Baisa, Ondřej Herman, Miloš Jakubíček*
Ninth Workshop on Recent Advances in Slavonic Natural Language Processing. Karlova Studánka, Czech Republic, 2015.
We present a methodology proposal for finding changes in contextual behaviour of words. Assuming that a word sense is defined by actual usages of the word in a given context, this task corresponds to finding changes of word senses. We outline main ideas of our distributional approach based on word sketches and discuss preliminary results.

Annotation of Multi-word Expressions in Czech Texts - *Zuzana Nevěřilová*
Ninth Workshop on Recent Advances in Slavonic Natural Language Processing. Karlova Studánka, Czech Republic, 2015.
Multi-word expressions (MWEs) are difficult to define and also difficult to annotate. Many cases of incorrect annotation in Czech corpora are known. Our method is based on corpus data observation that indicates that people are unsure when writing a MWE whether it is

one word, a word with dashes, or several words. The result is a list of MWE candidates and also an application that classifies the input as MWE, probable MWE, or non-MWE.

DEBWrite: Free Customizable Web-based Dictionary Writing System - *Adam Rambousek and Aleš Horák*
Electronic lexicography in the 21st century: linking lexical data in the digital age, eLex 2015, United Kingdom
Today, lexicographers can avail themselves of several commercial and freely distributed dictionary writing systems (DWS). Nevertheless, there is still a group of users whose requirements are not satisfied by existing DWSs. In various lexicographic forums, there is a growing demand for freely available DWS that allows customization of the dictionary microstructure. In accordance with such requests, a new project was developed as part of the DEB (Dictionary Editor and Browser) platform. DEBWrite is implemented as a multi-platform web application based on open standards. It allows users to create and share a new dictionary without any difficult configuration or advanced technical skills. According to a defined entry structure, the editing form and the public dictionary browser are generated automatically. The dictionary may be published in an online form, or in formats suitable for print preparation.

Increasing Coverage of Translation Memories with Linguistically Motivated Segment Combination Methods - *Marek Medveď, Vít Baisa, Aleš Horák*
Proceedings of The Workshop on Natural Language Processing for Translation Memories (NLP4TM). Bulgaria, 2015.
Translation memories (TMs) used in computer-aided translation (CAT) systems are the highest-quality source of parallel texts since they consist of segment translation pairs approved by professional human translators. In this paper, we describe several methods for expanding translation memories using linguistically motivated segment combining approaches concentrated on preserving the high translational quality. The evaluation of the methods was done on a medium-size real-world translation memory and documents provided by a Czech translation company as well as on DGT TM published by EC.

Interactive Visualizations of Corpus Data in Sketch Engine - *Lucia Kocincová, Miloš Jakubíček, Vojtěch Kovář, Vít Baisa*
Proceedings of the Workshop on Innovative Corpus Query and Visualization Tools at NODALIDA 2015. Vilnius, Lithuania, 2015.
Automatic analysis of the large text corpora produces large amounts of figures as a result of various functions. These provide empirical evidence for a research hypothesis or serve in numerous practical applications of natural language processing. Usually, the results are presented in the form of tables containing raw data to be interpreted by domain experts. This paper describes an ongoing work on new visualizations and user interface enhancements in Sketch Engine corpus management system which aim at easing the interpretation of the data.

Longest-commonest Match - *Adam Kilgarriff, Vít Baisa, Miloš Jakubíček, Pavel Rychlý*
Electronic lexicography in the 21st century: linking lexical data in the digital age, eLex 2015, United Kingdom

Finding two-word collocations is a well-studied task within natural language processing. The result of this task for a given headword is usually a list of collocations sorted by a salience score. In corpus manager Sketch Engine, these pairs are extracted from data using a word sketch grammar relation rules and log-dice statistics resulting in a sorted list of triples . The longest–commonest match is a straightforward extension of these two-word collocations into multi-word expressions. We present an algorithm behind the longest–commonest match together with a simple evaluation.

SemEval-2015 Task 15: A CPA dictionary-entry-building task - *Vít Baisa, Jane Bradbury, Silvie Cinková, Ismail El Maarouf*
Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Denver, Colorado, USA, 2015.
This paper describes the first SemEval task to explore the use of Natural Language Processing systems for building dictionary entries, in the framework of Corpus Pattern Analysis. CPA is a corpus-driven technique which provides tools and resources to identify and represent unambiguously the main semantic patterns in which words are used. Dictionary entry building is split into three subtasks which all start from the same concordance sample. The task has produced a major semantic multidataset resource which includes data for 121 verbs and about 17,000 annotated sentences, and which is freely accessible.

Negation Scope Detection for Twitter Sentiment Analysis - *Johan Reitan, Jørgen Faret, Björn Gambäck, Lars Bungum.* The 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA) at the 2015 Conference on Empirical Methods on Natural Language Processing (EMNLP), Lisbon, Portugal. September 2015, pp. 99–108, Association for Computatioal Linguistics.
Inducing the scope of negation is a major problem when trying to analyse the semantic structure of social media text. The paper describes the first sophisticated negation scope detection system for Twitter sentiment analysis. The system has been evaluated both on existing corpora from other domains and on a corpus of English Twitter data (tweets) annotated for negation. It produces better results than what has been reported in other domains and improves the performance on tweets containing negation when incorporated into a state-of-the-art Twitter sentiment analyser.

Part-of-Speech Tagging for Code-Mixed English-Hindi Twitter and Facebook Chat Messages - *Anupam Jamatia, Björn Gambäck, Amitava Das*. The 10th Conference on Recent Advances in Natural Language Processing (RANLP), Hissar, Bulgaria. September 2015, pp. 239–248.
Annotating social media corpora poses several original problems, including that bilingual users mix (written) languages very much in the same way as in their speech. The paper reports joint work with researchers at National Institute of Technology, Agartala, Tripura, India and at Indian Institute of Information Technology, Sri City, Andhra Pradesh on collecting and annotating code-mixed English-Hindi social media text (Twitter and Facebook messages), and experiments on automatic tagging of these corpora, using both a coarse-grained and a fine-grained part-of-speech tag set. We compare the performance of a combination of language specific taggers to that of applying four machine learning algorithms to the task (Conditional Random Fields, Sequential Minimal Optimization,

Naïve Bayes and Random Forests), using a range of different features based on word context and word-internal information.

Multi-Domain Adapted Machine Translation Using Unsupervised Text Clustering - *Lars Bungum, Björn Gambäck.* Modeling and Using Context: 9th International and Interdisciplinary Conference, CONTEXT 2015, Lanarca, Cyprus, November 2-6, 2015. Proceedings. Editors: Henning Christiansen, Isidora Stojanovic, George A. Papadopoulos. Springer Verlag, Lecture Notes in Computer Science Volume 9405, pp. 201-213.
Domain Adaptation in Machine Translation means to take a machine translation system that is restricted to work in a specic context and to enable the system to translate text from a different domain. The paper presents a two-step domain adaptation strategy, by first making use of unlabeled training material through an unsupervised algorithm, the Self-Organizing Map, to create auxiliary language models, and then to include these models dynamically in a machine translation pipeline.

Self-Organizing Maps for Classification of a Multi-Labeled Corpus - *Lars Bungum, Björn Gambäck.* The 12th International Conference on Natural Language Processing (ICON), Trivandrum, Kerala, India.
A Self-Organizing Map was used to classify the Reuters Corpus, by assigning a label to each of the documents that cluster to a specific node in the Self-Organizing Map. The predicted label is based on the most frequent label among the training documents attributed to that particular node. Experiments were carried out on different grid sizes (node numbers) to determine their influence on classification results. Informative visualizations of the resulting Self-Organizing Maps are demonstrated. We argue that the Self-Organizing Map is well suited to classify a document collection in which many documents simultaneously belong to several categories.

Sentence Boundary Detection for Social Media Text - *Dwijen Rudrapal Anupam Jamatia Kunal Chakma, Amitava Das, Björn Gambäck*. The 12th International Conference on Natural Language Processing (ICON), Trivandrum, Kerala, India.
The paper presents a joint study on automatic sentence boundary detection in social media texts with researchers at National Institute of Technology, Agartala, Tripura, India and at Indian Institute of Information Technology, Sri City, Andhra Pradesh. We explore the limitations of using existing rule-based sentence boundary detection systems on social media text, and as an alternative investigate applying three machine learning algorithms (Conditional Random Fields, Naïve Bayes, and Sequential Minimal Optimization) to the task. The systems were tested on three corpora annotated with sentence boundaries, one containing more formal English text, one consisting of tweets and Facebook posts in English, and one with tweets in code-mixed English-Hindi. The results show that Naïve Bayes and Sequential Minimal Optimization were clearly more successful than the other approaches.

2.9 Project promotion and information activities about project (max. 2 pages A4)

A logo for the HaBiT project was created in two versions, so that it can be used for all communication purposes as the main object of recognition when promoting the project.
The colours and the font of the logo were chosen according to Norway grants logo, so the connection between those two is apparent.

Logo HaBiT - full name variant

The project website was updated not only with up-to-date information and last activities but the web presentation's appearance was also changed to be more user-friendly and eye-catching.

Logo HaBiT - acronym variant

Publicity events and activities were organised according to Publicity Plan (Annex VI). In the reporting period, two information activities were held:

1. Public information activity which was especially designed for students was held on Oct 22nd 2015 at Faculty of Informatics Masaryk University (FI MU), Czech Republic. The event was launched with information about the work of Natural Language Centre at FI MU. Then, the research within HaBiT project was thoroughly presented along with Norway grants. Different possibilities to engage students in the research activities were introduced. At the end, the discussion with students and the project researchers took place. Overall feedback and reactions were very positive.
   (selected photos from the event can be found in the Annex V along with informative text; more photos can be found at the project website at http://habit-project.eu/wiki/InformationEvents)

2. The Ninth Workshop on Recent Advances in Slavonic Natural Language Processing was held between 4tn and 6th December 2015 in Karlova Studánka, Czech Republic. The event is focused on exchanging of information between research teams working in the field of computer processing of Slavonic languages and also in related areas. Two HaBiT researchers -- Miloš Jakubíček and Vít Suchomel -- were actively participating in presenting the current outputs and progress of the project. Moreover, the event helped to promote the funding programme with roll-up and leaflets given to all participants.
   (selected photos from the event can be found in the Annex V along with informative text; more photos can be found at the project website at http://habit-project.eu/wiki/InformationEvents)

3. Björn Gambäck was invited to give a talk about language processing for under-resourced languages at the 7th meeting of Forum for Information Retrieval Evaluation (FIRE), Gandhinagar, Gujarat, India, on December 4-6, 2015 (see http://fire.irsi.res.in/fire/home. The invited talk abstracts are availiable at http://fire.irsi.res.in/fire/abstracts).

Next year the project promoter MU will organize the 19th International Conference on Text, Speech and Dialogue (www.tsdconference.org) which is a large scientific meeting attended by participants from all over the world. MU will also organize a specific pre-conference workshop named Community-based Building of Language Resources, CBBLR. The main topic of the workshop is directed at building new language resources, especially for languages with no or too small existing language resources. The workshop will be organized in cooperation with the HaBiT project Consortium, with submissions open to other resource development projects. The workshop submissions will undergo two separate review processes - the best papers which will succeed in both review processes (by the TSD 2016 Conference PC and CBBLR Workshop 2016 PC) will be published in the TSD 2016 Springer Proceedings, all other accepted CBBLR workshop papers will be published in a separate proceedings with ISBN. The CBBLR workshop will take place on September 12 2016 in the TSD conference venue.

2.9.1 Project website

www.habit-project.eu

2.10 Achievement of Programme outcome(s) and outputs (max. 1/2 page A4)

The HaBiT project contributes to research-based knowledge development by extending the cooperation within and outside of the Czech-Norwegian Research Programme. A number of PhD students are already part of the research and the project partners plan to include more female researchers, PhD students, and postdocs in order to actively engage them in best practices of both workplaces, and to transfer and share state-of-the-art development. This close cooperation between the partners becomes a basis for future successful cooperation, thus ensuring sustainability of the research or other future relevant activities.
The project results also contribute to the cultural development of less-developed countries by acquiring and sharing developed information technologies.

All project outputs are being presented to the international research community and are also disseminated via the project website in order to present the research to all possible types of audiences.

2.10.1 Programme output indicators in project (numbers per project, e.g. 7)

Number of PhD students	**5**
Number of postdocs	**4**
Number of female researchers (including after maternity leave)	**2**
Number of internationally refereed (joint) scientific publications	**12**

Note: Numbers of the target groups are for the whole project. These indicators are used for reporting about the Programme to the Financial Mechanism Office in Brussels.

3. FINANCIAL PART

3.1 Explanation of use of grant (max. 3 pages A4)

Expenses were according to project budget, as stated in detail in Annex I:

- Masaryk University:
  - personnel costs:1,355,869 CZK of 2015 budget (693,341 CZK of 2014 budget)
    The personnel costs cover salaries of 9 part-time researchers and 1 part-time project administrator, as stated in 2.5.1:
    - Karel Pala, 20%, 16,000 CZK/month
    - Aleš Horák, 20%, 16,000 CZK/month
    - Pavel Rychlý, 20%, 12,000 CZK/month
    - Vít Suchomel, 40%, 18,000 CZK/month
    - Vít Baisa, 25%, 11,250 CZK/month
    - Miloš Jakubíček, 25%, 13,250 CZK/month
    - Vojtěch Kovář, 25%, 13,250 CZK/month
    - Adam Rambousek, 20%, 10,600 CZK/month
    - Zuzana Nevěřilová, 10% 5,300 CZK/month
    - Lucia Kocincová, 30%, 11,400 CZK/month. L.Kocincová was for 4 months in an abroad stay in Norway, during this time her work was done by Marek Medveď.
  - travel costs: 8,976 CZK of 2015 budget (65,452 CZK of 2014 budget).
    The travel costs include: annual meeting of all partners was held in Oslo (Rychlý, Horák, Suchomel),
    NoDaLiDa conference workshop active participation (Kocincová),
    RASLAN workshop active participation (Jakubíček, Suchomel) and inland travelling to the programme operator seminars in Prague.
  - cost of consumables: 19,893 CZK of 2015 budget (16,500 CZk of 2014 budget).
    The consumables expenses were used for extra server specific disk storage used for the project data.
  - indirect costs: 830,843 CZK of 2015 budget (465,176 CZK of 2014 budget) computed as 60% flat rate

- NTNU (costs 2015):
  - personnel costs: 1,201,457 NOK
    - Hans Moen, 60% for 2.5 months
    - Lars Bungum, 100% for 6 months
    - prepayment to ERCIM for postdoc employment of Utpal Sikdar 2016
  - travel costs: 41,111 NOK
    - annual meeting in Oslo (Gambäck, Bungum)
    - meeting with Ethiopian partners in Rondane (Bungum)
    - conference travel (CONTEXT, Bungum)
    - more late 2015 travels will be added 2016
  - indirect costs: 745,540 NOK
    - calculated using a 60% flat rate

3.2 Budget changes justification for reporting period (max. 2 pages A4)

MU has made the following operative changes in the part of the budget planned for 2014 and spent in 2015:

- 20,000 CZK planned for part-time agreements were transferred to salaries, since the annotation work planned for external cooperators was done by project team members. The amount of the total personnel costs remains unchanged.
- 1,000 CZK planned for travel costs were transferred to consumables, which covered a purchase of server disk storage for the project data.
- 500 CZK planned in the approved budget for indirect costs had to be transferred to consumables, too, since the whole amount approved for indirect costs, i.e. 710,000 CZK, would overrun the 60% limit.

These changes do not influence the MU total project budget, neither the total personnel expenses.

NTNU has prepaid the 2016 salary for Utpal Sikdar to ERCIM. This does also not affect the total project budget nor the total personnel expenses.

3.3 VAT reclaim YES or NO

Full Name of Project promoter/partner	VAT reclaim (Yes/No)
Masaryk University	**NO**
Norwegian University of Science and Technology	**NO**

3.4 Indirect costs model – overheads (approved in the Project contract)

Name of Project Promoter/Project partner (abbreviation)	Participant Identification Code (PIC)	Overheads rate in % (per partner)	Using analytical accounting system (Yes/No)
Masaryk University	999880657	60	YES
Norwegian University of Science and Technology	999977851	60	YES

3.5 Procurement and small scale contracting (max. 2 pages A4)

No procurement or small scale contracting are planned for the project, so none were launched in the reporting period.

3.6 Fund for bilateral relations (max. 1 page A4)

No bilateral funds received.

4. MANDATORY AND VOLUNTARY ANNEXES

4.1 Overview of annexes required to project periodic report

No.	Annex	Mandatory online submission format
I.	***Annex I – Project Interim Financial Report*** Expenditure actually incurred in the reporting period in CZK. It is signed by statutory of the Project Promoter, or an attorney, and by principal investigator. Use the template.	xls(x) and pdf
II.	***Annex II - Report on Actual Incurred Expenditure*** It relates to all Czech and Norwegian entities. It is a record from an accounting system in a local currency. Actual expenditure per entity corresponds to annex I and/or III. It contents a stamp of the organisation and a name and a signature of a responsible person. For Norwegian partners it may be a copy. No template.	pdf
III.	***Annex III - Financial Statement by Norwegian Partner*** For Norwegian partners in NOK only. It contents a stamp of the organisation and a name and a signature of a responsible person. It may be a copy. Use the template.	pdf
IV.	***Annex IV - Confidentiality Declaration by Evaluator*** Related to each annex V. Signed by evaluator. A copy may be submitted. Use the template.	pdf
V.	***Annex V - Evaluation Report of the Project*** (at least 2 evaluators). Fill in English. Signed by evaluator. A copy may be submitted. Use the template.	pdf
VI.	***Annex VI - Project Publicity Plan*** Fill in English. Use the template.	pdf
VII.	***Annex VII - Letter of Attorney***, if applicable for Project promoters. Acceptable in Czech. A copy may be submitted. No template.	pdf
VIII.	***Voluntary annexes*** – e.g. photo documentation. No template.	CD/jpg

Note: For e-submission the required format for the project periodic report is doc(x) (you may also submit the undersigned pages in pdf as a separate file if no e-signature). Please, always indicate a revised document.
Please tie up the documents in this order: periodic report and annexes. Use Calibri font, size 12.

5. OTHER ANNEXES

INFORMATION EVENT

The RASLAN Workshop is an event dedicated to exchange of information between research teams working on projects of computer processing of Slavonic languages and related areas. RASLAN is focused on theoretical as well as technical aspects of the project work, presentations of verified methods are welcomed together with descriptions of development trends. The workshop also serves as a place for discussion about new ideas.
The ninth workshop was held on Dec 4-6 2015 in Karlova Studánka, Czech Republic where HaBiT researches also actively participated.

INFORMATION EVENT

The evening of Oct 22 2015 at Faculty of Informatics Masaryk University (FI MU), Czech Republic was dedicated to an information event for students where information about the work of Natural Language Centre at FI MU was displayed. The research within HaBiT project was thoroughly presented along with Norway grants. Possibilities to engage students in the research activities were introduced and the overall feedback was very positive.

PROMOTIONAL LEAFLET

The promotional leaflet was made for information purposes and was distributed at information events. It states basic facts about the project and the project consortium along with the project aims and information about funding.

HaBiT

HARVESTING BIG TEXT DATA FOR UNDER-RESOURCED LANGUAGES

PROJECT SUMMARY

MAIN AIMS AND OUTPUTS

The main objectives of the HaBiT project are to gather large-scale text data (corpora) from the Web for under-resourced languages, involving Norwegian, partly Czech and the major languages of Ethiopia – Amharic, Afaan Oromo, Tigrinya, Somali – and to build shallow processing applications.

The gathered data will be processed to make it usable in many language applications, such as information extraction or machine translation.

Furthermore, in the process of collecting corpora data, existing tools for building web text resources will be further developed and improved since the Ethiopian languages are quite different from most European languages.

Applications for the given languages will be built to allow for the separation and disambiguation of multiple senses of words.

One of the project's aims is to build a multi-billion word Norwegian corpus using the tools co-developed by the team of NLP Centre at FI MU and utilized in a previous joint EU-funded project with NTNU ("PRESEMT: Pattern REcognition-based Statistically Enhanced MT", 2010-2012).

Second, NTNU collaborates with University of Oslo and two Ethiopian universities in a project to support linguistic resource building in Ethiopia funded by Norad ("Linguistic Capacity Building – tools for the inclusive development of Ethiopia", NORHED 2013-2018).

It is natural to link the activities and to include processing of the four major languages in Ethiopia in the project: the HaBiT project will be able to feed into and leverage on the NORHED project, thoroughly testing the technologies and thus addressing the call topics on technology assessment, verification and testing, as well as on ICT meeting societal challenges, hence obtaining added value also in the political respect through cooperation with a less-developed country.

http://habit-project.eu

info@habit-project.eu

Harvesting big text data for under-resourced languages

PROJECT CONSORTIUM

PROJECT PROMOTER

The Natural Language Processing Centre is part of Faculty of Informatics, Masaryk University. The Centre focuses on obtaining practical results in the field of information technologies and linguistics.

Its main activities comprise above all: morphological and syntactic analysis, grammar development, corpus linguistics, semantic nets and ontology acquisition, semantic web and visual lexicons, production of various dictionaries and software tools for editing them, development of lexicographer's workbench and machine translation.

The key expertise of the NLP Centre related to the project is building and processing of huge corpora. The Centre has tools for fast querying, annotation and parsing of multi billion corpora.

PROJECT PARTNER

Norwegian University of Science and Technology is Norway's premier academic institution for technology and the natural sciences, with equally strong programmes in the social sciences, the arts and humanities.

NTNU's research has an international focus with an interdisciplinary approach. Due to their central position in the European Research Consortium for Informatics and Mathematics (ERCIM) network, staff at NTNU are given easy access to the European research community and the European ICT industry.

The language processing team at IDI collaborated with Masaryk University in the FP7 project PRESEMT ("Pattern REcognition-based Statistically Enhanced MT") where NTNU designed and implemented the corpus-processing modul.

The research leading to the results has received funding from the Norwegian Financial Mechanism 2009-2014 and the Ministry of Education, Youth and Sports under Project Contract no. MSMT-28477/2014

PROMOTIONAL PRESENTATION

The promotional presentation was made for information purposes and is to be used on information events. It states basic facts about the project and the project consortium along with the project aims and information about funding.