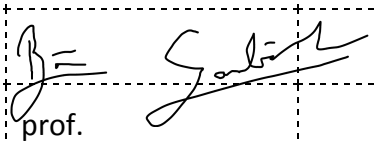
PROPOSAL FOR ADDITIONAL RESEARCH ACTIVITIES WITHIN SOCIAL SCIENCES AND HUMANITIES AREA	
Czech-Norwegian Research Programme (CZ09)	
Norwegian Financial Mechanism 2009-2014	
Programme area	**Bilateral Research Cooperation**
Project ID number	7F14047
Project title in English	Efficient Allocation of Human Resources for Linguistic Annotation of Texts
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
Name of Principal Investigator (PI)	Karel Pala
Signature of PI	
Statement	*I hereby declare that the information I state in the project proposal is accurate, true and complete. I am aware that if the information has been reversed in the opposite, I will face disqualification of the project proposal from the selection process.*
Done in	Brno, Czech Republic
Date	10/03/2016

On behalf of Project Promoter				
Stamp of Project Promoter	MASARYKOVA UNIVERZITA Žerotínovo nám. 9 601 77 BRNO [19]			
Statutory authority of Project Promoter	**Name(s):**	Mikuláš Bek		
	Signature(s):			
	Position:	rector		
On behalf of Project Partner				
Stamp of Project Partner				
Statutory authority of Project Partner	**Name(s):**			
	Signature(s):			
	Position:			

On behalf of Project Promoter				
Stamp of Project Promoter				
Statutory authority of Project Promoter	**Name(s):**	Mikuláš Bek		
	Signature(s):			
	Position:	rector		
On behalf of Project Partner				
Stamp of Project Partner				
Team leader of Project Partner	**Name(s):**	Björn Gambäck		
	Signature(s):			
	Position:	prof.		

1. GENERAL INFORMATION ABOUT PROJECT

1.1 Project ID

7F14047

1.2 Project acronym

AHuRAT (extension of HaBiT project)

1.3 Project title in English

Efficient Allocation of Human Resources for Linguistic Annotation of Texts.

1.4 Project title in Czech

Efektivní alokace lidských zdrojů pro jazykovou anotaci textů

1.5 Activity of research and development in project

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

Note: Tick one or more options to identify the relevant activity or combination of activities. For combination mark the percentage of each type of activity. If you tick only one option, it is meant that your project is 100 % of the indicated R and D activity.

1.5.1 Project Promoter: XXX

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

1.5.2 Project Partner: XXX

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

1.6 Programme thematic area

Social Sciences and Humanities ☒

1.7 Additional activities starting date (dd/mm/yyyy)

01/09/2016

Note: This date must not precede 1 June, 2016.

1.8 Additional activities ending date (dd/mm/yyyy)

30/04/2017

Note: Deadline for all additional activities is 30 April, 2017.

1.9 Additional activities duration in months (number, e.g. 10)

8

1.10 Total costs of additional activities (in CZK)

1 125 280 Kč

1.10.1 Grant for additional activities requested from Programme Operator (in CZK)

1 125 280 Kč

1.11 Number of partners

2

Note: Number of all partners including Project Promoter.

1.12 Abstract in English (max. ½ A4)

The aim of the project is to verify a methodology of PoS tagging based on automaticc extraction of texts from public sources (Wikipedia) for annotation by users, tagset refinement based on annotator agreement. Further, a simple user interface for annotating small parts of the data by novice and/or anonymous users, exporting the tagger language model and a new tagger or adaptation of an existing one to support lemmatization and tagging of unseen word forms will be tested . The methodology will be verified on PoS tagging of corpora of under-resourced languages spoken in Ethiopia. The output -- tagged corpora -- will be created on special workshop and will supplement outputs of HaBiT project. These corpora will be used for research and further development of under-resourced languages. The output of this project is also handing over the know-how thus socially support the community and creating platform for further co-operation.

1.12.1 Key words in English (max. 15 y words)

corpora, taggers, annotation of textual data, Amharic, Afaan Oromo, Tigrinya, Somali, Natural Language Processing

1.13 Abstract in Czech (max. ½ A4)

Cílem projektu je ověřit metodologii značkování slovních druhů založenou na automatickém získávání textů z veřejných zdrojů (Wikipedie), anotací od uživatelů, na jejichž shodě budou zároveň založeny značky. Dále bude testováno jednoduché uživatelské rozhraní k anotování malých částí dat i nezkušenými či anonymními uživateli, k exportu jazykového modelu a vytvoření nového taggeru či adaptace existujícího k podpoře lemmatizace a značkování neznámých slovních tvarů. Tato metodologie bude ověřena na označkování a značkách samých v korpusu jazyků s nedostatečnými zdroji, jimiž se mluví v Etiopii. Výstup ve formě označkovaných korpusů vzniklých na workshopu bude doplňovat výstupy projektu HaBiT. Tyto korpusy budou dále využitelné pro výzkum a další vývoj jazyků s nedostatečnými zdroji. Současně bude výstupem sociální podpoření komunity předáním know-how a vytvořením platformy k další spolupráci.

1.13.1 Key words in Czech (max. 15 key words)

korpusy, morfologické analyzátory, anotace textových dat, zpracování přirozeného jazyka

1.14 Ethical issues (max ¼ page A4)

NO

1.15 VAT reclaim YES or NO

NO

2. PROJECT INTRODUCTION

2.1 Introduction to supported project (max. ½ page A4)

Part of speech (PoS) tagging is a long-time and widely studied problem in the field of natural language processing. For many languages, PoS tagging is even considered as a solved problem since there are many taggers with relatively high accuracy (97% for English). On the other hand, most of the existing PoS taggers face problems in industrial-grade usage. Moreover, the task of building a tagger for a new language or even training an existing tagger for a new language or new language variant is still difficult, especially for languages with rich morphology.

The aim of the project is to test methodology of PoS tagging based on automatic extraction of texts from public sources (Wikipedia) for annotation by users, tagset refinement based on annotator agreement. Moreover, a simple user interface for annotating small parts of the data by novice and/or anonymous users, exporting the tagger language model and a new tagger or adaptation of an existing one to support lemmatization and tagging of unseen word forms will be evaluated as well.

The methodology will be verified on PoS tagging ang tags of corpora of the under-resourced languages spoken in Ethiopia (Amharic, Afaan Oromo, Tigrinya, Somali). The output -- tagged corpora -- will be created on special workshop and will supplement outputs of HaBiT project. These corpora can be used for research and further development in the area of the under-resourced languages. The output of this project will also be handing over the know-how, thus socially support the community and building a platform for further co-operation.

2.2 Brief Project Promoter introduction (max. ½ page A4)

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. Results of the projects are frequently published at various conferences, the NLP Centre also cooperates with similarly oriented institutes and companies in the Czech Republic and abroad. 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.

2.3 Brief Project Partner(s) introduction (max. ¼ page A4 each)

NTNU 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. The Department of Computer

and Information Science (IDI) has some 180 employees including about 40 permanent research staff. 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.

2.4 Management of the project (max. ½ page A4)

The key project management tasks are the monitoring of the technical content and progresss of each work package, coordination of the different project activities, and performing quality control to ensure appropriate project standards. The project is run by a Project Management Board consisting of the team leaders at the two partner sites, where administrative issues and day-to-day operations are mainly handled by the Project Promoter. The Project Management Board handles most scientific issues and middle- and long-term planning. It is responsible for approving the project workplan, reviewing all project deliverables, publications and demonstrations. It is also the Board's responsibility to discuss problems and changes to the plan, and to help the Project Promoter in resolving problems and conflicts, while the Project Promoter is responsible for the day-to-day management. For the financial management of the project the MU Principal Investigator will be assisted by MU's designated EU Project Administrative and Financial Support Team.

Each work package has a Work Package Leader who is responsible for managing and coordinating activities among the partners within the WP. The Work Package Leader establishes, in cooperation with the participating partners, the detailed schedule of the WP and organises the production and internal review of the WP deliverables. The Work Package Leader will also organise small WP-internal meetings if necessary, and keep the Coordinator informed of the work in progress, and of delays or changes to the original plan. All Work Package Leaders have the right to attend Board meetings, and to participate in the deliberations. Unless also members of the Board, they do not have the right to vote.

2.5 Risk management and quality assurance (max. ½ page A4)

Quality in the project will be monitored according to a Quality Assurance Plan to be drafted at the project's inception. It will describe quality standard requirements for deliverables, research performed, and experiments. Other reviews will be internal but controlled and monitored by the Project Promoter.

The Project Promoter will ensure that the risks assessment is a continuous process throughout the entire project duration, and will allocate a dedicated slot to address risk assessment in every Board meeting. The Project Promoter will work with each of the WP Leaders to establish contingency plans in the event of delays in work package deliverables,

reduced quality and delays between work packages. The Project Promoter will pay particular attention to potential risks that can have a “snowball effect” for work packages that are dependent on each other. Instrumental in the contingency planning will be a simple but state-of-the-art Risk Management Plan which will be developed within the first six months of the project. The Risk Management Plan will assess the likely severity of each risk and its potential impact on the project; assess the potential probability of the risk and identify the measures that may be necessary to minimize the impact of the risk should it nevertheless occur. The accuracy of identified risks will be reviewed bi-monthly and the Risk Management Plan will be changed, improved and completed accordingly.

2.6 7F project web page

www.habit-project.eu

3. ADDITIONAL ACTIVITIES FRAMEWORK

3.1 Description of proposed additional activities (max. 3 A4 pages for 3.1.1 to 3.1.4)

3.1.1 Current state of art including your relevant previous work

Currently available morphologically annotated texts for Czech, Norwegian and main Ethiopean languages (Amharic, Afaan Oromo, Tigrinya, Somali) as well as for many other languages possess severe deficiencies in following aspects:

1. annotation consistency
2. annotation scheme consistency
3. data size

All of these three issues have huge impact on the performance of the automatic tools that are being trained on such data. While the effect of training data size is well-studied and obvious, the importance of the annotation scheme definition and high-agreement annotation is usually underestimated.

On one hand annotation schemes often describe marginal linguistic phenomena with low inter-annotator agreement that increase data sparseness issues, on the other hand there is only little research in how particular annotation scheme changes (e.g. more or less fine-grained scheme) contribute to the performance of machine learning tools built on resources using these schemes.

Recent research shows that even the most developed languages like English still lack tools with sufficient performance on tasks like part-of-speech tagging, and that further development must include revisiting of the training data annotation in first place.

3.1.2 Objective of additional activities

The main objective of the additional activities within this project lies in the development of an open data portal that will enable easy crowdsourced creating of annotated language resources for the languages of interest in a controlled environment.

The annotation will be under supervision of linguistic experts that will continually evaluate inter-annotator agreement of the contributors and suitability of the annotation scheme for machine-learning tools. This will be achieved by iterative training of these tools on already annotated data, evaluating their performance and repeating with a modified annotation scheme.

Besides quantitatively and qualitatively better language resources this will provide better insights into the studied linguistic phenomena with regard to practical language engineering applications.

Particular objectives of the project are as follows:

- design and implementation of a crowdsourcing portal for morphological linguistic annotation of open data

- methodology for continuous evaluation of the consistency of the annotation and annotation scheme
- practical verification of the methodology and annotation portal on Ethiopean languages by annotating at least 100,000 words during a one-week evaluation session.

3.1.3 Coherence with thematic area Social Sciences and Humanities

The project will contribute to better understanding of formal linguistic aspects of morphology of Ethiopean languages, foster creating new resources for these languages and hereby support preserving cultural heritage of the region as well as improve educational potential of the language technology tools for those languages. This has, besides a cultural dimension, an important economic impact on the development of the region.

3.1.4 Methods and approaches

Linguistic research conducted within this project will be largely driven by an empirical corpus-based paradigm -- representative real world text samples of current language will be collected to be analyzed and subject to annotation and annotation scheme development.

Validation of the annotated data will be performed by training automatic tools using state-of-the-art machine-learning methods and evaluating their performance in terms of part-of-speech tagging accuracy.

For this purpose a one-week intensive workshop will take place where participants will use the implemented system to annotate large amounts of data and evaluate their inter-annotator agreement and possibly revise the annotation schemes.

3.1.5 Description of project plan (max. ½ page A4)

The project work will be closely connected to the HaBiT project work plan that consists of 6 work packages. Within the additional activity, the project will be extended with one complex work package that covers the following tasks:

- design and analysis of a new methodology for intelligent allocation of human annotating task with previously uncovered lingustic data
- verification of the methodology by implementation of a community-based annotation framework
- development and testing of the designed processes in connection with Czech and Norwegian as pivot languages
- evaluation and verification of the framework assets in connection with new testing languages (Amharic, et al.) performed at one-week workshop at the University of Addis Ababa

3.2 Project outputs (max. 1 A4 pages)

3.2.1 Intended short-term outcome(s)

1. Creating PoS-annotated language resources for Czech, Norwegian and main Ethiopean languages: Amharic, Afaan Oromo, Tigrinya, Somali.
2. Design and implementation of efficient semi-automated PoS-annotations, using Czech and Norwegian data
3. Evalatuation of the semi-automated framework on Amharic, Oromo.
4. One-week workshop in Addis Abeba

3.2.2 Intended long-term application of outcome(s)

1. Releasing annotated language resources under some non-restrictive open licence, e.g. Creative Commons (CC-BY-SA) through Lindat/CLARIN infrastructure.
2. Releasing the fast PoS-annotation framework as open source.
3. The project results will make it possible to annotate language data of low resourced languages and contribute to their cultural development.

3.2.3 Additional output(s)

Type of output	Title	Date of accomplishment (mm/yyyy)	Date of realization (mm/yyyy)
D	Research paper or technical report	03/2017	04/2017
O	2 PoS-annotated text corpora	02/2017	03/2017
R	dynamic PoS-annotation framework	11/2016	12/2016
W	workshop at University of Addis Ababa	01/2017	02/2017

3.3 Intellectual property rights management (max. ½ page A4)

The partners foresee that patentable results may come out of the proposed research. As a general rule, each partner will own the patents, which it has generated, or jointly own the patents with responsible collaborators. However, the aim will be that the tools and resources produced in the project should be open; and also in the case of patentable results, access rights should be given to the other partner as well as to the Ethiopian associates. Partners are required to inform their work package leaders of any intellectual property rights acquired or applied for resulting from work in the project. The Project Management Board will be the established forum that will allow the IPR to be fairly distributed back amongst the consortium, whether this be owned jointly by the consortium, licensed out to the most relevant member or divided up for each participant. For IPR that is created during a joint effort the partners, a Technology Management Plan will be developed. It will regulate the right of ownership and exploitation of results, and patents and access rights. The Technology Management Plan will consider the relative contributions of the participants and cover strategies of licensing by territory or for fields and take into consideration any requirements imposed by the partners' domestic law. The Technology Management Plan will be integrated into the Partnership Agreement (Annex III).

4. WORK PACKAGES (WPS) AND TASKS

4.1. Work Packages (WPs) (max. 3 A4 pages)

Note: Present the work packages in detail, using the table provided below. Please copy 4.1.1 to 4.1.11 if applicable.

4.1.1 Project working packages (WP)

WP number	Title	Date of start (mm/yyyy)	Date of end (mm/yyyy)
WP1	Efficient Annotation Framework	September 2016	April 2017

4.1.2 WP number

WP1

4.1.3 WP title

Efficient Annotation Framework

4.1.4 WP leader

Pavel Rychlý

4.1.5 WP start date

2016-09-01

4.1.6 WP end date

2017-04-30

4.1.7 WP objective

The work package objective aims at tasks connected to the design and analysis of a new methodology for crowdsourcing annotation of previously uncovered language data as well as verification of the methodology by implementation of a new annotating framework. The framework will be developed on well-known languages and evaluated on less-covered languages at the University of Addis Ababa workshop.

4.1.8 WP task

1) Design crowd-focused system for manual annotation of corpora.
2) Implementation the system, including the user interface.
3) Testing the annotation system for selected language.
4) Evaluation usability for Ethiopian languages.

4.1.9 WP deliverable

M4 - R - prototype implementation of first version of the annotation framework **M8** - D - publication of the framework and methodology evaluation results

4.1.10 WP milestone

M4 - implementation of testing version of annotating tool and providing the feedback **M8** - final evaluation of crowd-sourced annotation of selected Ethiopian languages

4.1.11 WP Human resources

Qualification level: 8 - Karel Pala (docent), Aleš Horák (docent), Pavel Rychlý (docent), Miloš Jakubíček (RNDr., PhD student), Vít Suchomel (RNDr., PhD student), Vít Baisa (Mgr., PhD student), Marek Medveď (Mgr., PhD student), Jan Bušta (Mgr., PhD student?). Person-months: 5,6

5. PARTNERS AND TEAMS

5.1 Project Promoter

5.1.1 Project Promoter identification

5.1.1.1 Role	Project Promoter
5.1.1.2 Organization full legal name (in Czech)	Masarykova univerzita
5.1.1.2.1 Full legal name in English	Masaryk university
5.1.1.3 Abbreviation	MU
5.1.1.4 ID number (IČ)	00216224
5.1.1.5 VAT number	CZ00216224
5.1.1.6 Organization legal form	public university
5.1.1.7 Registration in Commercial Register	
5.1.1.8 Status of organization by Framework 2014/C 198/01	Research organization
5.1.1.9 Participant identification code (PIC) (if relevant)	999880657
5.1.1.10 Full legal headquarters' address	
5.1.1.10.1 Street, number	Žerotínovo nám. 617/9
5.1.1.10.2 Place/location	Brno
5.1.1.10.3 Post code	601 77
5.1.1.10.4 Country	Czech Republic
5.1.1.11 Bank details	
5.1.1.11.1 Bank full name	Česká národní banka
5.1.1.11.2 Bank code	0710
5.1.1.11.3 Account number	94-41924621
5.1.1.11.4 Specific symbol	
5.1.1.11.5 Variable symbol	
5.1.1.12 Contacts	
5.1.1.12.1 Telephone number	+420 549 49 1111
5.1.1.12.2 E-mail	info@muni.cz
5.1.1.12.3 Official web page(s)	www.muni.cz

5.1.2 Statutory authority of Project Promoter

Degree	First name	Surname	Degree	Position	Personal telephone	Personal email
Doc. PhDr	Mikuláš	Bek	Ph.D.	rector	+420 549 49 1001	rektor@muni.cz

5.1.3 Introduction of Project Promoter's team

5.1.3.1 Project Promoter's team composition and competence (max. ¾ page A4)

The team of the NLP Centre FI MU has a leading position in the area of corpus development and supply as well as the design of systems providing fast corpus access. A scalable algorithm for supporting very large corpora has been proposed (Pomikálek, Rychlý, Kilgarriff, 2009, Suchomel, 2012)". The MU team has developed several other linguistic tools and techniques that can be most useful in applications related to the present project. These range from methodologies for building new comprehensive lexica (Horák & Pala, 2007) to methods for the automatic identification of specialised domain-specific terms (Pala et al., 2008). A system for parsing Czech text has been proposed by MU team members (Kovář et al., 2008).

Thus, it can be concluded that MU team is well able to perform the tasks envisaged in the proposed project, especially with regard to the creation of large corpora and generation of the linguistic resources to be used within the actual assumed applications.

K. Pala and A. Horák will be responsible for supervising the team members who are Ph.D. students and for coordinating their research work and also for the overall management. A. Horák will be as well involved in the design of annotation techniques. P. Rychlý will function in the project as a leading researcher in the area of the building large corpora, corpus tools and their effective exploitation. As a leading programmer he will be helping other team members with the technical tasks. V. Suchomel's role will consist in implementing tools for gathering large text data from the Web. V. Baisa will take care of the design and implementation of the specialized user interfaces. M. Jakubíček will work on creating the program interfaces for processing large text data. V. Kovář's task will be to design and implement automatic annotation techniques. The capacity of the particular team member is clearly expressed in the percentages given in the brackets and it is based on our experience with similar tasks in our previous projects (PRESEMT).

5.1.3.2 List of Project Promoter's team staff (all qualified key members)

First name	Surname	Position in project	*PhD/ post doc	Female researcher (Y/N)**
Karel	Pala	Principal investigator	-	N
Aleš	Horák	Researcher	-	N
Pavel	Rychlý	Researcher	-	N
Vít	Suchomel	Researcher	PhD	N
Vít	Baisa	Researcher	PhD	N
Miloš	Jakubíček	Researcher	PhD	N
Marek	Medveď	Researcher	PhD	N
Jan	Bušta	Researcher	PhD	N

5.1.4 Principal investigator

5.1.4.1 Principal investigator identification

Role in the project	Principal investigator
Degree(s)	Ph.D., docent
First name	Karel
Surname	Pala
Citizenship	Czech
Position in organization	Associate Prof.
Work load in project (0-1.0)	0.2
Telephone	+420 549 49 5616
E-mail	pala@fi.muni.cz
Personal web page*	https://is.muni.cz/person/pala

5.1.4.2 Principal investigator's core activities in project (max. ½ page A4)

Prof. K. Pala has dedicated himself for many years to computer and corpus linguistics, and to the computer processing of natural language, especially Czech. He has achieved the significant success in this area at Faculty of Informatics, Masaryk University (FI MU), the main results include work on Czech morphological analysers Lemma and Ajka, partial syntactic analyser DIS/VADIS and Czech WordNet. He created a Dictionary of Czech surface verb valencies with 15,000 items. He is currently the supervisor of the Natural Language Processing Specialization at FI MU and the head of Natural Language Processing Centre at FI MU. He is a co-chair of the international conference Text, Speech and Dialog and co-editor of the TSD Proceedings printed by Springer Verlag.

Within the project: he will participate in the project management as well as in the research activities, related to corpora building and processing. He will be coordinating the tasks concerned with annotating and evaluating issues. He also will be responsible for supervising young members of the MU team who are Ph.D. students.

5.1.4.3 Principal investigator's internationally refereed (joint) scientific publications (max. ½ page A4)

SOJKA, Petr, Aleš HORÁK, Karel PALA a Pavel RYCHLÝ. MTW 2012 -- **Hybrid Machine Translation, Machine Translation Workshop, Brno, Czech Republic, September 3, 2012**. Edited by Sojka P., Horák A., Kopeček I., Pala K. prvni. Brno: Tribun EU a Springer Verlag, 2012. 68 s. ISBN 978-80-263-0266-7.

SOJKA, Petr, Aleš HORÁK, Ivan KOPEČEK a Karel PALA. T**ext, Speech and Dialogue: 15th International Conference TSD 2012, Brno, Czech Republic, September 3-7, 2012**. Edited by Sojka P., Horák A., Kopeček I., Pala K. Berlin Heidelberg: Springer Verlag, 2012. 697 s. ISBN 978-3-642-32789-6. doi:10.1007/978-3-642-32790-2.

PALA, Karel a Pavel RYCHLÝ. A Case Study in Word Sketches - Czech Verb vidět 'see'. In **A Way with Words: Recent Advances in Lexical Theory and Analysis**. Uganda: Menha Publishers Ltd., 2010. s. 187-198, 12 s. Neuveden. ISBN 978-9970-10-101-6.

PALA, Karel, Pavel RYCHLÝ a Pavel ŠMERK. A**utomatic Identification of Legal Terms in Czech Law Texts. In Semantic Processing of Legal Texts**. Berlin: Springer, 2010. s. 83-94, 12 s. ISBN 978-3-642-12836-3.

HORÁK, Aleš, Karel PALA a Dana HLAVÁČKOVÁ. Preparing VerbaLex Printed Edition. In **Seventh Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2013**. Brno: Tribun EU, 2013. s. 3-11, 9 s. ISBN 978-80-263-0520-0.

PALA, Karel a Ondřej SVOBODA. Semi-automatic Theme-Rheme Identification. In **Seventh Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2013**. Brno: Tribun EU, 2013. s. 39-48, 10 s. ISBN 978-80-263-0520-0.

PALA, Karel a Pavel RYCHLÝ. **Do we need very large corpora?** Praha (Prague): Nakladatelství Lidové Noviny, 2011. s. 33-39, 379 s. ISBN 978-80-7422-114-9.

5.2 Project Partner(s)

5.2.1 Project Partner identification

5.2.1.1 Role	Project Partner
5.2.1.2 Organization legal name (in Czech/Norwegian or other)	Norges teknisk-naturvitenskapelige universitet
5.2.1.2.1 Full legal name in English	Norwegian University of Science and Technology
5.2.1.3 Abbreviation	NTNU
5.2.1.4 ID number	974 767 880
5.2.1.5 VAT number	NO 974 767 880 MVA
5.2.1.6 Organization legal form	Organization section, Higher education, Public entity
5.2.1.7 Registration in Commercial Register	Brønnøysundregistrene: 974 767 880
5.2.1.8 Status of organization by Framework 2014/C 198/01	Public body. Higher education establishment
5.2.1.9 Participant identification code (PIC) (if relevant)	999977851
5.2.1.10 Full legal headquarters' address	
5.2.1.10.1 Street, number	Høgskoleringen 1
5.2.1.10.2 Place/location	Trondheim
5.2.1.10.3 Post code	7491
5.2.1.10.4 Country	Norway
5.2.1.11 Contacts	
5.2.1.11.1 Telephone number	+47 73 59 50 00 (switchboard)

5.2.1.11.2 E-mail	postmottak@adm.ntnu.no
5.2.1.11.3 Official web page(s)	www.ntnu.no

5.2.2 Statutory authority of Project Partner

Degree	First Name	Surname	Degree	Position	Personal telephone	Personal email
PhD	Kari	Melby		Pro-Rector for research	+4773598011	kari.melby@ntnu.no
PhD	Johan E.	Hustad		Pro-Rector for Innovation	+4773598011	johan.e.hustad@ntnu.no

5.2.3 Introduction of Project Partner's team

5.2.3.1 Project Partner's team composition and competence (max. ¾ page A4)

Note: Since the Project Partner does not request any part of the funding for the Additional Activities, its team composition is not included here. For the relevant details, please refer to the original HaBiT project contract.

5.2.3.2 List of Project Partner's team staff (all qualified key members)

First name	Surname	Position in project	*PhD/ post doc	Female researcher (Y/N)**
-	-	-	-	-

5.2.4 Leader of Norwegian Project Partner(s)

5.2.4.1 Norwegian leader identification

Role in the project	Head of research team
Degree(s)	Dr.Sci.
First name	Björn
Surname	Gambäck
Citizenship	Sweden
Position in organization	Professor of Language Technology
Work load in project (0-1.0)	0 (Additional Activities)
Telephone	+46 70 568 1535
E-mail	gamback@idi.ntnu.no
Personal web page*	

5.2.4.2 Norwegian leader's core activities in project (max. ½ page A4)

The Norwegian team does not directly participate in the development of the additional activity, however, they will participate at the evaluation workshop at the University of Addis Ababa.
The Norwegian team within the HaBiT project is lead by Björn Gambäck who has been a professor at NTNU since 2008 and 2010-2012 was NTNU's principal investigator in the FP7 project PRESEMT (Pattern REcognition-based Statistically Enhanced Machine Translation, ICT-248307; Language Technology), and currently leads NTNU's activities in the NORHED collaboration with Ethiopia and NTNU's involvement in the Norwegian national language resource project CLARINO. Previously, Prof. Gambäck was a project leader at SICS for over 20 years and there coordinated a number of national and international projects, including projects both in FP5 (DUMAS: Dynamic Universal Mobility for Adaptive Speech Interfaces, IST-2000-29452; Language Technology) and FP6 (the IP EVERGROW: Ever-growing global scale-free networks: their provisioning, repair and unique functions, IST-2004-001935; Complex Systems). He has also been the principal investigator at SICS in projects such as the FP6 IP COMPANIONS (Intelligent, Persistent, Personalised Multimodal Interfaces to the Internet, IST-034434; Language Technology) and FP5 Research Infrastructure project SCHOLNET (A Digital Library Testbed to Support Networked Scholarly Communities, IST-1999-20664; Digital Libraries).

5.2.4.3 Norwegian leader's internationally refereed (joint) scientific publications (max. ½ page A4)

Asker, Lars, Atelach Alemu Argaw, Björn Gambäck, and Magnus Sahlgren. 2007. Applying Machine Learning to Amharic Text Classification. In *Proceedings of the 5th World Congress of African Linguistics Addis Ababa 2006*. Rüdiger Köppe Verlag.

Asker, Lars, Atelach Alemu Argaw, Björn Gambäck, Samuel Eyassu, and Lemma Nigussie. 2009. Classifying Amharic Webnews. *Information Retrieval*, **12**(3):416-435.

Bungum, Lars and Björn Gambäck. 2012. Efficient N-Gram Language Modeling for Billion Word Web-Corpora. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey. ELRA. Workshop on Challenges in the Management of Large Corpora.

Gambäck, Björn. 2012. Tagging and Verifying an Amharic News Corpus. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation,* pages 79-84. European Language Resources Association.

Gambäck, Björn and Lars Asker. 2010. Experiences with Developing Language Processing Tools and Corpora for Amharic. In *Proceedings of the 5th Conference on Regional Impact of Information Society Technologies in Africa*, Durban, South Africa.

Marsi, Erwin, André Lynum, Lars Bungum, and Björn Gambäck. 2011. Word Translation Disambiguation without Parallel Texts. In *Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation*, pp. 66-74, Barcelona, Spain.

Marsi, Erwin, Hans Moen, Lars Bungum, Gleb Valerjevich Sizov, Björn Gambäck, and André Lynum. 2013. NTNU-CORE: Combining strong features for semantic similarity. In: *Second Joint Conference on Lexical and Computational Semantics (*SEM)*. Association for Computational Linguistics.

Moen, Hans, Erwin Marsi, and Björn Gambäck. 2013. Towards Dynamic Word Sense Discrimination with Random Indexing. In: *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*. Association for Computational Linguistics.

6. ADDITIONAL ACTIVITIES BUDGET

6.1 Additional budget rates

Rate of requested grant in the total additional budget (in %)	100%
Project Promoter share of total requested grant (in %)	Masaryk university, 100 %
Project Partner share of total requested grant (in %)	Norwegian University of Science and Technology, 0 %
Used indirect cost percentage (overheads) in project contract – Project Promoter	Masaryk university, 100 %
Used indirect cost percentage (overheads) in project contract – Project Partner	Norwegian University of Science and Technology, 0 %

6.2 Additional budget and requested funding justification (max. 3 pages A4)

The requested funding corresponds in case of the Project Promoter to 100% of the planned budget as it is public research organization (university) and the proposed project does not include commercial activities.

The planned budget for Masaryk University as presented in the Annex II of this proposal consists of the following parts (in terms according to the Eligible cost structure from the Guide for Applicants):

Preparatory costs - there are no preparatory costs planned in the budget proposal

Personnel cost

The MU Personnel costs consists of actual salaries (including usual remuneration) staff assigned to the project. The amounts are computed from average income tables for the corresponding positions (associate professor, assistant professor, researcher) at the Faculty of Informatics MU plus social security charges and other statutory costs. All the proposed staff members are MU employees at the moment. The particular personnel costs are based on the following roles and working loads of the project staff members:

- doc.PhDr. Karel Pala, CSc., 5%, associate professor (4 000 CZK/month) - Principal Investigator, project team leader
- doc.RNDr. Aleš Horák, Ph.D., 5%, associate professor (4 000 CZK/month) - design and analysis of annotation techniques and structures
- Pavel Rychlý, Ph.D., 5%, assistant professor (4 000 CZK/month) - design and analysis of effective processing of large textual data
- Mgr. Vít Suchomel, 5%, researcher (2 250 CZK/month) - implementation of tools for obtaining language data from Internet
- Mgr. Vít Baisa, 5%, researcher (2 250 CZK/month) - design and implementation of specialized user interfaces

- RNDr. Miloš Jakubíček, 5%, researcher (2 650 CZK/month) - creation of program interfaces for big data processing
- Mgr. Marek Medveď, 30%, resercher (11 400 CZK/month) - design and implementation of automatic annotation techniques
- Mgr. Jan Bušta, 10%, reasercher (3 800 CZK/month) - design and implementation of automatic annotation techniques
- Bc. Marie Stará, 0%, project administrator (0 CZK/month)- project administrator. The administration tasks will be included in the whole HaBiT project administration, that is why no additional funding is requested for the project administrator.

Travel and subsistence

The travel and subsistence costs include:

- active participation of MU research team members at one-week workshop in Addis Abeba with the cost of 332 320 CZK.

New or second hand equipment

The MU NLP Centre is equipped with large computer servers and storage facilities for processing very big text databases, which is regularly upgraded. Therefore we do not plan any new equipment for the project.

Consumables and supplies

The MU does not request any extra funding for consumables and supplies.

Other costs

No other costs are planned in the project. Dissemination and publication costs are included in the travel costs stated above. Subcontracting is not planned in the MU budget.

Indirect cost (overheads)

According to the MU statutory rules, the MU institutional overhead costs are computed as a flat rate of 60% of the total direct eligible costs (no subcontracting and no costs of third party resources is planned).

7. MANDATORY ANNEXES

7.1 Overview of annexes required to proposal

Completed proposal form		Doc(x), signed pages in pdf
No.	**Mandatory Annexes**	**Mandatory electronic (CD) submission format**
I.	*Annex I – The draft of budget for additional research activities* in Czech Crowns (CZK).	xls(x)
II.	*Annex II – The draft of total project budget* in Czech Crowns (CZK).	xls(x)
III.	*Annex III - Statutory declarations* (Project Promoter and each Czech Project Partner).	pdf
IV.	*Annex IV - Consent to processing personal data* – for all entities	pdf
V.	*Annex V – The draft of a contract of a new project team member*, if applicable (a copy is sufficient, no template available).	pdf
VI.	*Annex VI – The resume of a new research team member* if applicable (Europass form, no template available).	pdf
VII.	*Annex VII – The revision of the Partnership Agreement*, if applicable (signed original must be delivered before issuing revised Project Contract).	pdf
VIII.	*Annex VIII - The power of attorney for authorized person of the Project Promoter*, if applicable.	pdf

Note: For electronic submission the required format for the proposal form is docx (you can also submit the undersigned pages in pdf in a separate file).
Please tie up the documents in this order: completed proposal form, mandatory annexes (I-VIII), any voluntary annexes.

8. OTHER

You may add other information you think necessary.