FINAL PROJECT REPORT 2017			
The Czech-Norwegian Research Programme (CZ09)			
The Norwegian Financial Mechanism 2009-2014			
Programme area	**Bilateral Research Cooperation**		
Project ID number	7F14047		
Project contract no.	MSMT-19538/2016-1		
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		
Thematic area (choose)	Social Sciences and Humanities		
Project website	www.habit-project.eu		
Project Promoter (Full name and address)	Masarykova univerzita Žerotínovo nám. 617/9, 601 77 Brno Czech Republic		
Project Partner 1 (Full name and address)	Norges teknisk-naturvitenskapelige universitet Høgskoleringen 1, 7491 Trondheim Norway		
Full name of Principal Investigator (PI)	doc. PhDr. Karel Pala, CSc.		
PI Contacts (@ + phone)	pala@fi.muni.cz ; +420 549 49 5616		
Signature of PI			
Statement	*I hereby declare that the information I state in the Final Project Report is accurate, true and complete. I am aware that, if this is not the case, I will face sanctions from the Programme Operator.*		
Done in	Brno, Czech Republic		
Date	30/06/2017		
On behalf of Project Promoter			
Stamp of Project Promoter			
Statutory authority of Project Promoter	**Name(s):**	doc. PhDr. Mikuláš Bek, Ph.D.	
	Signature(s):		
	Position:	rector	

1. GENERAL INFORMATION ABOUT PROJECT

1.1 Main research activity in the project

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

Note: Tick the same research activities as specified in the project contract to identify the relevant activity, or combination of activities. For combination, mark the percentage of each type of the activity. If you tick only one option, it means that the project is 100% of the indicated R&D activity.

1.2 Project start date

01/10/2014

1.3 Project end date

30/04/2017

1.4 Total approved project costs per project (in CZK)

25 593 000

1.4.1 Total approved grant per project (in CZK)

25 593 000

Note: The amount excludes the payment from the Fund for bilateral relations (preparatory costs in Measure I.; or bilateral activities costs in Measure II. (B)).

2. SCIENTIFIC OUTCOMES AND MANAGEMENT

2.1 Publishable summary in English (max. ½ page A4)

The main objective of the HaBiT project was to gather large-scale text data (corpora) from web and to process them so they can be used in language applications for e.g. information extraction or machine translation. The project focused on Norwegian, Czech and Amharic, Afaan Oromo, Tigrinya and Somali (these being four major Ethiopian languages).

Large annotated corpora were created for all given languages and software modules, such as taggers, parsers and sketch grammars were created. The results were presented to the scientific community via conference and journal papers as well as the project web page.

The outputs are freely accessible for further research via the HaBiT system created with cooperation of project partners -- Masaryk University and Norwegian University of Technology -- and University of Oslo and two Ethiopian universities, which cooperate with NTNU on NORHED project.

The HaBiT project leveraged 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 a relevant added value also in the political respect through cooperation with a less-developed country.

2.1.1 Publishable summary in Czech (max. ½ page A4)

Hlavním cílem projektu HaBiT bylo shromáždit velká textová data (textové korpusy) z webu a zpracovat je tak, aby mohla být použita v jazykových aplikacích jako např. extrakce informací nebo strojový překlad. Projekt se zaměřil na norštinu, češtinu a dále na amharštinu, afaan oromštinu, tigrinštinu a somálštinu, což jsou čtyři hlavní etiopské jazyky.

Pro všechny zmíněné jazyky byly vytvořeny velké anotované korpusy a spolu s nimi softwarové moduly jako taggery (značkovače), parsery (syntaktické analyzátory) a sketchové grammatiky. Získané výsledky byly představeny odborné veřejnosti na mezinárodních konferencích a

v odborných časopiseckých článcích a zejména na webové stránce projektu.
Výstupy jsou volně přístupné pro další výzkum prostřednictvím systému HaBiT, který vznikl v kooperaci s projektovými partnery -- Masarykovou univerzitou a Norskou technickou univerzitou (NTNU) a univerzitou v Oslo a dále dvěma etiopskými universitami, které spolupracují s NTNU na projektu NORHED.
Projekt HaBiT se napojil na NORHED projekt podrobným testováním vzniklých technologií a adresoval tak aktuální výzvu k hodnocení vyvinutých technologií, jejich verifikaci a testování. Reaguje rovněž na splňování společenských výzev ze strany informačních technologií (ICT) a poskytuje též relevantní přidanou hodnotu i v politickém ohledu díky kooperaci s méně rozvinutou zemí (Etiopií).

2.2 Achievement of objectives, outcomes and bilateral contribution for DoRIS (max. 1000 characters)

1. Why was the project needed? How will the outcomes be sustained?
Intelligent language processing applications (such as information extraction or machine translation) need processed (annotated and parsed) large text data. Big languages (English, German, ...) have enough resources available, however, less-covered languages such as the Ethiopian languages but also hundreds of other languages, are in urgent need of methodologies and techniques for linguistic resource creation. The HaBiT project developed such methodologies and techniques and presented them with the case of four main Ethiopian languages.
2. What was the project outcome, and to what extent was it reached?
The expected outcomes were large annotated corpora for the participating under-resourced languages and Part-of-Speech annotation framework as well as taggers, parsers and sketch grammars for the involved languages. All the outcomes were achieved as planned and are accessible via the publicly available interface of the HaBiT system.
3. What objective was achieved, and to what extent was it reached? What was the impact?
The objective of the project was to create annotated corpora, annotation framework and to help acquiring information technology in a less-developed country. This objective was fulfilled in both aspects: a) general methodologies for obtaining linguistic resources (corpora, taggers, sketch grammars) for a new language were developed and published, and b) the methodologies were explicated in the form of the respective linguistic resources for the four Ethiopian languages, where they represent the state-of-the-art for these languages.
4. Which outputs were delivered?
All of planned outputs were delivered: annotated corpora and sketch grammars for all involved languages were created. Semantic search interface was developed as well as dynamic concept matching. The PoS annotation framework was developed. The HabiT system was finished.
5. How were the beneficiaries involved? What was their main benefit?
Scientist from the University of Addis Ababa participated within the evaluation of the newly built resources and they highly appreciate the unbeatable quality of these project outcomes.
6. What did the donor partnership achieved? If applicable.
7. What did the project partner/donor project partner contribute to the project at a technical/professional level?

Masaryk university offered the world-leading expertise in corpus preparation and natural language processing. The Norges teknisk-naturvitenskapelige universitet introduced new results within the area of word-level semantic matching and disambiguation and multi-lingual word spaces. The cooperating institutions of the University of Oslo arched over the linguistic evaluation of the project results.
8. What did the partnership contribute to the project outcome and outputs?
Within the partnership, three research meetings in Oslo and Brno were organized, an information meeting for the students of the Faculty of Informatics and the Faculty of Arts about the Czech-Norwegian project was held and two international workshops devoted to building new language resources for languages with no or too little existing language resources and to practical evaluation of the efficient annotation framework were supported.
9. What did the partnership achieve concerning strengthened bilateral relations?
The cooperation between MU and NTNU continued from the previous EU project PRESEMT and it was institutionalized within the HaBiT project leading to new state-of-the-art results in the area of linguistic resources of under-resourced languages, which are being exploited by the beneficiaries even after the actual end of the project.
10. What are the socio-economic impact and the wider societal implications of the project, including gender equality actions, ethical issues, and wider awareness?
The two international workshops organized within the HaBiT project received high ratio of attendants outside the project partners from all over the world. The interim results of the project have been published in 40 scientific conference papers and 2 international journal articles. The application results are publicly available in the form of software tools and linguistic databases.

2.3 Achievement of indicators of the Programme at project level (numbers per project, e.g. 21)

1. Total number of Ph.D. students involved	**6**
2. Total number of postdocs involved	**4**
3. Total number of female researchers	**2**
3.1 How many total number female researchers returned after maternity leave	**0**
4. Total number of internationally refereed (joint) scientific publications	**44**

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

2.4 Summary of achieved outcomes at project level

Type of outcome	Title	Accomplished/or Submitted (mm/yyyy)
D	Annotation of Multi-Word Expressions in Czech Texts	10/2015
D	DEBWrite: Free Customizable Web-based Dictionary Writing System	04/2015
D	Increasing Coverage of Translation Memories with Linguistically Motivated Segment Combination Methods	06/2015
D	Interactive Visualizations of Corpus Data in Sketch Engine	01/2015
D	Longest-commonest Match	04/2015
D	SemEval-2015 Task 15: A CPA dictionary-entry-building task	08/2014

D	Towards Automatic Finding of Word Sense Changes in Time	10/2015
D	D Matching logic for mono- and multi-lingual word space models	08/2016
D	Semantic Search in Large Word Space Models	01/2017
D	Multi-sense Random Indexing	03/2017
D	Annotated Amharic Corpora	06/2016
D	Annotation of Czech Texts with Language Mixing	06/2016
D	AQA: Automatic Question Answering System for Czech	06/2016
D	Czech Grammar Agreement Dataset for Evaluation of Language Models	10/2016
D	DSL Shared task 2016: Perfect Is The Enemy of Good Language Discrimination Through Expectation-Maximization and Chunk-based Language Model	09/2016
D	English-French Document Alignment Based on Keywords and Statistical Translation	05/2016
D	European Union Language Resources in Sketch Engine	03/2016
D	Evaluating Natural Language Processing Tasks with Low Inter-Annotator Agreement: The Case of Corpus Applications	10/2016
D	Evaluation and Improvements in Punctuation Detection for Czech	06/2016
D	Finding Definitions in Large Corpora with Sketch Engine	03/2016
D	Graded and Word-Sense-Disambiguation Decisions in Corpus Pattern Analysis: a Pilot Study	03/2016
D	Large Scale Keyword Extraction using a Finite State Backend	10/2016
D	Multilingual CPA: Linking Verb Patterns across Languages	10/2015
D	On Evaluation of Natural Language Processing Tasks: Is Gold Standard Evaluation Methodology a Good Solution	12/2015
D	RuSkELL: Online Language Learning Tool for Russian Language	10/2015
D	VPS-Grade-Up: Graded Decisions on Usage Patterns	03/2016
J	Lexicographic Tools to Build New Encyclopaedia of the Czech Language	07/2016
J	Sketch Engine for Bilingual Lexicography	03/2016
C	Walking the tightrope between linguistics and language engineering	08/2016
R	HaBiT system	04/2017
R	PoS-annotation framework	12/2016
R	Set of Ethiopian Web Corpora	08/2016
W	Workshop at the Masaryk university	02/2017

2.5 Summary of outcomes dissemination and publicity (max. ½ page A4)

The information about the project are presented online on the project web page http://habit-project.eu/ and also on the Masaryk University web page (both in Czech and English). Other communication tools actually used were active participation on conferences and workshops and display of posters and roll-ups. These tools were used to target research community and students as well as broader public.
The list of conferences includes: The XVII EURALEX International congress, 8th International Conference on Agents and Artificial Intelligence, LREC 2016, The First Conference on Machine Translation, VarDial3, TSD 2016, RASLAN 2016, eLex 2015, NODALIDA 2015, SemEval 2015, the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA) at EMNLP 2015, RANLP 2015, NLP4TM 2015, CONTEXT 2015, FIRE 2015, ICON 2015, RASLAN 2015.

2.5.1 Reporting on internationally referred (joint) scientific publications

Reference: HORÁK, Aleš a Adam RAMBOUSEK. **Lexicographic Tools to Build New Encyclopaedia of the Czech Language**. The Prague Bulletin of Mathematical Linguistics, Prague (Czech Republic): Charles University, 2016, roč. 2016, č. 106, s. 205-213. ISSN 0032-6585. doi:10.1515/pralin-2016-0019.
Name of scientific journal: The Prague Bulletin of Mathematical Linguistics
Date of publication/OR submission: 10/2016
Abstract: The first edition of the Encyclopaedia of the Czech Language was published in 2002 and since that time it has established as one of the basic reference books for the study of the Czech language and related linguistic disciplines. However, many new concepts and even new research areas have emerged since that publication. That is why a preparation of a complete new edition of the encyclopaedia started in 2011, rather than just reprinting the previous version with supplements. The new edition covers current research status in all concepts connected with the linguistic studies of (prevalently, but not solely) the Czech language. The project proceeded for five years and it has finished at the end of 2015, the printed edition is currently in preparation. An important innovation of the new encyclopaedia lies in the decision that the new edition will be published both as a printed book and as an electronic on-line encyclopaedia, utilizing the many advantages of electronic dictionaries. In this paper, we describe the lexicographic platform used for the Encyclopaedia preparation and the process behind the work flow consisting of more than 3,000 pages written by nearly 200 authors from all over the world. The paper covers the process of managing entry submissions, the development of tools to convert word processor files to an XML database, tools to cross-check and connect bibliography references from free text to structured bibliography entries, and the preparation of data for the printed publication.

Reference: KOVÁŘ, Vojtěch, Vít BAISA a Miloš JAKUBÍČEK. **Sketch Engine for Bilingual Lexicography**. International Journal of Lexicography, Oxford: Oxford University Press, 2016, roč. 29, č. 3, s. 339-352. ISSN 0950-3846. doi:10.1093/ijl/ecw029.
Name of scientific journal: International Journal of Lexicography
Date of publication/OR submission: 09/2016
Abstract: Sketch Engine is a leading corpus query and corpus management tool that has been used for many large dictionary projects. The paper summarizes its features supporting bilingual lexicography and the creation of bilingual learner's dictionaries. Some of these features have been added recently; some of them have been part of the software for a rather long time, but they have been recently improved.

KUMAR, Upendra, Aishwarya N. REGANTI, Tushar MAHESHWARI, Tanmoy CHAKROBORTY, Björn GAMBÄCK and Amitava DAS. **Inducing Personalities and Values from Language Use in Social Network Communities.** To appear in Information Systems Frontiers, Special Issue on Mining Human Psycholinguistic Behaviour from Social Media. ISSN 1387-3326. Springer.

2.5.2 Reporting on other (joint) scientific publications/articles/patents/prototype/thesis etc.

D - Vít Baisa, Jane Bradbury, Silvie Cinková, Ismaïl El Maarouf, Adam Kilgarriff, Octavian Popescu. SemEval-2015 Task 15: A CPA dictionary-entry-building task. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Denver, Colorado: Association for Computational Linguistics, 2015. s. 315-324, 10 s. ISBN 978-1-941643-40-2. https://is.muni.cz/publication/1308719

D - Adam Kilgarriff, Vít Baisa, Miloš Jakubíček, Pavel Rychlý. Longest-commonest Match. In Kosem, I., Jakubíček, M., Kallas, J., Krek, S.. Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom. Ljubljana: Trojina, Institute for Applied Slovene Studies, 2015. s. 397-404, 8 s. ISBN 978-961-93594-3-3. https://is.muni.cz/publication/1308616

D - Lucia Kocincová, Miloš Jakubíček, Vojtěch Kovář, Vít Baisa. Interactive Visualizations of Corpus Data in Sketch Engine. In Gintaré Grigonyté, Simon Clematide, Andrius Utka, Martin Volk. Proceedings of the Workshop on Innovative Corpus Query and Visualization Tools at NODALIDA 2015. Vilnius, Lithuania: Linköping University Electronic Press, Linköpings universitet, 2015. s. 17-22, 6 s. ISBN 978-91-7519-035-8. https://is.muni.cz/publication/1299713

D - Adam Rambousek, Aleš Horák. DEBWrite: Free Customizable Web-based Dictionary Writing System. In Kosem, I., Jakubiček, M., Kallas, J., Krek, S.. Electronic lexicography in the 21st century: linking lexical data in the digital age. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd., 2015. s. 443-451, 9 s. ISBN 978-961-93594-3-3. https://is.muni.cz/publication/1308365

D - Vít Baisa, Ondřej Herman, Miloš Jakubíček. Towards Automatic Finding of Word Sense Changes in Time. In Aleš Horák, Pavel Rychlý, Adam Rambousek. Ninth Workshop on Recent Advances in Slavonic Natural Language Processing. Brno: Tribun EU, 2015. s. 33-41, 9 s. ISBN 978-80-263-0974-1. https://is.muni.cz/publication/1318600

D - Zuzana Nevěřilová. Annotation of Multi-Word Expressions in Czech Texts. In Aleš Horák, Pavel Rychlý, Adam Rambousek. Ninth Workshop on Recent Advances in Slavonic Natural Language Processing. Brno: Tribun EU, 2015. s. 103-112, 10 s. ISBN 978-80-263-0974-1. https://is.muni.cz/publication/1320593

D - Marek Medveď, Vít Baisa, Aleš Horák. Increasing Coverage of Translation Memories with Linguistically Motivated Segment Combination Methods. In Constantin Orasan and Rohit Gupta. Proceedings of The Workshop on Natural Language Processing for Translation Memories (NLP4TM). Bulgaria: INCOMA Ltd. Shoumen, 2015. s. 31-35, 5 s. ISBN 978-954-452-032-8. https://is.muni.cz/publication/1311833

D - 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 Computational Linguistics. http://aclweb.org/anthology/W15-2914

D - 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. http://aclweb.org/anthology/R15-1033

D - 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. http://link.springer.com/chapter/10.1007/978-3-319-25591-0_15

D - 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. December 2015.
D - 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. December 2015.
D - Collecting and Annotating Indian Social Media Code-Mixed Corpora - Anupam Jamatia, Björn Gambäck, Amitava Das. 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing), Konya, Turkey. April 2016.
D - Comparing the Level of Code-Switching in Corpora - Björn Gambäck, Amitava Das. The 10th International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia. May 2016. http://www.lrec-conf.org/proceedings/lrec2016/summaries/669.html
D - NTNUSentEval at SemEval-2016 Task 4: Combining General Classifiers for Fast Twitter Sentiment Analysis - Brage Ekroll Jahren, Valerij Fredriksen, Björn Gambäck, Lars Bungum. 10th International Workshop on Semantic Evaluation (SemEval) at the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'2016), San Diego, California. June 2016, pp. 103–108. http://aclweb.org/anthology/S/S16/S16-1014.pdf
D - Linguistic Domains and Adaptable Companionable Agents - Björn Gambäck, Lars Bungum. 1st International Workshop on Domain Adaptation for Dialog Agents at the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD), Riva del Garda, Italy. September 2016.
D - Language Identification in Code-Switched Text Using Conditional Random Fields and Babelnet - Utpal Kumar Sikdar, Björn Gambäck. The 2nd Workshop on Computational Approaches to Code Switching at the 2016 Conference on Empirical Methods on Natural Language Processing (EMNLP), Austin, Texas. November 2016. http://www.aclweb.org/anthology/W/W16/W16-5817.pdf
D - Feature-Rich Twitter Named Entity Extraction and Classification - Utpal Kumar Sikdar, Björn Gambäck. The 2nd Workshop on Noisy User-generated Text (W-NUT) at the 26th International Conference on Computational Linguistics (COLING), Osaka, Japan. December 2016. http://www.aclweb.org/anthology/W/W16/W16-3922.pdf
D - Twitter Named Entity Extraction and Linking Using Differential Evolution - Utpal Kumar Sikdar, Björn Gambäck. The 13th International Conference on Natural Language Processing (ICON), Varanasi, Uttar Pradesh, India. December 2016, pp. 198-207.
D - Pavel Rychlý, Vít Suchomel. Annotated Amharic Corpora. In Petr Sojka, Aleš Horák, Ivan Kopeček, Karel Pala. Text, Speech, and Dialogue 19th International Conference, TSD 2016 Brno, Czech Republic, September 12–16, 2016 Proceedings. Switzerland: Springer International Publishing, 2016. s. 295-302, 8 s. ISBN 978-3-319-45509-9. https://is.muni.cz/publication/1353390
D - Zuzana Nevěřilová. Annotation of Czech Texts with Language Mixing. In Petr Sojka; Aleš Horák; Ivan Kopeček; Karel Pala. Text, Speech, and Dialogue 19th International Conference, TSD 2016 Brno, Czech Republic, September 12–16, 2016 Proceedings. Switzerland: Springer International Publishing, 2016. s. 279-286, 8 s. ISBN 978-3-319-45509-9. doi:10.1007/978-3-319-45510-5_32. https://is.muni.cz/publication/1358121
D - Marek Medveď, Aleš Horák. AQA: Automatic Question Answering System for Czech. In Sojka Petr, Horák Aleš, Kopeček Ivan, Pala Karel. Text, Speech, and Dialogue 19th International Conference, TSD 2016 Brno, Czech Republic, September 12–16, 2016 Proceedings. Switzerland: Springer International Publishing, 2016. s. 270-278, 9 s. ISBN 978-3-319-45510-5. doi:10.1007/978-3-319-45510-5_31. https://is.muni.cz/publication/1353405
D - Vít Baisa. Czech Grammar Agreement Dataset for Evaluation of Language Models. In RASLAN 2016 Recent Advances in Slavonic Natural Language Processing. Brno: Tribun EU, 2016. s. 63-67, 5 s. ISBN 978-80-263-1095-2. https://is.muni.cz/publication/1362555
D - Ondřej Herman, Vít Suchomel, Vít Baisa, Pavel Rychlý. DSL Shared task 2016: Perfect Is The Enemy of Good Language Discrimination Through Expectation-Maximization and Chunk-based Language Model. In Preslav Nakov, Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, Shervin Malmasi.

Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3). Osaka: The COLING 2016 Organizing Committee, 2016. s. 114-118, 5 s. ISBN 978-4-87974-716-7. https://is.muni.cz/publication/1366107

D - Marek Medveď, Vojtěch Kovář, Miloš Jakubíček. English-French Document Alignment Based on Keywords and Statistical Translation. In Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers. Berlin: Association for Computational Linguistics, 2016. s. 728-732, 5 s. ISBN 978-1-945626-10-4. https://is.muni.cz/publication/1352922

D - Vít Baisa, Jan Michelfeit, Marek Medveď, Miloš Jakubíček. European Union Language Resources in Sketch Engine. In Nicoletta Calzolari (Conference Chair) and Khalid Choukri and Thierry Declerck and Marko Grobelnik and Bente Maegaard and Joseph Mariani and Asuncion Moreno and Jan Odijk and Stelios Piperidis. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia: European Language Resources Association (ELRA), 2016. s. 2799-2803, 5 s. ISBN 978-2-9517408-9-1. https://is.muni.cz/publication/1346032

D - Vojtěch Kovář. Evaluating Natural Language Processing Tasks with Low Inter-Annotator Agreement: The Case of Corpus Applications. In Aleš Horák, Pavel Rychlý, Adam Rambousek. Tenth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2016. Brno: Tribun EU, 2016. s. 127-134, 8 s. ISBN 978-80-263-1095-2. https://is.muni.cz/publication/1365039

D - Vojtěch Kovář, Jakub Machura, Kristýna Zemková, Michal Rott. Evaluation and Improvements in Punctuation Detection for Czech. In Petr Sojka; Aleš Horák; Ivan Kopeček; Karel Pala. ext, Speech, and Dialogue 19th International Conference, TSD 2016 Brno, Czech Republic, September 12–16, 2016 Proceedings. Switzerland: Springer, 2016. s. 287-294, 8 s. ISBN 978-3-319-45509-9. https://is.muni.cz/publication/1358120

D - Vojtěch Kovář, Monika Močiariková, Pavel Rychlý. Finding Definitions in Large Corpora with Sketch Engine. In Nicoletta Calzolari (Conference Chair) et al.. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia: European Language Resources Association (ELRA), 2016. s. 391-394, 4 s. ISBN 978-2-9517408-9-1. https://is.muni.cz/publication/1360550

D - Silvie Cinkova, Ema Krejčová, Anna Vernerová, Vít Baisa. Graded and Word-Sense-Disambiguation Decisions in Corpus Pattern Analysis: a Pilot Study. In Nicoletta Calzolari (Conference Chair) and Khalid Choukri and Thierry Declerck and Marko Grobelnik and Bente Maegaard and Joseph Mariani and Asuncion Moreno and Jan Odijk and Stelios Piperidis. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia: European Language Resources Association (ELRA), 2016. s. 848-854, 7 s. ISBN 978-2-9517408-9-1. https://is.muni.cz/publication/1346038

D - Miloš Jakubíček, Pavel Šmerk. Large Scale Keyword Extraction using a Finite State Backend. In Aleš Horák, Pavel Rychlý, Adam Rambousek. Tenth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2016. Brno: Tribun EU, 2016. s. 143-146, 4 s. ISBN 978-80-263-1095-2. https://is.muni.cz/publication/1365139

D - Vít Baisa, Sara Može, Irene Renau. Multilingual CPA: Linking Verb Patterns across Languages. In Tinatin Margalitadze, George Meladze. Proceedings of the XVII EURALEX International congress. Tbilisi: Ivane Javakhishvili Tbilisi State University, 2016. s. 410-417, 8 s. ISBN 978-9941-13-542-2. https://is.muni.cz/publication/1352903

D - Vojtěch Kovář, Miloš Jakubíček, Aleš Horák. On Evaluation of Natural Language Processing Tasks: Is Gold Standard Evaluation Methodology - a Good Solution In Jaap van den Herik and Joaquim Filipe. Proceedings of the 8th International Conference on Agents and Artificial Intelligence. Rome: SCITEPRESS, 2016. s. 540-545, 6 s. ISBN 978-989-758-172-4. https://is.muni.cz/publication/1322854

D - Valentina Apresjan, Vít Baisa, Olga Buivolova, Olga Kultepina. RuSkELL: Online Language Learning Tool for Russian Language. In Tinatin Margalitadze, George Meladze. Proceedings of the XVII EURALEX International congress. Tbilisi: Ivane Javakhishvili Tbilisi State University, 2016. s. 292-299, 8 s. ISBN 978-9941-13-542-2 https://is.muni.cz/publication/1352900

D - Vít Baisa, Silvie Cinkova, Ema Krejčová, Anna Vernerová. VPS-GradeUp: Graded Decisions on Usage Patterns. In Nicoletta Calzolari (Conference Chair) and Khalid Choukri and Thierry Declerck and Marko Grobelnik and Bente Maegaard and Joseph Mariani and Asuncion Moreno and Jan Odijk and Stelios

Piperidis. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia: European Language Resources Association (ELRA), 2016. s. 823-827, 5 s. ISBN 978-2-9517408-9-1. https://is.muni.cz/publication/1347072

R - Vít Suchomel, Pavel Rychlý. Set of Ethiopian Web Corpora (software). 2016. https://is.muni.cz/publication/1381970

D - GAMBÄCK, Björn and Utpal Kumar SIKDAR. Named Entity Recognition for Amharic Using Deep Learning. In Paul Cunningham and Miriam Cunningham (Eds). IST-Africa 2017 Conference Proceedings. IIMC International Information Management Corporation, Windhoek, Namibia, June 2017. ISBN: 978-1-905824-56-4.

D - GAMBÄCK, Björn and Utpal Kumar SIKDAR. Using Convolutional Neural Networks to Classify Hate-Speech. To appear in the 1st Workshop on Abusive Language Online to be held at the 55th Annual Meeting of the Association of Computational Linguistics, Vancouver, Canada, August 2017.

D - MAHESHWARI, Tushar, Aishwarya N. REGANTI, Samiksha GUPTA, Anupam JAMATIA, Upendra KUMAR, Björn GAMBÄCK and Amitava DAS. A Societal Sentiment Analysis: Predicting the Values and Ethics of Individuals by Analysing Social Media Content. In Mirella Lapata, Phil Blunsom and Alexander Koller (eds.): Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 731–741, Valencia, Spain, April, 2017. ISBN 978-1-945626-34-0. ACL.

D - RÆDER, Johan G. Cyrus M. and Björn GAMBÄCK. Sarcasm Annotation and Detection in Tweets. In Alexander Gelbukh (ed.): The 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2017), Budapest, Hungary, May 2017, Springer Lecture Notes in Computer Science.

D - SIKDAR, Utpal Kumar and Björn GAMBÄCK. Named Entity Recognition for Amharic Using Stack-Based Deep Learning. In Alexander Gelbukh (ed.): The 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2017), Budapest, Hungary, May 2017, Springer Lecture Notes in Computer Science.

D - STEINSKOG, Asbjørn Ottesen, Jonas Foyn THERKELSEN and Björn GAMBÄCK. Twitter Topic Modeling by Tweet Aggregation. In Jörg Tiedemann (ed.): Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa), pp. 77-86, Göteborg, Sweden, May 2017. ISBN 978-91-7685-601-7. NEALT.

R - Pavel Rychlý. Corpus Annotation Tool (software). 2017. https://is.muni.cz/publication/1381994

R - Karel Pala, Aleš Horák, Pavel Rychlý, Vít Suchomel, Vít Baisa, Miloš Jakubíček, Vojtěch Kovář, Zuzana Nevěřilová, Adam Rambousek, Björn Gambäck, Utpal Sikdar, Lars Bungum. HaBiT system (software). 2017. https://is.muni.cz/publication/1381969

2.5.3 Public events dedicated to the project (organised or/and participated)

Type	Title of event	Date	Place
Launch:	Kick-Off Meeting	20-21.11. 2014	Brno
Interim:	Information Event for Students of FI and FA faculties	22.10.2015	Brno
	Recent Advances in Slavonic Natural Language Processing, RASLAN 2015	04-06.12. 2015	Karlova Studánka
	Community-based Building of Language Resources Workshop, CBBLR 2016	12.09.2016	Brno
	Recent Advances in Slavonic Natural Language Processing, RASLAN 2016	02-04.12. 2016	Karlova Studánka
	HaBiT Evaluation Workshop 2017	16-21.02. 2017	Brno
Closing:	HaBiT Closing Workshop 2017	26-28.04. 2017	Oslo

Other:	Project meeting in Oslo	05-06.09. 2015	Oslo
	Project meeting in Brno	12.09.2016	Brno

Note: In compliance with annex 4 of the Regulation on implementation of Norwegian Financial Mechanism 2009-2014.

2.6 Planned/further cooperation

Will cooperation with the donor project partner(s) continue after the project is completed? Please tap twice to choose one option below.			
☐ YES – a formal agreement	☒ YES – planned	☐ MAYBE – no plans yet	☐ NO
Total number of joint applications for further collaboration			**0**
Total number of long-term cooperation (new projects) resulting from the partnership			**0**

2.6.1 Further projects

Title of new project	Partners	Programme	Prep*/Submitted /Funded	Grant amount

*Note: *Prep means Preparation.*

2.7 Risk management (max. ½ pages A4)

In the course of the HaBit project there were two risks successfully mitigated.
The first risk arose with the delayed financing at the beginning of the project, which forced shortening the timeplan from 35 to 31 months. This risk was solved by hiring two more researchers, which was possible thanks to unreduced project budget, and redesigning the planned milestones.
The second risk was related to the planned organization of an evaluation workshop in Addis Ababa in February 2017, devoted to expert annotation of the obtained corpora. However, during the final workshop preparation, there was an emergency state declared in Ethiopia and after consulting the Czech Ministry of Education, the workshop was moved to Czech Republic with inviting Ethiopian linguistic experts to participate. 8 researchers from the University of Addis Ababa and the Jijiga University, Ethiopia participated at the workshop - all included languages and their corpora were successfully annotated and the positive outcomes have been achieved.

3. FINANCIAL PART

3.1 Actually used Indirect Costs Model (Choose one option per entity only)

Entity	Used PIC rate in %	Full cost - used analytical accounting system (Yes/No)	Used flat rate in % (60/20)
MU	-	-	**60**
NTNU	-	-	**60**

Note: According to 5.4 of Annex 12 – Rules for the establishment and implementation of donor partnership programmes falling under the Programme Areas "Research within Priority sectors" and "Bilateral Research Cooperation" of the Regulation on implementation of Norwegian Financial Mechanism 2009-2014.

3.2 Explanation of grant use (max. 2 pages A4)

The requested funding corresponded in case of both the Project Promoter and the Project Partner to 100% of the planned budget as they are public research organizations (universities) and the proposed project did not include commercial activities.

The overall budget justification is divided into the part describing the Masaryk University (MU, Project Promoter) budget and the Norges teknisk-naturvitenskapelige universitet (NTNU, Project Partner) budget.

MU Budget

The budget for Masaryk University as presented in the Annex I of this report consists of the following parts (in terms according to the Eligible cost structure):

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

doc.PhDr. Karel Pala, CSc., associate professor - Principal Investigator, project team leader

doc.RNDr. Aleš Horák, Ph.D., associate professor - design and analysis of annotation techniques and structures

doc.Mgr. Pavel Rychlý, Ph.D., associate professor - design and analysis of effective processing of large textual data

RNDr. Vít Suchomel, researcher - implementation of tools for obtaining language data from Internet

Mgr. Vít Baisa, Ph.D., researcher - design and implementation of specialized user interfaces

RNDr. Miloš Jakubíček, researcher - creation of program interfaces for big data processing

RNDr. Vojtěch Kovář, Ph.D., researcher - design and implementation of automatic annotation techniques

RNDr. Adam Rambousek, Ph.D., researcher - design and implementation of efficient processing and presentation of the given language lexicon

RNDr. Zuzana Nevěřilová, Ph.D., researcher - analysis of collocational characteristics of the given languages

Mgr. Ondřej Herman, researcher - preparation of data and tools for annotations

Mgr. Marek Medveď, researcher - preparation of data and tools for annotations

Bc. Marie Stará/Mgr. Lucia Kocincová, project administrator – dissemination manager, administrative processing and checking of project clerical papers

Some of the personnel costs were paid in the form of part-time job agreements according to the internal rules of MU.

Travel and subsistence. The travel and subsistence costs included: active participation of MU research team members at scientific international conferences designated to the computer processing of natural languages, such as NoDaLiDa, TSD, LREC, Translating and the Computer LTC, or COLING. This cost included also the respective conference fees. The travel costs also covered visits of research team members to the Project Partner institution.

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 no new equipment was planned for the project.

Consumables and supplies. The only consumables for the computer processing project work included special server disk storage for the project data. The number of disks increased due to the need of safe storage of extra data in RAID and to be able to keep the project resources running after the end of the project. The amount for this extra storage did not change the overall project expenses and was within the limit of 60.000 CZK per year.

Other costs. No other costs were originally planned in the project. Dissemination and publication costs were included in the travel costs stated above. Subcontracting was not planned in the MU budget. Since in 2016, the project was enhanced with the Additional Activity related to the implementation and evaluation of the Efficient Annotation Framework, MU started to be obliged to organize an audit. The expenditures for audit were then replanned as other/subcontracting costs.

Indirect cost (overheads). According to the MU statutory rules, the MU institutional overhead costs were computed as a flat rate of 60% of the total direct eligible costs (no subcontracting and no costs of third party resources were planned, the audit expenditures are excluded from the flat rate).

The overall amount of MU expenditures was 10 401 969,47 CZK of 10 611 000 CZK received payments. The remaining amount of 209 030,53 CZK is going to be returned to the Project Promoter. The amount consists of 64 710,62 CZK personnel costs, 57 230,56 CZK travel costs and 87 089,35 CZK indirect costs. This amount corresponds to 2% of the planned expenditures, which was caused by savings in the personnel and travel expenses.

Changes at MU:
All changes were either in the limit of 60,000 CZK/year or were approved by the Project Operator. The changes did not increase or decrease the overall project costs. The changes included:
in 2014:
The only budget changes were related to the delayed beginning of the project, i.e., the project expenditures correspond to a 3 months period instead of 7 months as planned.
20,000 CZK were moved from A1.2 Agreements to A1.1 Salaries due to internal regulations for part-time jobs.
1400 CZK were moved from A2 Travel to A4 Consumables to cover the server disk storage purchase.
in 2015:
the bank charges for the payments to the Project Partner have been declared as ineligible and thus moved out of the project budget
8145,04 CZK were moved from A2.Travel to A4.Consumables to cover the server disk storage purchase.
8406 CZK were moved from A1.1 Salaries to A1.2 Agreements due to internal regulations for part-time jobs.
in 2016:
The amount of 65 000 CZK in A5.4 Other costs was moved from A2 Travel costs in 2016 due to the expected audit expenditures. This change was approved in MSMT-24209/2016-38.
137 000 CZK was moved from A1.1 Salaries to A1.2 Agreements due to internal regulation for part-time jobs. This change was within the A1 Personnel costs and was approved by e-mail.
42445,60 CZK were moved from A5.4 Other do A4 Consumables and used for server disk storage.
in 2017:
26876,49 CZK were moved from A2 Travel do A4 Consumables and used for server disk storage to allow storing and backup of the final project data.

4. MANDATORY ANNEXES

4.1 Annexes related to Final Project Report

No.	Annexes	Mandatory online submission format
I.	***Annex I – Final Project Financial Report*** This annex relates to expenditure actually incurred in CZK **by all entities in the whole implementation period of the project**. It must be stamped and signed by the Principal Investigator and by the statutory of the Project Promoter, or an attorney. Please use the template.	xls(x) and pdf
II.	***Annex II - Report on Actual Incurred Expenditure*** This annex relates **to all Czech entities** only. It is a record from an accounting system which reports costs in CZK per entity from 1 January 2017 to 30 April 2017. It contains a stamp of the organisation and a full name and signature of the person responsible for financial matters of the organization. No template.	pdf

III.	***Annex III – Final Financial Statement by Norwegian Project Partner*** This annex relates **to Norwegian project partners only** and is filled in in NOK for **the whole implementation period.** It must contain a stamp of the organisation and a full name and signature of the person responsible for financial matters. It may be a copy. Use the template please.	pdf
IV.	***Annex IV - Letter of Attorney***, if applicable for the Project Promoter. Acceptable in Czech. A copy may be submitted. No template.	pdf

Note: For e-submission the required format for the Final Project Report is doc(x) (you may also submit the undersigned pages in PDF as a separate file, if no e-signature). In case of revisions, please, always indicate the corrected version of the report.
Please staple the documents in the following order: The Final Project Report and annexes I, II, III, IV. Use Calibri font, size 11.