PROJECT PERIODIC REPORT	
Czech-Norwegian Research Programme (CZ09)	
Norwegian Financial Mechanism 2009-2014	
Programme area	**Bilateral Research Cooperation**
Periodic report	**3/2016-17**
Period covered	**1 January 2016 to 30 April 2017**
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	dd/mm/yyyy

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 Activity of research and development in project

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

Project Promoter: Masarykova univerzita (MU)

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

Project partner: Norges teknisk-naturvitenskapelige universitet (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)
25,593,000

1.4.1 Total grant in Project contract (in CZK)
25,593,000

1.4.1.1 Grant for reporting period approved by Programme Operator (in CZK)
11,448,000

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 were:

- to gather large-scale text data (corpora) for under-resourced languages from the Web and process them so they can be used in language applications, such as information extraction or machine translation; these languages include Norwegian, Czech and four major Ethiopian languages: Amharic, Afaan Oromo, Tigrinya and Somali;
- to build a multi-billion word Norwegian corpus using the tools co-developed by Masaryk University and utilized in a joint EU-funded project with NTNU ("PRESEMT: Pattern REcognition-based Statistically Enhanced MT", 2010-2012);
- to feed into and leverage on the NORHED project ("Linguistic Capacity Building – tools for the inclusive development of Ethiopia", NORHED 2013-2018, collaboration of NTNU, University of Oslo and two Ethiopian universities) and include the processing of the Ethiopian languages mentioned above into the HaBiT project;
- to build shallow processing applications for Czech, Norwegian and Amharic for investigating and separating multiple senses of the words in the corpora, as well as for creating multi-sense vector spaces and parallel multi-lingual vector spaces for word translation disambiguation.

The aims above were accomplished in delivering the final HaBiT system and addressing the 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. 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)

Hlavní cíle projektu HaBiT byly:

- shromáždit velká textová data (textové korpusy) pro jazyky s nedostatečnými zdroji z webu a zpracovat je tak, aby mohla být využita v jazykových aplikacích jako např. extrakce informací nebo strojový překlad. Mezi tyto jazyky patří norština, čeština a čtyři hlavní etiopské jazyky: amharština, afaan oromština, tigrinština a somálština;
- vytvořit multibilionový norský korpus s použitím nástrojů spoluvyvinutých FI MU v rámci společného evropského projektu s NTNU ("PRESEMT: Pattern REcognition-based Statistically Enhanced MT", 2010-2012);
- navázat na projekt NORHED ("Linguistic Capacity Building – tools for the inclusive development of Ethiopia", NORHED 2013-2018, spolupráce NTNU, University of Oslo a dvou etiopských univerzit) a zahrnout zpracování zmíněných etiopských jazyků do projektu HaBiT;
- vytvořit mělké aplikace pro zpracování češtiny, norštiny a amharštiny umožňující zkoumat a vyčlenit vícenásobné významy slov v korpusech a také vytvářet vícevýznamové vektorové prostory a paralelní multilinguální vektorové prostory umožňující zjednoznačnění při překladu slov.

Cíle uvedené výše byly dosaženy vytvořením finálního systému HaBit a naplněním témat evaluace, verifikace a testování technologie a informační technologie (ICT) naplňující společenské výzvy. Díky tomu byla získána i relevantní přidaná hodnota v politickém ohledu

prostřednictvím spolupráce s méně rozvinutou zemí (Etiopií). Výsledky dosažené v projektu jsou dále prezentovány v periodické zprávě. Aktuální informace je k dispozici na webové stránce projektu www.habit-project.eu.

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

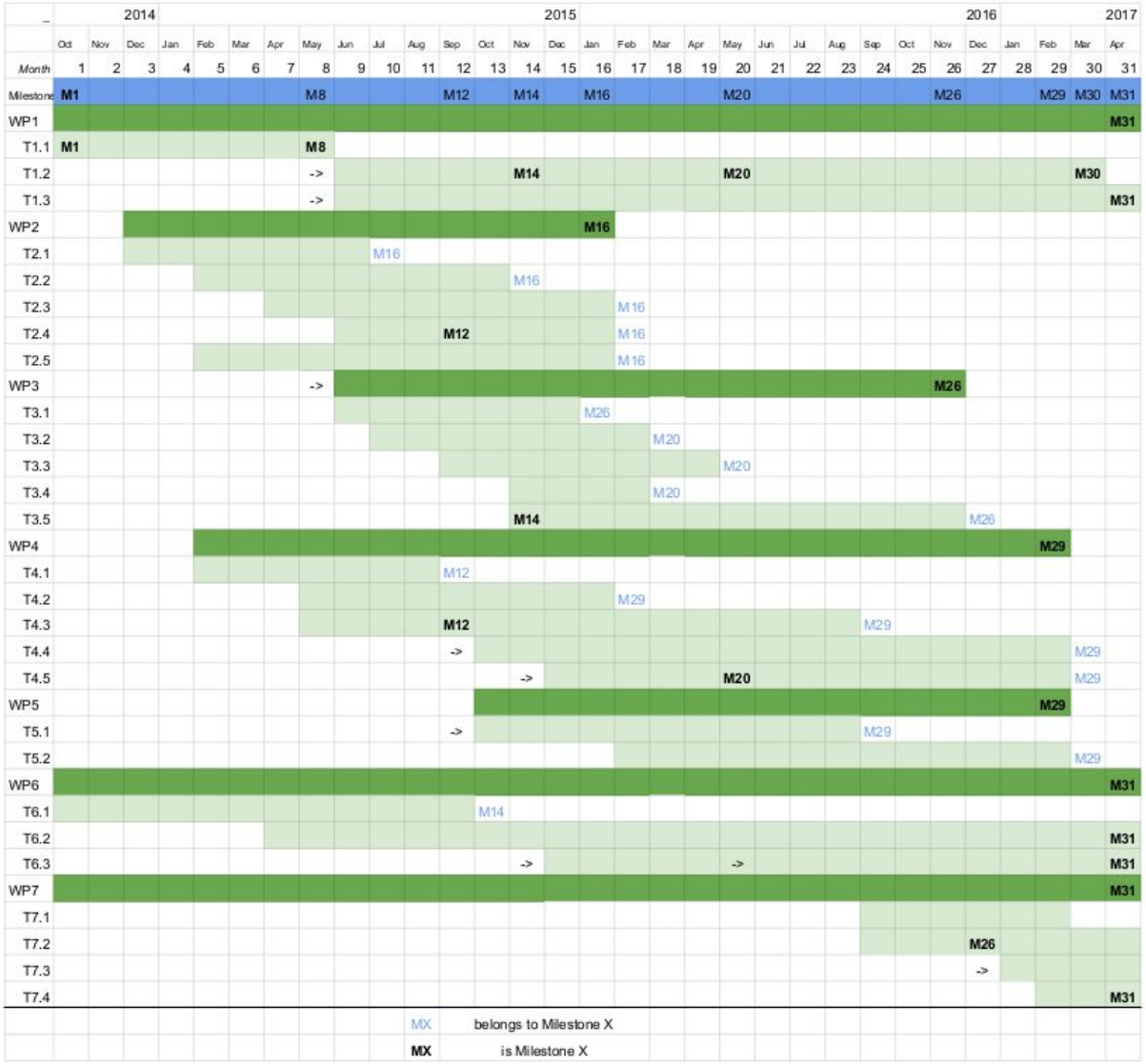
In the reported period the main objectives fall into the following work packages and respective tasks:

- 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
- WP7: Efficient annotation framework

From these, the work packages WP1-WP6 were originally planned from the project start, while WP7 was added in 2016 as a part of the approved Additional Activity.
The interim results of the HaBiT project are presented in the deliverables mentioned in section 2.3.3, as planned for the reported years 2016 and 2017.

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 below. The main efforts were devoted to morphological annotation of the new corpora (Amharic, Afaan Oromo, Tigrinya, Somali, Czech and Norwegian), creation (or improvement) of the sketch grammars for given languages and the completion of the HaBiT system. The achievements in these tasks are detailed further in this section, split to individual work packages.



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

Work package 1

Based on the system and corpora specification (D1.1.1-4) the second and third system prototype was created (D1.2.2 and D1.2.3, respectively). During the work the system specifications (D1.1.1) were adjusted leading to integration of the modules developed in WP4, WP5 and WP6 and to creation of the final HaBiT system including the respective system demonstrator (D1.3).

Work package 2

Data gathered in D2.1 were morphologically annotated and a new Czech corpus of almost 8 billion words and parallel Czech-Norwegian corpus of 25 million words were created and incorporated into the HaBiT system.

Work package 3

New versions of the four Ethiopian language corpora were created in D3.1-4a. All these corpora were then morphologically annotated to allow further linguistic analysis (D3.1-4b).

Work package 4
New methodology for Sketch Grammar development was proposed and tested (D4.2) with the help of Sketch Grammar visualization methods (D4.3). A new Sketch Grammar was created for Norwegian and the Sketch Grammar for Czech was improved (D4.4a,b). New definitions of word sketches were created for Amharic, Afaan Oromo, Tigrinya and Somali (D4.4c, d). These instances were evaluated and led to a new language Sketch Grammar methodology, which allows to generate an initial Sketch Grammar for a new language based on a set of selected language features (D4.5).

Work package 5
In line with the latest developments in language technology, the word space models were implemented in the form of word embeddings to be fed to a deep learner, that is, a neural network architecture. Experiments were carried out with both evolutionary algorithms and different network setups, and combinations of more traditional machine learners with deep learners. The system was in particular applied to the task of identifying and classifying named entities in Amharic text, as reported in D5.2. The different versions of the system architecture are described in the joint deliverable D5.1, D5.3 & D5.4.

Work package 6
The coverage of the created corpora for Norwegian (Bokmål and Nynorsk), Amharic, Somali, Oromo and Tigrinya was evaluated, as well as the ease of use of the system modules, and reported in the joint usability and system evaluation deliverable (D6.2 & D6.3).

Work package 7
The system for manual annotation of corpora was designed and implemented (D7.1), tested and the usability for the Ethiopian languages was evaluated. The framework and evaluation results were published (D7.2).

Please present a detailed review of work, divided into work packages for the reporting period. The number of used work packages should be appropriate to the complexity of the project.

2.3.2 Milestones achievement (cumulative)

No.	Milestone title	WP no.	Lead partner (abbreviation)	Planned achievement date dd/mm/yyyy	Actual/Forecast achievement date dd/mm/yyyy
1	M1: Kick-Off Meeting	WP1	MU	31/10/2014	21/11/2014
2	M8: System specification	WP1	MU	30/11/2014	31/05/2015
3	M12: The first version of Norwegian corpus	WP2	MU	30/09/2015	30/09/2015
4	M12: Methodology of Sketch Grammar evaluation	WP4	MU	31/05/2015	30/09/2015
5	M14: HaBiT system v1	WP1	MU	30/11/2015	30/11/2015
6	M14: Parallel Corpora from Web data	WP3	MU	30/11/2015	30/11/2015

7	M20: HaBiT system v2	WP1	MU	31/05/2016	31/05/2016
8	M20: Semantic search interface v1	WP5	NTNU	31/05/2016	31/05/2016
9	M26: Semantic search interface v2	WP5	NTNU	31/12/2016	31/12/2016
10	M26: Implementation of testing version of annotating tool and providing the feedback	WP7	MU	31/12/2016	31/12/2016
11	M29: New and/or improved Word Sketeches for given languages	WP4	MU	18/02/2017	18/02/2017
12	M29: Semantic search interface v3	WP5	NTNU	18/02/2017	18/02/2017
13	M30: HaBiT system v3	WP1	MU	28/02/2017	31/03/2017
14	M31: Final version of the HaBiT system	WP1 WP6	MU	30/04/2017	30/04/2017
15	M31: Final evaluation of crowd-sourced annotation of selected Ethiopian languages	WP7	MU	30/04/2017	30/04/2017

Note: This table is cumulative. It should show all milestones in the whole project period. Reporting and publicity actions are not milestones.

2.3.3 Deliverables achievement (cumulative)

No.	Deliverable title	WP no.	Lead partner (abbreviation)	Planned delivery date dd/mm/yyyy	Actual/Forecast delivery date dd/mm/yyyy
1	D1.1.1 System specifications	WP1	MU	31/03/2015	25/05/2015
2	D1.1.2 Specification of corpora and the corpus building module	WP1	MU	31/03/2015	28/02/2015
3	D1.1.3 Specification of word-sketch grammars and tools	WP1	MU	31/03/2015	28/02/2015
4	D1.1.4 Specification of the semantic content matching and wordspace module	WP1	MU	31/03/2015	31/03/2015
5	D1.2.1 The HaBiT system v1	WP1	MU	31/03/2015	30/11/2015
6	D1.2.2 The HaBiT system v2: Second integrated system prototype	WP1	MU	30/04/2016	23/05/2016
7	D1.2.3 The HaBiT system v3: Third and	WP1	MU	28/02/2017	27/03/2017

	pre-final integrated system prototype				
8	D1.3 The final HaBiT system: Tested and evaluated system demonstrator	WP1	MU	30/04/2017	30/04/2017
9	D2.1: An improvement of web crawler SpiderLing	WP2	MU	31/07/2015	25/01/2016
10	D2.2: A Norwegian corpus, sized 5 billion words	WP2	MU	31/10/2015	25/01/2016
11	D2.3: A new Czech corpus, sized 8 billion words	WP2	MU	31/01/2016	25/01/2016
12	D2.4: Parallel Czech-Norwegian corpus, size 10 million tokens	WP2	MU	31/01/2016	25/01/2016
13	D3.1a: An Amharic corpus, sized 20 million words	WP3	MU	31/08/2015	31/01/2016
14	D3.2a: An Afaan Oromo corpus, sized 3 million words	WP3	MU	31/10/2015	30/05/2016
15	D3.3a: A Tigrinya corpus, sized 3 million words	WP3	MU	31/12/2015	30/05/2016
16	D3.4a: A Somali corpus, sized 10 million words	WP3	MU	29/02/2016	30/05/2016
17	D3.1b: An Amharic corpus, sized 20 million words	WP3	MU	30/11/2015	28/11/2016
18	D3.2b: An Afaan Oromo corpus, sized 3 million words	WP3	MU	30/04/2016	28/11/2016
19	D3.3b: A Tigrinya corpus, sized 3 million words	WP3	MU	30/07/2016	28/11/2016
20	D3.4b: A Somali corpus, sized 10 million words	WP3	MU	31/11/2016	28/11/2016
21	D4.1 Methodology of Sketch Grammar evaluation	WP4	MU	30/06/2015	30/09/2015

22	D4.2: Sketch Grammar development module	WP4	MU	31/01/2016	20/02/2017
23	D4.3: Visualization tool for Sketch Grammar queries	WP4	MU	31/08/2016	20/02/2017
24	D4.4a: An improved definition of Word Sketches for Czech	WP4	MU	30/04/2016	20/02/2017
25	D4.4b: A new definition of Word Sketches for Norwegian	WP4	MU	30/11/2015	20/02/2017
26	D4.4c: A new definition of Word Sketches for Amharic	WP4	MU	30/04/2016	20/02/2017
27	D4.4d: A new definition of Word Sketches for Afaan Oromo, Tigrinya, and Somali	WP4	MU	28/02/2017	20/02/2017
28	D4.5: New language Sketch Grammar module	WP4	MU	28/02/2017	20/02/2017
29	D5.1 Semantic search interface v1	WP5	NTNU	31/07/2015	23/05/2016
30	D5.2 Dynamic concept matching	WP5	NTNU	30/04/2016	20/02/2017
31	D5.3 Semantic search interface v2	WP5	NTNU	30/11/2016	20/02/2017
32	D5.4 Semantic search interface v3	WP5	NTNU	28/02/2017	10/04/2017
33	D6.1 Project evaluation plan	WP6	MU	31/07/2015	31/01/2016
34	D6.2 Final Usability report	WP6	MU	30/04/2017	10/04/2017
35	D6.3 System evaluation	WP6	MU	30/04/2017	10/04/2017
36	D7.1 Prototype implementation of first version of the annotation framework	WP7	MU	31/12/2016	31/12/2016
37	D7.2 Publication of the framework and methodology evaluation results	WP7	MU	30/04/2016	30/04/2016

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

Note: Please present the work packages in detail for the reporting period, using the table provided below. If you have more WPs please copy the section 2.4.1.1-2.4.1.10.

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
WP7	Efficient annotation framework	09/2016	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

01/10/2014

2.4.1.5 WP end date

30/04/2017

2.4.1.6 WP objective

To integrate and test produced software modules, to keep the demo website updated, to create overall integrated system.

Partners involved: all, both teams dealt with the process of integration and testing of the developed modules.

2.4.1.7 WP task

Task T1.2: Module integration and testing and Task T1.3: Coordination of the platform development were in progress during preparation of the second and third HaBiT system prototype and the final HaBiT system.

Partners involved: all

2.4.1.8 WP deliverable
D1.2.2 The HaBiT system v2 [M20]: Second integrated system prototype
D1.2.3 The HaBiT system v3 [M29]: Third and pre-final integrated system prototype
D1.3 The final HaBiT system [M31]: Tested and evaluated system demonstrator

2.4.1.9 WP milestone
MS5 HaBiT system v2 [M20]: Second system prototype, including complete versions of the different software modules
MS6 HaBiT system v3 [M30]: Third, pre-final system prototype, ready for usability studies and final system evaluation
MS7 Project end [M31]: Release of the final system

2.4.1.10 WP Human resources
Qualification level: 9 - Karel Pala (docent), Aleš Horák (docent), Pavel Rychlý (docent), Miloš Jakubíček (RNDr., PhD student), Adam Rambousek (PhD, postdoc), Björn Gambäck (professor), Janne Bondi Johannessen (professor), Lars Bungum (researcher), Utpal Kumar Sikdar (researcher)
Person-months: 4,2

2.4.1.1 WP number
WP2

2.4.1.2 WP title
Multi-billion word corpus building

2.4.1.3 WP leader
Pavel Rychlý

2.4.1.4 WP start date
01/12/2014

2.4.1.5 WP end date
31/01/2016

2.4.1.6 WP objective
To finish morphological annotation and indexing of corpora.

Partners involved: all, both teams developed the corpora.

2.4.1.7 WP task
Task T2.3: The gathered data have been 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 have been 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 have been processed as well.

Partners involved: all

2.4.1.8 WP deliverable
D2.3: A new Czech corpus, sized 8 billion words, cleaned (without boilerplate, de-duplicated), morphologically annotated, indexed for fast searching. [M16]
D2.4: Parallel Czech-Norwegian corpus, size 32 million tokens [M16]

2.4.1.9 WP milestone
No milestones in the reported period.

2.4.1.10 WP Human resources
Qualification level: 10 - Pavel Rychlý (docent), Vít Suchomel (Mgr., PhD student), Vít Baisa (Mgr., PhD student), Miloš Jakubíček (RNDr., PhD student), Adam Rambousek (PhD, postdoc), Björn Gambäck (professor)
Person-months: 1

2.4.1.1 WP number
WP3

2.4.1.2 WP title
Corpora for under-resourced languages

2.4.1.3 WP leader
Pavel Rychlý

2.4.1.4 WP start date
01/01/2015

2.4.1.5 WP end date
30/11/2016

2.4.1.6 WP objective
To mine the Internet to gather multi million word corpora in Amharic, Afaan Oromo, Tigrinya and Somali. A web crawler SpiderLing has been used. The corpora have been morphologically annotated and indexed for fast searching.

Partners involved: all, both teams developed the corpora.

2.4.1.7 WP task
Task T3.3: Since presence of texts in these languages on the Internet is very scarce, an approach requiring an extra effort has been needed. Methods proposed by the Brno team and successfully used in the past for the Tajik language (Dovudov, 2012) have been followed: Web pages yielding a lot of texts in these languages are identified in data gathered in previous steps. Custom scripts for harvesting specific web domains were created. Less data

were discarded by the boilerplate cleaning tool in this way and thus a larger corpus size (compared to standard fully automated crawling) has been obtained.
Task T3.4: Data gathered in tasks T3.2 and T3.3 have been merged together, de-duplicated using tool Onion, tokenized, split to sentences and transformed into the XML format with separate structures for document, paragraph, and sentence. The metadata (language, source, date, size, headings) were stored as XML attributes of respective structures.
Task T3.5: The processed corpus data has been morphologically annotated to allow further linguistic analysis.

Partners involved: all

2.4.1.8 WP deliverable
D3.3a: A Tigrinya corpus, sized 3 million words, cleaned (without boilerplate, de-duplicated), indexed for fast searching. [M20]
D3.4a: A Somali corpus, sized 10 million words, cleaned (without boilerplate, de-duplicated), indexed for fast searching. [M20]
D3.2b: An Afaan Oromo corpus, sized 3 million words, cleaned (without boilerplate, de-duplicated), morphologically annotated, indexed for fast searching. [M26]
D3.3b: A Tigrinya corpus, sized 3 million words, cleaned (without boilerplate, de-duplicated), morphologically annotated, indexed for fast searching. [M26]
D3.4b: A Somali corpus, sized 10 million words, cleaned (without boilerplate, de-duplicated), morphologically annotated, indexed for fast searching. [M26]

2.4.1.9 WP milestone
The milestone for this WP was passed in year 2015. (MS4: delivery of D3.1b, an annotated corpus for Amharic. [M14])

2.4.1.10 WP Human resources
Qualification level: 11 - 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), Janne Bondi Johannessen (professor)
Person-months: 15

2.4.1.1 WP number
WP4

2.4.1.2 WP title
Shallow processing grammars and tools

2.4.1.3 WP leader
Aleš Horák

2.4.1.4 WP start date
01/10/2014

2.4.1.5 WP end date
28/02/2017

2.4.1.6 WP objective

A Word Sketch 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 basic Sketch Grammar for a new language and tools and methods for enhancing, evaluating and debugging existing Sketch Grammars.
A Word Sketch grammar (language specific definition of semantic relations) has been devised for each language in the project. The morphological tagging applied in WPs 2 and 3 has been exploited. The aim of Word Sketch grammars is to enable to study grammatical and collocational behaviour of words in further research.

Partners involved: all, both teams dealt with the sketch grammars development.

2.4.1.7 WP task

Task T4.3: Visualization of Sketch Grammar queries. A corpus query is the basic component of a sketch grammar, this task provides an easy and interactive tool for query visualization.
Task T4.4: New or updated Sketch Grammars for Czech, Norwegian and 4 new languages.
Task T4.5: Sketch Grammar template for a new language. A Sketch engine module.

Partners involved: all, both teams dealt with the sketch grammars development.

2.4.1.8 WP deliverable

D4.2: Sketch Grammar development module for listing Sketch Grammar statistics of Sketch Grammar definition and Word Sketches for selected corpus, and listing differences in two versions of Sketch Grammar or Sketch Grammars for two different languages. [M16]
D4.3: Visualization tool for Sketch Grammar queries. [M23]
D4.4a: An improved definition of Word Sketches for Czech. [M20]
D4.4b: A new definition of Word Sketches (grammatical and semantic relations) for Norwegian. [M20]
D4.4c: A new definition of Word Sketches (grammatical and semantic relations) for Amharic. [M20]
D4.4d: A new definition of Word Sketches (grammatical and semantic relations) for Afaan Oromo, Tigrinya, and Somali. [M29]
D4.5: New language Sketch Grammar module generating initial Sketch grammar for a new language from selected language features. [M29]

2.4.1.9 WP milestone

MS3 delivery of D4.4b, a new definition of Word Sketches for Norwegian. [M20]

2.4.1.10 WP Human resources

Qualification level: 11 - 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), Marek Medveď (Mgr., PhD student), Ondřej Herman (Mgr., PhD student), Björn Gambäck (professor)
Person-months: 18,3

2.4.1.1 WP number
WP5

2.4.1.2 WP title
Multi-sense and multi-lingual word spaces

2.4.1.3 WP leader
Björn Gambäck

2.4.1.4 WP start date
01/06/2015

2.4.1.5 WP end date
28/02/2017

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

Partners involved: all, both teams dealt with the model development.

2.4.1.7 WP task
Task T5.1: Semantic search and disambiguation: Development of semantic search models and development of matching logic for mono- and multi-lingual word space models.
Task T5.2: Word space modeling: Development of large word space models to be integrated into the semantic search; similarity measures and thresholds for multi-sense random indexing.

Partners involved: all

2.4.1.8 WP deliverable
D5.2 Dynamic concept matching [M20]: Report on the information matching logic
D5.3 Semantic search interface v2 [M26]: Integrating multi-sense and multi-lingual matching
D5.4 Semantic search interface v3 [M29]: Tested and evaluated search interface

2.4.1.9 WP milestone
M20: Semantic search interface v1
M26: Semantic search interface v2
M29: Semantic search interface v3

2.4.1.10 WP Human resources
Qualification level: 9 - Karel Pala (docent), Aleš Horák (docent), Pavel Rychlý (docent), Björn Gambäck (professor), Lars Bungum (researcher), Utpal Kumar Sikdar (PhD, researcher)
Person-months: 30.7

2.4.1.1 WP number
WP6

2.4.1.2 WP title
Requirements and evaluation

2.4.1.3 WP leader
Karel Pala

2.4.1.4 WP start date
01/10/2014

2.4.1.5 WP end date
30/04/2017

2.4.1.6 WP objective
To ensure that the system meets user demands, to ensure the usability of the system and monitor usage, to ensure availability of fully-illustrated support documentation for the use of the system and to define and assess the overall project success criteria.

Partners involved: all, both teams dealt with the usability studies and evaluation of the system.

2.4.1.7 WP task
Task T6.2: **Usability studies:** Quantitative user testing: A sample of test persons provides an early indication of whether the design meets end-user requirements; Qualitative usability study: Longitudinal study of the system usage.
Task T6.3: **System evaluation:** Integrability tests: Evaluate ease of system integration at external sites; System evaluation: Performance evaluation of the complete integrated system prototype.

Partners involved: all

2.4.1.8 WP deliverable
D6.2 Final Usability report [M31]: Report on the overall usability of the system
D6.3 System evaluation [M31]: Report on overall system performance

2.4.1.9 WP milestone
MS7 delivery of D6.2 and D6.3, final version of the system [M31]

2.4.1.10 WP Human resources
Qualification level: 11 - Karel Pala (docent), Aleš Horák (docent), Pavel Rychlý (docent), Vojtěch Kovář (PhD, postdoc), Adam Rambousek (PhD, postdoc), Vít Suchomel (Mgr., PhD student), Vít Baisa (Mgr., PhD student), Miloš Jakubíček (RNDr., PhD student), Marek Medveď (Mgr., PhD student), Ondřej Herman (Mgr., PhD student), Björn Gambäck (professor), Janne Bondi Johannessen (professor), Lars Bungum (researcher), Utpal Kumar Sikdar (PhD, researcher), with evaluation support from the following researchers at AAU: Derib Ado Jekale (PhD), Feda Negesse Beyene (PhD), Endalew Assefa Temesgen (PhD), Shimelis Mazengia Beyene (PhD), Girma Mengistu Desta (PhD), Tsegay Weldemariam Beyene, Tadesse Woldegebreal Baymot, Ahmed Yusuf Hirad.

Person-months: 8.5

2.4.1.1 WP number
WP7

2.4.1.2 WP title
Efficient annotation framework

2.4.1.3 WP leader
Pavel Rychlý

2.4.1.4 WP start date
01/09/2016de

2.4.1.5 WP end date
30/04/2017

2.4.1.6 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 has been developed on well-known languages and evaluated on less-covered languages at the workshop originally planned to take place at the University of Addis Ababa. Because of the emergency state in Ethiopia the workshop was finally moved to Brno and organized at the FI MU.
Partners involved: MU team dealt with the annotation framework development and publication.

2.4.1.7 WP task
To design a crowd-focused system for manual annotation of corpora, to implement the system, including the user interface, to test the annotation system for selected language and to evaluate its usability for Ethiopian languages.

Partners involved: MU

2.4.1.8 WP deliverable
D7.1 [M27] Prototype implementation of first version of the annotation framework
D7.2 [M31] Publication of the framework and methodology evaluation results

2.4.1.9 WP milestone
Implementation of the testing version of annotating tool and providing the feedback [M27].
Final evaluation of crowd-sourced annotation of selected Ethiopian languages [M31]

2.4.1.10 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), Ondřej Herman (Mgr., PhD student).
Person-months: 5,6

2.4.2 Project output(s)

Type of output	Title	Date of accomplishment (mm/yyyy)	Date of realization (mm/yyyy)
D	Matching logic for mono- and multi-lingual word space models	08/2016	05/2017
D	Semantic Search in Large Word Space Models	01/2017	09/2017
D	Multi-sense Random Indexing	03/2017	10/2017
D	Annotated Amharic Corpora	06/2016	09/2016
D	Annotation of Czech Texts with Language Mixing	06/2016	09/2016
D	AQA: Automatic Question Answering System for Czech	06/2016	09/2016
D	Czech Grammar Agreement Dataset for Evaluation of Language Models	10/2016	12/2016
D	DSL Shared task 2016: Perfect Is The Enemy of Good Language Discrimination Through Expectation-Maximization and Chunk-based Language Model	09/2016	12/2016
D	English-French Document Alignment Based on Keywords and Statistical Translation	05/2016	08/2016
D	European Union Language Resources in Sketch Engine	03/2016	05/2016
D	Evaluating Natural Language Processing Tasks with Low Inter-Annotator Agreement: The Case of Corpus Applications	10/2016	12/2016
D	Evaluation and Improvements in Punctuation Detection for Czech	06/2016	09/2016
D	Finding Definitions in Large Corpora with Sketch Engine	03/2016	05/2016
D	Graded and Word-Sense-Disambiguation Decisions in Corpus Pattern Analysis: a Pilot Study	03/2016	05/2016
D	Large Scale Keyword Extraction using a Finite State Backend	10/2016	12/2016
D	Multilingual CPA: Linking Verb Patterns across Languages	10/2015	09/2016
D	On Evaluation of Natural Language Processing Tasks: Is Gold Standard Evaluation Methodology - a Good Solution?	12/2015	02/2016
D	RuSkELL: Online Language Learning Tool for Russian Language	10/2015	09/2016

D	VPS-GradeUp: Graded Decisions on Usage Patterns	03/2016	05/2016
J	Lexicographic Tools to Build New Encyclopaedia of the Czech Language	07/2016	10/2016
J	Sketch Engine for Bilingual Lexicography	03/2016	09/2016
C	Walking the tightrope between linguistics and language engineering	08/2016	2017
R	HaBiT system	04/2017	04/2017
R	PoS-annotation framework	12/2016	02/2017
R	Set of Ethiopian Web Corpora	08/2016	10/2016
W	Workshop at the Masaryk university	02/2017	02/2017

Note: You may add lines. Use the types of outputs listed in RIV registry (in Czech: Rejstřík informací o výsledcích).

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

The project was managed in a standard way. The Project Management Board consisted of the team leaders of each party. The Board monitored the technical content and progress of the project. The process has been going on without any problems. The Board also handled the intellectual property rights following the aim to make project outputs (resources and tools) openly accessible.
The Risk Management Plan and the Quality Assurance Plan were controlled by the project promoter (see 2.6.1). The Technology Management Plan has been followed to ensure the project components are of the intended quality, for detail, see the sections describing project results. The mentioned plans are described in the project documentation. For that reason we do not repeat them here.

Personnel changes that occurred in the Czech and Norwegian teams are described and justified in the relevant section of the report.

The regular project meeting took place in September 12-13, 2016 in Brno, Hotel Continental (during the TSD Conference).
The attendees:

- Masaryk University (MU) - Aleš Horák, Karel Pala, Pavel Rychlý
- NTNU Trondheim - Björn Gambäck, Utpal Kumar Sikdar

The final project meeting took place in April 26-28, 2017 in Oslo, Professorboligen, University of Oslo.
The attendees:

- Masaryk University (MU) - Aleš Horák, Pavel Rychlý, Vít Suchomel, Vít Baisa
- NTNU Trondheim - Björn Gambäck, Utpal Kumar Sikdar, Lars Bungum
- University of Oslo (UiO) - Janne B. Johannessen, Kristin Hagen, Anders Nøklestad, Joel Priestley, Anne Renette Askeland

The additional activity -- Efficient allocation of human resources for linguistic annotation of texts -- has been added. The only deviations from the planned milestones and deliverable dates were caused by the later project start due to delayed financing at the beginning of the project. After the replanning of the milestones, all deliverables were prepared on time.

No changes to the legal status of any of the Project Promoter/partners, in research organisations and SMEs has occurred.
The project website was developed according to the plan. (The web page is referred in section 2.9.1.)

Please use this section to summarise the management of the consortium activities during the reporting period.
Amongst others, this section should include the following:

- *Consortium management tasks and achievements;*
- *Problems which have occurred and how they were (not) solved or envisaged solutions;*
- *Changes in the consortium, if any (changing or adding new partners not allowed);*
- *List of important project meetings, dates and venues;*
- *Project planning and status;*
- *Impact of possible deviations from the planned milestones and deliverables, if any;*
- *Any changes to the legal status of any of the Project Promoter/partners, in research organisations and SMEs;*
- *Development of the project website.*

The section should also provide short comments and information on co-ordination activities during the period in question, such as communication between partners, possible co-operation with other projects.

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	--
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	--
Staff changes in this reporting period					

Mgr. Lucia Kocincová	project administrator	part-time	0.30	01/01/2016	--
Mgr. Lucia Kocincová	project administrator	part-time	0.15	01/03/2016	30/04/2016
Bc. Marie Stará	project administrator	part-time	0.15	01/03/2016	--
Bc. Marie Stará	project administrator	part-time	0.30	01/05/2016	--
Bc. Marie Stará	project administrator	part-time	0.60	01/04/2017	--
Mgr. Marek Medveď	researcher, PhD student	part-time	0.50	01/11/2016	--
Mgr. Ondřej Herman	researcher, PhD student	part-time	0.40	01/11/2016	--
doc.PhDr. Karel Pala, CSc.	project team leader	part-time	0.25	01/11/2016	--
doc.RNDr. Aleš Horák, Ph.D.	researcher	part-time	0.25	01/11/2016	--
Mgr. Pavel Rychlý, Ph.D.	researcher	part-time	0.25	01/11/2016	--
Mgr. Vít Suchomel	researcher, PhD student	part-time	0.50	01/11/2016	--
Mgr. Vít Baisa	researcher, PhD student	part-time	0.20	01/11/2016	--
Mgr. Miloš Jakubíček	researcher, PhD student	part-time	0.20	01/11/2016	--
RNDr. Vojtěch Kovář, Ph.D.	researcher	part-time	0.20	01/11/2016	--

Note: You may add more lines if necessary. Full-time work means he/she works only on the project. All staff changes had to be reported to the Programme Operator during the period. Quit on means that an employee is no longer participating in the project.

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	--
Lars Bungum	researcher	full-time	1.0	01/06/2015	30/04/2017
Staff changes in this reporting period					
Utpal Kumar Sikdar	researcher	part-time	0.80	01/04/2016	30/04/2017

2.5.3 Personnel changes justification (max. 1 page A4)

Personnel changes did not affect the project budget. Lucia Kocincová returned to her work as a project administrator in January 2016. Because of Lucia Kocincová's planned leave Marie Stará was hired in March 2016 and in May she took over the position.
In November 2016 Marek Medveď and Ondřej Herman were hired to design and implement automatic annotation techniques, a task relevant because of additional activities for the project. The workload of the MU research team was increased in November 2016 so the work has been finished in time.
Utpal Kumar Sikdar was employed as the postdoctoral researcher at NTNU according to plan.

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

The technical content and progress was monitored by the Project Management Board, which consisted 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. It is the Boards responsibility to resolve any important issue or problem that arises; the Board also reviews all project deliverables and publications. The day-to-day operations, project management and administrative issues handles the Project Promoter.
The project was audited by external auditors. The MU was audited by INTEREXPERT BOHEMIA with result 1: correct implementation. The audit focused on year 2016. The results for years 2014 and 2015 will be added.
The audit at NTNU was not yet finished. It will be done by the end of June.
In January 2016 an External Interim Examination was held; the output was handed to the Ministry in January 2016. The project report will be evaluated by the independent experts proposed by each of project partners. These experts are Prof. Václav Matoušek from the West Bohemia University in Pilsen on the MU side and Prof. Lars Borin from the University of Gothenburg, Sweden for the NTNU side.

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

The risk management and quality assurance were monitored and evaluated by the Project Management Board in accordance to the Quality Assurance Plan and the Risk Management Plan.
The core of the Quality Assurance Plan is that any issues are entered into the ticket system; each ticket is assigned to a team member. The WP leaders overview the process of implementations of deliverables and control the quality of outputs with help of mentioned ticket system. Every deliverable has a due date, which ensures the lack of delays in the work progress.
The project promoter controls the risk assessment and ensures the process will continue without any major disruption throughout the whole project duration. The relevant risk issues are discussed during the regular Board meetings, which ensures that potential risks are taken seriously. The Risk Management Plan points out potential risks and defines measures needed in order to minimize the consequences of the risks. This plan was reviewed and improved regularly to ensure that any new risks are taken into account.
We have met the following risk: within the Additional Activity, MU planned to organize the HaBit Evaluation Workshop with the cooperation the project partner and the University of

Addis Ababa, Ethiopia. During the final preparations in October 2016, Ethiopian government has declared a state of emergency and the Czech Ministry of Foreign Affairs strongly recommended not to travel to Ethiopia. However, MU managed to fulfil the original purpose of the evaluation of the developed Efficient Annotation Framework with a prompt change in the Workshop organization and inviting 8 Ethiopian linguistic experts to Brno, Czech Republic. In this way the risk has been successfully handled.

2.6.2 Irregularities in reporting period (max. 1 page A4)
No irregularities occurred.

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

The consortium's main aim is to make all project outputs open access, the resources as well as the tools. All the project partners including the collaborators from Ethiopia have access rights. The distribution of intellectual property rights is controlled by the Project Management Board in accordance with the Partnership Agreement; all decisions about intellectual property rights as well as the plans for the exploitation of results and licensing strategies follow the project Technology Management Plan.

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

The progress of the HaBiT project was published and presented at scientific conferences with topics related to the topics of the HaBiT project. These conferences include:

- 8th International Conference on Agents and Artificial Intelligence, Rome, Italy
- The XVII EURALEX International congress, Tbilisi, Georgia
- International Workshop on Semantic Evaluation (SemEval-2016), San Diego, California
- The Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia
- The First Conference on Machine Translation, Berlin, Germany
- The Second Workshop on Computational Approaches to Code-Switching, Austin, Texas, USA
- The Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3), Osaka, Japan
- The Second Workshop on Noisy User-Generated Text (W-NUT 2016), Osaka, Japan
- Text, Speech, and Dialogue 19th International Conference (TSD 2016), Brno, Czech Republic
- The Thirteenth International Conference on Natural Language Processing (ICON-2016), Varanasi, India
- The Fifteenth Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain
- The Eighteenth International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2017), Budapest, Hungary
- The 21st Nordic Conference on Computational Linguistics (NoDaLiDa 2017), Göteborg, Sweden
- The Twelfth Annual Ministerial Level Research Conference on Information Sciences and Technology for Africa (IST-Africa 2017), Windhoek, Namibia

- The 10th Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2016), Karlova Studánka, Czech Republic. This workshop was partly organized as a HaBiT informational event. (For more information see the project web page.)

2.8.1 Reporting on scientific (joint) publications (length according your need)

APRESJAN, Valentina, Vít BAISA, Olga BUIVOLOVA a Olga KULTEPINA. R**uSkELL: 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.

RuSkELL ("Russian + Sketch Engine for Language Learning") is a new online resource intended for researchers and learners of Russian. It incorporates a specially pre-processed corpus and the interface which allows users to search for phrases in sentences, extract salient collocates and show similar words. The tool builds upon its English counterpart SkELL (Baisa & Suchomel 2014). The aim of the project is to adapt the existing SkELL tool to Russian, improve its performance and make it user-friendly to Russian users. The existing problems include errors in query output and insufficiently transparent interface. The project aspires to solve them by 1) modifying Sketch grammar rules to exclude irrelevant output and to add informative collocations unaccounted for in the existing Sketch grammar; 2) providing collocation groups with easy-to-understand labels in Russian. The authors describe the process of building the language data and problems needed to be addressed to accommodate the tool for the specifics of the Russian language.

BAISA, Vít. **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. MUNI/A/0863/2015, project 7F14047.

AGREE is a dataset and task for evaluation of language models based on grammar agreement in Czech. The dataset consists of sentences with marked suffixes of past tense verbs. The task is to choose the right verb suffix which depends on gender, number and animacy of subject. It is challenging for language models because 1) Czech is morphologically rich, 2) it has relatively free word order, 3) high out-of-vocabulary (OOV) ratio, 4) predicate and subject can be far from each other, 5) subjects can be unexpressed and 6) various semantic rules may apply. The task provides a straightforward and easily reproducible way of evaluating language models on a morphologically rich language.

BAISA, Vít, Jan MICHELFEIT, Marek MEDVEĎ a 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.

Several parallel corpora built from European Union language resources are presented here. They were processed by state-of-the-art tools and made available for researchers in the Sketch Engine corpus management system. A completely new resource is introduced: EUR-Lex corpus, being one of the largest parallel corpus available at the moment, containing

840 million tokens of English and having the largest language pair (English-French) with more than 25 million aligned segments (paragraphs).

BAISA, Vít, Silvie CINKOVÁ, Ema KREJČOVÁ and Anna VERNEROVÁ. **VPS - Grade-Up: Graded Decisions on Usage Patterns.** *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia: European Language Resources Association (ELRA), 2016. ISBN 978-2-9517408-9-1.

The paper presents VPS-Grade-Up — a set of 11,400 graded human decisions on usage patterns of 29 English lexical verbs from the Pattern Dictionary of English Verbs by Patrick Hanks. The annotation contains, for each verb lemma, a batch of 50 concordances with the given lemma as KWIC, and for each of these concordances the authors provide a graded human decision on how well the individual PDEV patterns for this particular lemma illustrate the given concordance, indicated on a 7-point Likert scale for each PDEV pattern. With their annotation, they were pursuing a pilot investigation of the foundations of human clustering and disambiguation decisions with respect to usage patterns of verbs in context. The data set is publicly available at http://hdl.handle.net/11234/1-1585.

BAISA, Vít, Sara MOŽE a 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.

This paper presents the results of a pilot study in linking corresponding English and Spanish verb patterns using both automatic and manual procedures. Our work is rooted in Corpus Pattern Analysis (CPA) (Hanks 2004, 2013), a corpus-driven technique that was used in the creation of existing monolingual pattern dictionaries of English and Spanish verbs, which were used in our experiment to design a gold standard of manually annotated verb pattern pairs. Research in CPA has inspired parallel projects in English, Spanish, Italian and German. Our study represents the first attempt to build a multilingual lexical resource by linking verb patterns in these languages. Verbs display special difficulties related to grammar and argument structure that are not found in other parts-of-speech, and for that reason it is necessary to create a specific resource for them. After applying the automatic matching to a set of 87 Spanish verbs linked to 176 English verbs, an evaluation of a random selection of 50 of these pairs shows 80% precision.

CINKOVÁ, Silvie, Ema KREJČOVÁ, Anna VERNEROVÁ and Vít BAISA. **Graded and Word – Sense - Disambiguation Decisions in Corpus Pattern Analysis: a Pilot Study**. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia: European Language Resources Association (ELRA), 2016. ISBN 978-2-9517408-9-1

The authors present a pilot analysis of a new linguistic resource, VPS-Grade-Up (available at http://hdl.handle.net/11234/1-1585). The resource contains 11,400 graded human decisions on usage patterns of 29 English lexical verbs, randomly selected from the Pattern Dictionary of English Verbs (Hanks, 2000 2014) based on their frequency and the number of senses their lemmas have in PDEV. This data set has been created to observe the interannotator agreement on PDEV patterns produced using the Corpus Pattern Analysis (Hanks, 2013). Apart from the graded decisions, the data set also contains traditional Word-Sense-Disambiguation (WSD) labels. They analyze the associations between the

graded annotation and WSD annotation. The results of the respective annotations do not correlate with the size of the usage pattern inventory for the respective verbs lemmas, which makes the data set worth further linguistic analysis.

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.

The paper describes a named entity recognition system for Amharic, an under-resourced language, using a recurrent neural network, a bidirectional long short term memory model to identify and classify tokens into six predefined classes: Person, Location, Organization, Time, Title, and Other (non-named entity tokens). Word vectors based on semantic information are built for all tokens using an unsupervised learning algorithm, *word2vec*. The word vectors were merged with a set of specifically developed language independent features and together fed to the neural network model to predict the classes of the words. When evaluated by 10-fold cross-validation, the created Amharic named entity recogniser achieved good average precision (77.2%), but did worse on recall (63.4%), for a 69.7% F_1-score.

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.

The paper introduces a deep learning-based Twitter hate-speech text classification system. The classifier assigns each tweet to one of four predefined categories: racism, sexism, both (racism and sexism) and non-hate-speech. Four Convolutional Neural Network models were trained on resp. character 4-grams, word vectors based on semantic information built using word2vec, randomly generated word vectors, and word vectors combined with character n-grams. The feature set was down-sized in the networks by maxpooling, and a softmax function used to classify tweets. Tested by 10-fold cross-validation, the model based on word2vec embeddings performed best, with higher precision than recall, and a 78.3% F-score.

HERMAN, Ondřej, Vít SUCHOMEL, Vít BAISA and 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: Association for Natural Language Processing (ANLP), Osaka, Japan, 2016. s. 114-118, 5 s. ISBN 978-4-87974-716-7.

In this paper the authors investigate two approaches to discrimination of similar languages: Expectation--maximization algorithm for estimating conditional probability P(word|language) and byte level language models similar to compression-based language modelling methods. The accuracy of these methods reached respectively 86.6 % and 88.3 % on set A of the DSL Shared task 2016 competition.

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.

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.

JAHREN, Brage Ekroll, Valerij FREDRIKSEN, Björn GAMBÄCK and Lars BUNGUM. **NTNUSentEval at SemEval-2016 Task 4: Combining General Classifiers for Fast Twitter Sentiment Analysis**. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pp. 103-108, San Diego, California, June 2016. ISBN 978-1-941643-95-2. ACL.

The paper describes experiments on sentiment classification of microblog messages using an architecture allowing general machine learning classifiers to be combined either sequentially to form a multi-step classifier, or in parallel, creating an ensemble classifier. The system achieved very competitive results in the shared task on sentiment analysis in Twitter, in particular on non-Twitter social media data, that is, input it was not specifically tailored to.

JAKUBÍČEK, Miloš, Vít BAISA, Jan BUŠTA, Vojtěch KOVÁŘ, Jan MICHELFEIT, Pavel RYCHLÝ a Vít SUCHOMEL. **Walking the tightrope between linguistics and language engineering**. 2017. GA15-13277S, project 7F14047.

Very many state-of-the-art solutions in language technology owe their success to the right balance between a wide range of linguistic introspection and theory neutral computer engineering. And Sketch Engine is undoubtedly one of them. In this chapter we elaborate on both the theoretical and practical issues we have faced in the thirteen years of Sketch Engine development and argue for both the linguistic and computer science oriented decisions we have taken. We also discuss Sketch Engine's current challenges from which many can be extrapolated to any language technology software aiming at industrial strength impact.

JAKUBÍČEK, Miloš a 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. 7F14047.

The authors present a novel method for performing fast keyword extraction from large text corpora using a finite state backend. The FSA3 package has been adopted for this

purposes. They outline the basic approach and present a comparison with previous hash-based method as used in Sketch Engine.

JAMATIA, Anupam, Björn GAMBÄCK and Amitava DAS. **Collecting and Annotating Indian Social Media Code-Mixed Corpora**. In Alexander Gelbukh (ed.): The 17th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2016), Konya, Turkey, April 2016, Springer Lecture Notes in Computer Science.

The pervasiveness of social media in the present digital era has empowered the 'netizens' to be more creative and interactive, and to generate content using free language forms that often are closer to spoken language and hence show phenomena previously mainly analysed in speech. One such phenomenon is code-mixing, which occurs when multilingual persons switch freely between the languages they have in common. Code-mixing presents many new challenges for language processing and the paper discusses some of them, taking as a starting point the problems of collecting and annotating three corpora of code-mixed Indian social media text: one corpus with English-Bengali Twitter messages and two corpora containing English-Hindi Twitter and Facebook messages, respectively. We present statistics of these corpora, discuss part-of-speech tagging of the corpora using both a coarse-grained and a fine-grained tagset, and compare their complexity to several other code-mixed corpora based on a Code-Mixing Index.

KOVÁŘ, Vojtěch. **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. 7F14047.

In Low inter-annotator agreement = an ill-defined problem?, the authors have argued that tasks with low inter-annotator agreement are really common in natural language processing (NLP) and they deserve an appropriate attention. They have also outlined a preliminary solution for their evaluation. In On evaluation of natural language processing tasks: Is gold standard evaluation methodology a good solution?, the authors have agitated for extrinsic application-based evaluation of NLP tasks and against the gold standard methodology which is currently almost the only one really used in the NLP field. This paper brings a synthesis of these two: for three practical tasks, that normally have so low inter-annotator agreement that they are considered almost irrelevant to any scientific evaluation, the authors have introduced an application-based evaluation scenario which illustrates that it is not only possible to evaluate them in a scientific way, but that this type of evaluation is much more telling than the gold standard way.

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.

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.

KOVÁŘ, Vojtěch, Miloš JAKUBÍČEK a 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.

The paper discusses problems in state of the art evaluation methods used in natural language processing (NLP). Usually, some form of gold standard data is used for evaluation of various NLP tasks, ranging from morphological annotation to semantic analysis. The authors discuss problems and validity of this type of evaluation for various tasks, and illustrate the problems on examples. Then they propose using application-driven evaluations, wherever it is possible. Although it is more expensive, more complicated and not so precise, it is the only way to find out if a particular tool is useful at all.

KOVÁŘ, Vojtěch, Jakub MACHURA, Kristýna ZEMKOVÁ a 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 International Publishing, 2016. s. 287-294, 8 s. ISBN 978-3-319-45509-9. doi:10.1007/978-3-319-45510-5_31

Punctuation detection and correction belongs to the hardest automatic grammar checking tasks for the Czech language. The paper compares available grammar and punctuation correction programs on several data sets. It also describes a set of improvements of one of the available tools, leading to significantly better recall, as well as precision.

KOVÁŘ, Vojtěch, Monika MOČIANKOVÁ and Pavel RYCHLÝ. **Finding Definitions in Large Corpora with Sketch Engine.** *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia: European Language Resources Association (ELRA), 2016. ISBN 978-2-9517408-9-1.

The paper describes automatic definition finding implemented within the leading corpus query and management tool, Sketch Engine. The implementation exploits complex pattern-matching queries in the corpus query language (CQL) and the indexing mechanism of word sketches for finding and storing definition candidates throughout the corpus. The approach is evaluated for Czech and English corpora, showing that the results are usable in practice: precision of the tool ranges between 30 and 75 percent (depending on the major corpus text types) and we were able to extract nearly 2 million definition candidates from an English corpus with 1.4 billion words. The feature is embedded into the interface as a concordance filter, so that users can search for definitions of any query to the corpus, including very specific multi-word queries. The results also indicate that ordinary texts (unlike explanatory texts) contain rather low number of definitions, which is perhaps the most important problem with automatic definition finding in general.

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.

A community in social networks is generally assumed to be composed of a group of individuals with similar characteristics. Although there has been a plethora of work on understanding network topologies (edge density, clustering coefficient, etc.) within an online community, the psycho-sociological compositions of social network communities have hardly been studied. The present paper aims to analyse the communities as composition of

induced psycholinguistic and sociolinguistic variables (Personalities, Values and Ethics) across individuals in social media networks. The motivation behind this analysis is to understand the behavioural characteristics at individual as well as societal level in social networks. To this end, three studies were carried out on six different datasets: four Twitter corpora, a Facebook corpus, and an Essay corpus, annotated with Values and Ethics of the users. First, experiments on creating automatic models to determine the Personality and Values of individuals by analysing their language usage and social media behaviour. Second, experiments on understanding the characteristics or blend of characteristics of individuals within an online community. Finally, generation of a map of values and ethics for India, a multi-lingual and multicultural country. Striking similarities to general intuitive perception could be observed, i.e., the results obtained in the study resemble our general perception about the cities/towns of India.

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.

To find out how users' social media behaviour and language are related to their ethical practices, the paper investigates applying Schwartz' psycholinguistic model of societal sentiment to social media text. The analysis is based on corpora collected from user essays as well as social media (Facebook and Twitter). Several experiments were carried out on the corpora to classify the ethical values of users, incorporating Linguistic Inquiry Word Count analysis, n-grams, topic models, psycholinguistic lexica, speech-acts, and non-linguistic information, while applying a range of machine learners (Support Vector Machines, Logistic Regression, and Random Forests) to identify the best linguistic and non-linguistic features for automatic classification of values and ethics.

MEDVEĎ, Marek a 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.

Question answering (QA) systems have become popular nowadays, however, a majority of them concentrates on the English language and most of them are oriented to a specific limited problem domain. In this paper, a new question answering system called AQA (Automatic Question Answering) is presented. AQA is an open-domain QA system, which allows users to ask all common questions related to a selected text collection. The first version of the AQA system is developed and tested for the Czech language, but the authors also plan to include more languages in future versions. The AQA strategy consists of three main parts: question processing,answer selection and answer extraction. All modules are syntax-based with advanced scoring obtained by a combination of TF-IDF, tree distance between the question and candidate answers and other selected criteria. The answer extraction module utilizes named entity recognizer which allows the system to catch entities that are most likely to answer the question. Evaluation of the AQA system is performed on a previously published Simple Question-Answering Database, or SQAD, with more than 3,000

question-answer pairs.

MEDVEĎ, Marek, Vojtěch KOVÁŘ a 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.

In this paper the authors present their approach to the Bilingual Document Alignment Task (WMT16), where the main goal was to reach the best recall on extracting aligned pages within the provided data. The approach consists of three main parts: data preprocessing, keyword extraction and text pairs scoring based on keyword matching. For text preprocessing the authors use the TreeTagger pipeline that contains the Unitok tool (Michelfeit et al., 2014) for tokenization and the TreeTagger morphological analyzer (Schmid, 1994). After keywords extraction from the texts according TF-IDF scoring the system searches for comparable English-French pairs. Using a statistical dictionary created from a large English-French parallel corpus, the system is able to find comparable documents. At the end this procedure is combined with the baseline algorithm and best one-to-one pairing is selected. The result reaches 91.6% recall on provided training data. After a deep error analysis (see section 5) the recall reached 97.4%.

NEVĚŘILOVÁ, Zuzana. **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.

Language mixing (using chunks of foreign language in a native language utterance) occurs frequently. Foreign language chunks have to be detected because their annotation is often incorrect. In the standard pipelines of Czech texts annotation, no such detection exists. Before morphological disambiguation, unrecognized words are processed by Czech guesser which is successful on Czech words (e.g. neologisms, typos) but its usage makes no sense on foreign words. A new pipeline is proposed that adds foreign language chunk and multi-word expression (MWE) detection. The author experimented with a small corpus where she compared the original (semi-automatic) annotation (including foreign words and MWEs) with the results of the new pipelines. As a result, she reduced the number of incorrect annotations of interlingual homographs and foreign language chunks in the new pipeline compared to the standard one. She also reduced the number of tokens that have to be processed by the guesser. The aim was to use the guesser solely on potentially Czech words.

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.

Identifying sarcasm in text is a challenging task which can be difficult also for humans, in particular in very short texts with little explicit context, such as tweets (Twitter messages). The paper presents a comparison of three sets of tweets marked for sarcasm, two annotated manually and one annotated using the common strategy of relying on the authors correctly using hashtags to mark sarcasm. To evaluate the difficulty of the datasets, a state-of-the-art system for automatic sarcasm detection in tweets was implemented. Experiments on the two manually annotated datasets show comparable results, while

deviating considerably from results on automatically annotated data, indicating that using hashtags is not a reliable approach to creating Twitter sarcasm corpora.

RYCHLÝ, Pavel and 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. doi:10.1007/978-3-319-45510-5

Amharic is one of under-resourced languages. The paper presents two text corpora. The first one is a substantially cleaned version of existing morphologically annotated WIC Corpus (210,000 words). The second one is the largest Amharic text corpus (17 million words). It was created from Web pages automatically crawled in 2013, 2015 and 2016. It is part-of-speech annotated by a tagger trained and evaluated on the WIC Corpus.

SIKDAR, Utpal Kumar and Björn GAMBÄCK. **Language Identification in Code-Switched Text Using Conditional Random Fields and Babelnet**. In Mona Diab, Pascale Fung, Mahmoud Ghoneim, Julia Hirschberg and Thamar Solorio (eds.): Proceedings of the Second Workshop on Computational Approaches to Code-Switching at EMNLP 2016, pp. 127-131, Austin, Texas, USA, November 2016. ISBN 978-1-945626-28-9. ACL.

The paper outlines a supervised approach to language identification in code-switched data, framing this as a sequence labeling task where the label of each token is identified using a classifier based on Conditional Random Fields and trained on a range of different features, extracted both from the training data and by using information from Babelnet and Babelfy. The method was tested on the development dataset provided by organizers of the shared task on language identification in code-switched data, obtaining tweet level monolingual, code-switched and weighted F1-scores of 94%, 85% and 91%, respectively, with a token level accuracy of 95.8%. When evaluated on the unseen test data, the system achieved 91%, 87% and 89% monolingual, code-switched and weighted tweet level F1-scores, and a token level accuracy of 96.5%.

SIKDAR, Utpal Kumar and Björn GAMBÄCK. **Feature-Rich Twitter Named Entity Recognition and Classification**. In Proceedings of the 2nd Workshop on Noisy User-generated Text (W-NUT 2016), pp. 164-170. Osaka, Japan December, 2016. ISBN 978-4-87974-707-5. ACL.

Twitter named entity recognition is the process of identifying proper names and classifying them into some predefined labels/categories. The paper introduces a Twitter named entity system using a supervised machine learning approach, namely Conditional Random Fields. A large set of different features was developed and the system was trained using these. The Twitter named entity task can be divided into two parts: i) Named entity extraction from tweets and ii) Twitter name classification into ten different types. For Twitter named entity recognition on unseen test data, our system obtained the second highest F1 score in the shared task: 63.22%. The system performance on the classification task was worse, with an F1 measure of 40.06% on unseen test data, which was the fourth best of the ten systems participating in the shared task.

SIKDAR, Utpal Kumar and Björn GAMBÄCK. **Twitter Named Entity Extraction and Linking Using Differential Evolution**. In D.S. Sharma, R. Sangal and A.K. Singh (eds.): Proceedings of the 13th International Conference on Natural Language Processing (ICON-2016), pp. 198-207, Varanasi, India, December 2016, NLPAI.

Systems that simultaneously identify and classify named entities in Twitter typically show poor recall. To remedy this, the task is here divided into two parts: i) named entity *identification* using Conditional Random Fields in a multi-objective framework built on Differential Evolution, and ii) named entity *classification* using Vector Space Modelling and edit distance techniques. Differential Evolution is an evolutionary algorithm, which not only optimises the features, but also identifies the proper context window for each selected feature. The approach obtains F-scores of 70.7% for Twitter named entity extraction and 66.0% for entity linking to the DBpedia database.

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.

In order to improve the performance of a deep-learning neural network, the paper outlines a stack-based approach incorporating various information sources. A named entity recognition system for Amharic was implemented using a recurrent neural network, a bi-directional long short term memory model. Word vectors based on semantic information were built using an unsupervised learning algorithm, *word2vec*, while a Conditional Random Fields (CRF) classifier was trained on language independent features to predict each token's named entity class. The predictions, features and word vectors were fed to the deep neural network to assign labels to the words. This stack-based approach reached an 74.26% F-score, outperforming various other deep-learning set-ups, as well as a baseline CRF classifier, and an ensemble method incorporating the same information sources.

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.

Conventional topic modeling schemes, such as Latent Dirichlet Allocation, are known to perform inadequately when applied to tweets, due to the sparsity of short documents. To alleviate these disadvantages, we apply several pooling techniques, aggregating similar tweets into individual documents, and specifically study the aggregation of tweets sharing authors or hashtags. The results show that aggregating similar tweets into individual documents significantly increases topic coherence.

Please provide a list of references of scientific (joint) publications dedicated to the research project which were published in the reporting period if any. Please add a short abstract to each reference as well.

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

The project web page was regularly updated with up-to-date information and activities. According to publicity plan four events were held:

1) **Community-based Building of Language Resources Workshop**, CBBLR 2016 was held on September 12th, 2016 as a pre-conference workshop of the 19th International Conference of Speech, Text and Dialogue. The main topic of the workshop was directed at building new language resources, especially for languages with no or too little existing language resources. The workshop was organized in cooperation with the HaBiT CZ-NO project Consortium, with submissions open to other resource

development project. (Photos can be found at the project web page: http://habit-project.eu/wiki/InformationEvents)

2) **The Tenth Workshop on Recent Advances in Slavonic Natural Language Processing** (RASLAN 2016) was held from December 2nd to December 4th, 2016 in Karlova Studánka, Czech Republic. 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. Topics of the Workshop have included (but have not been limited to): text corpora and tagging, syntactic parsing, sense disambiguation, machine translation, semantic networks and ontologies, semantic web, knowledge representation and reasoning and applied systems and software. Three HaBiT researchers -- Vít Baisa, Vojtěch Kovář and Miloš Jakubíček -- actively participated presenting the outputs and progress of project. The event also helped to promote the HaBiT project with a roll-up. (Photos can be found at the project web page: http://habit-project.eu/wiki/InformationEvents)
3) **HaBiT Evaluation Workshop** was held between February 16th and 21st, 2017 at Masaryk University, Brno, Czech Republic. The goal of the Workshop was an evaluation of our new Efficient Annotation Framework for under-resourced languages. The objective aimed 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. MU has prepared the tools with corpora in the 4 Ethiopian languages HaBiT project deals with plus Czech and Norwegian. These tools allowed for efficient preparation of the new Part-of-Speech (PoS) taggers of these languages. The texts have been annotated (during the workshop) by the native speakers of the four Ethiopian languages for PoS tags, and the tool was acquired and improved itself during this process. (Photos can be found at the project web page: http://habit-project.eu/wiki/InformationEvents)
4) **HaBiT Closing Workshop** was held between April 26th and 28th at University of Oslo, Oslo, Norway. The closing event targeted at presenting project's outputs and achievements, informing about the HaBiT project consortium, and promoting HaBiT project, Programme, Norwegian financial mechanism. (Photos can be found at the project web page: http://habit-project.eu/wiki/InformationEvents)

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 succeeded in creation of a repository for the investigated languages, making them accessible for further research, enabling to acquire information technologies in a less-developed countries (such as Ethiopia) and thus contributing to its cultural development. The accessibility of the results contributes to further development of research in the area of under-resourced languages.

Number of PhD students and postdocs were actively participating in research, helping to transfer and share state-of-the-art development. The close cooperation between partners led to numerous project outputs that were presented to the international research community and have been disseminated via the project web page aiming at presenting the research to any kind of audience.

Please describe how your project contributes in the medium to long term to the objectives of the Programme. You should describe the steps that were necessary in the project (and outside the projects) to bring about these impacts (e.g. dissemination and exploitation of project results, stakeholders involvement). For more see section 1.1 in the Guide for Applicants from the Call for proposals 2013.

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

Number of PhD students	**7**
Number of postdocs	**4**
Number of female researchers (including after maternity leave)	**2**
Number of internationally refereed (joint) scientific publications	**32**

Note: Numbers of the target groups are for the whole project. These indicators are stipulated in the Programme Agreement and 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)

Please provide an explanation and a justification of eligibility of expenditure and to spending of the grant in total and per partner. Put short comments to every single item from the partial budget in the Annex I. Please mention leftovers from previous payments in particular items if applicable.

MU expenditures

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 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. The particular personnel costs are 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, 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 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.
- 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 expenses in 2016 was 3 836 190.48 CZK, which consisted of 1 033 418.78 CZK transferred from 2015 and 2 802 771.7 CZK from 2016 payment. 743 228,30 CZK were transferred to 2017. The expenses in 2017 in total of 2 458 197.77 CZK consisted of 1 714 969.47 from 2017 payment plus the leftover of 743 228,30 CZK from the 2016 payment. Detailed composition of the transfers between years are stated in the Annex II of the report.

NTNU expenditures

The eligible costs for the Norwegian University of Science and Technology as presented in the Annex I of this report, mainly consists of the personnel costs for the following researchers:

- Dr. Utpal Kumar Sikdar
- Lars Bungum

The direct eligible costs furthermore include travel expense and conference registration fees for the two above-mentioned researchers and for:

- Prof. Björn Gambäck, NTNU team leader
- Prof. Janne Bondi Johannessen, UiO team leader

In addition, indirect cost were computed at a flat rate of 60% of the total direct eligible costs.

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

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 increased or decreased the overall project costs. The changes included:

- 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.
- in 2016, 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.

Please explain and justify the project budgetary items changes per category if any in this period. Please relate them to planned deliverables, milestones, outputs, risks, grant utilization, and impacts on the project.
Details have to be provided about all budgetary items which were changed during the reporting period – see Guide for Applicants – section 1.8 – Budget and eligible costs.
Note: You must also fill in the mandatory Annexes I and III.

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

Note: If you reclaim Value Added Tax at state financial authorities, write Yes. If you do not reclaim VAT at state financial authorities, write No. Fill in for each partner.

3.4 Indirect costs model – overheads (approved in annex 2 of 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 were planned for the project, so none were launched in the reporting period.

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

Not relevant.

4. MANDATORY AND VOLUNTARY ANNEXES

4.1 Overview of annexes required to project periodic report

No.	Annexes	Mandatory online submission format
I.	***Annex I – Project Interim Financial Report*** This annex relates to actual expenditure in CZK incurred by all entities from 1 January to 31 December. It is signed by statutory of the Project Promoter, or an attorney, and by principal investigator. Please use the template.	xls(x) and pdf
II.	***Annex II - Report on Actual Incurred Expenditure*** It relates to all Czech entities only. It is a record from an accounting system in CZK reporting costs from 1 January and 31 December per entity corresponds to annex I. It contents a stamp of the organisation and a full name and a signature of a person responsible for financial matters of the organization. No template.	pdf
III.	***Annex III - Financial Statement by Norwegian Partner*** It relates to Norwegian partners in NOK only. It contents a stamp of the organisation and a full name and a signature of a person	pdf

	responsible for financial matters. It may be a copy. Use the template please.	
IV.	***Annex IV - Confidentiality Declaration by Evaluator*** It is related to each annex V. Signed by evaluators. A copy may be submitted. Use the template please.	pdf
V.	***Annex V - Evaluation Report of the Project*** (at least 2 evaluators). Fill in English. Signed by evaluators. Copies may be submitted. Use the template.	pdf
VI.	***Annex VII - Letter of Attorney***, if applicable for Project promoters. Acceptable in Czech. A copy may be submitted. No template.	pdf
VII.	***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

This is a voluntary section. You may add other information you think it is necessary.