PROJECT PROPOSAL				
Czech-Norwegian Research Programme (CZ09)				
Norwegian Financial Mechanism 2009-2014				
Programme area	**Bilateral Research Cooperation**			
Project ID number	7F14047			
Project title in English	Harvesting big text data for under-resourced languages			
Project Promoter (name, full address)	Masarykova univerzita Žerotínovo nám. 617/9, 601 77 Brno Czech Republic			
Project Partner(s) (name, full address)	Norges teknisk-naturvitenskapelige universitet Høgskoleringen 1, 7491 Trondheim Norway			
Name of Principal Investigator	Karel Pala			
Statement	*I hereby declare that the information I state in the project proposal is accurate, true and complete. I am aware that if the information has been reversed in the opposite, I will face disqualification of the project proposal from the Call for proposals.*			
Done in	Brno, Czech Republic			
Date	27/01/2014			
On behalf of Project Promoter				
Stamp of Project Promoter				
Statutory authority of Project Promoter	**Name(s):**	Mikuláš Bek		
	Signature(s):			
	Position:	rector		

On behalf of Project Partner				
Stamp of Project Partner				
Statutory authority of Project Partner	**Name(s):**	Kari Melbi		
	Signature(s):			
	Position:	Pro-rector for Research		

1. GENERAL INFORMATION ABOUT PROJECT

1.1 Project ID

7F14047

1.2 Project acronym

HaBiT

1.3 Project title in English

Harvesting big text data for under-resourced languages

1.4 Project title in Czech

Získávání velkých textových dat pro jazyky s nedostatečným množstvím jazykových zdrojů

1.5 Activity of research and development

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

1.6 Programme thematic area(s)

Environment	Health	Social sciences and Humanities
☐	☐	☒

1.7 Thematic subarea (detailed)

1. Reinforcing knowledge-sharing between universities and society at large
2. Large-scale data gathering
3. ICT meeting societal changes

1.8 Project starting date (dd/mm/yyyy)

01/06/2014

1.9 Project ending date (dd/mm/yyyy)

30/04/2017

1.10 Project duration (mm/yyyy – mm/yyyy)

06/2014 – 04/2017

1.11 Project duration in months (number, e.g. 36)

35

1.12 Total project costs (in CZK)

24,468,000

1.12.1 Total grant request (in CZK)

24,468,000

1.13 Number of partners

2

Note: Number of all partners including Project Promoter.

1.14 Abstract in English (max. ½ A4)

The main goal of the project is to harvest from the Web big text data (corpora) for under-resourced languages, which includes Norwegian, partly Czech and also the major languages in Ethiopia (Amharic, Afaan Oromo, Tigrinya, Somali), hence in particular addressing the call topic "large-scale data gathering". The data will be annotated and parsed to make it usable in various language processing applications, such as information extraction and retrieval, machine translation, etc. The project consortium will include a team from the Czech Republic (Masaryk University, Brno), which will use its existing tools for building Web corpora and coordinate the project, and a Norwegian team (NTNU, Trondheim), which will deal with processing and utilizing the compiled corpora.

One of the project's aims will be 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). Second, NTNU collaborate with University of Oslo and two Ethiopian universities in a project to support linguistic resource building in Ethiopia funded by Norad ("Linguistic Capacity Building – tools for the inclusive development of Ethiopia", NORHED 2013-2018). It is natural to link these activities and to include processing of the four major languages in Ethiopia in the present project: The HaBiT project would be able to feed into and leverage on the NORHED project, thoroughly testing the technologies and thus addressing also the call topics on technology assessment, verification and testing, as well as on ICT meeting societal challenges, hence obtaining a relevant added value also in the political respect through cooperation with a less-developed country. Third, shallow processing applications for Czech and Norwegian, and at least one Ethiopian language, would be built, for investigating and separating multiple senses of the words in the corpora, that is, for word sense induction, as well as for creating multi-sense vector spaces and parallel multi-lingual vector spaces for word translation disambiguation.

1.14.1 Key words in English (max. 20 key words)

Big text data, Web, corpora, parallel corpora, taggers, parsers, corpus managers, Word Sketch Engine, annotation of textual data, collocations, disambiguation, Czech, Norwegian, Amharic, Afaan Oromo, Tigrinya, Somali, Natural Language Processing, vector-space modelling

1.15 Abstract in Czech (max. ½ A4)

Cílem projektu je získat z webu velká textová data (korpusy) pro jazyky s nedostatečnými zdroji, mezi něž patří norština, zčásti čeština a také některé etiopské jazyky (amharština, afaan oromština, tigrinština, somálština). Data budou anotována, parsována tak, aby byla použitelná pro různé aplikace v oblasti počítačového zpracování přirozeného jazyka, např. extrakce informací, strojový překlad a další.

Konsorcium bude tvořeno jedním norským týmem (NTNU Trondheim), který se bude věnovat zpracování vzniklých korpusů, a jedním českým týmem (MU Brno), který využije

svých již existujících nástrojů pro budování korpusů z webu. Projekt bude koordinován brněnským týmem.
Jedním cílem projektu bude vytvoření velkého norského korpusu čítajícího miliardy slovních tvarů s použitím nástrojů vyvinutých v rámci spolupráce s NTNU v EU projektu PRESEMT ("PRESEMT: Pattern REcognition-based Statistically Enhanced MT", 2010-2012). Za druhé, NTNU spolupracuje s Universitou v Oslo a dvěma etiopskými universitami v projektu na podporujícím budování jazykových zdrojů a fundovaném organizací Norad ("Linguistic Capacity Building – tools for the inclusive development of Ethiopia", NORHED 2013-2018). Je tedy přirozené propojit tyto aktivity a zahrnout zpracování čtyř velkých etiopských jazyků do předkládaného projektu: projekt HaBiT tak může podpořit a posílit projekt NORHED důkladným testováním technologií a tím adresovat témata evaluace a verifikace a také splnit společenskou výzvu pro informační technologie (ICT). Takto získáme relevantní přidanou hodnotu rovněž po politické stránce prostřednictvím kooperace s méně rozvinutou zemí. Za třetí, budou vytvořeny aplikace pro povrchové zpracování češtiny a norštiny a aspoň jednoho etiopského jazyka, umožňující vyčlenění a zkoumání mnohoznačnosti slov v korpusech, tj. indukci slovních významů a také tvorbu vícevektorových prostorů a paralelních multilinguálních prostorů pro desambiguaci významů slov při překladu.

1.15.1 Key words in Czech (max. 20 key words)

velká textová data, Web, korpusy, paralelní korpusy, korpusové manažery, morfologické analyzátory, syntaktické analyzátory, Word Sketch Engine, anotace textových dat, kolokace, čeština, norština, amharština, afaan oromština, tigrinština, somálština, zpracování přirozeného jazyka

1.16 Ethical issues (max ¼ page A4)

NO

1.17 VAT reclaim YES or NO

NO

2. PROJECT INTRODUCTION

2.1 Introduction to project; its coherence with Call topic (max. ½ page A4)

The Web is a fast-growing knowledge resource which (among other things) can be used to create very large language corpora, by downloading webpages and then cleaning, analysing, and possibly annotating the created language resources. The present project will in particular address the call topic ``large-scale data gathering´´ by doing exactly this, primarily for Norwegian and Czech. One of the aims of the project will be to build a multi-billion word Norwegian corpus. Furthermore, NTNU is collaborating with University of Oslo and the universities in Addis Ababa and Hawassa in Ethiopia in a project to support linguistic capacity building in Ethiopia funded by Norad through the NORHED programme. The present project would be able to feed into and leverage on the NORHED project, and aim to create corpora related to the four major languages in Ethiopia (Amharic, Afaan Oromo, Tigrinya, Somali). Creating language processing resources for these languages is a challenging and new task, thoroughly testing the previously developed technologies and hence addressing not only the large-scale data gathering topic, but also the topic on technology assessment: verification and testing, as well as the topic of ICT meeting societal challenges. In addition, it would, together with the NORHED project on linguistic capacity building for Ethiopia, address the topic of reinforcing knowledge-sharing between universities and society at large. Still, just building web corpora – even very large ones – rarely is useful unless they are usable in actual applications. Hence, a third aim of the project would be to build shallow processing applications for Czech and Norwegian. This work could also be extended to cover at least one of the Ethiopian languages (Amharic), and would minimally entail creating and/or extending present corpus tools, taggers, parsers and word-level grammars for each language. These can then be used for investigating and separating multiple senses of the words in the corpora, that is, for word sense induction, as well as for creating multi-sense vector spaces and parallel multi-lingual vector spaces for word translation disambiguation.

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

The Natural Language Processing Centre is part of Faculty of Informatics, Masaryk University. The Centre focuses on obtaining practical results in the field of information technologies and linguistics. Its main activities comprise above all: morphological and syntactic analysis, grammar development, corpus linguistics, semantic nets and ontology acquisition, semantic web and visual lexicons, production of various dictionaries and software tools for editing them, development of lexicographer's workbench and machine translation. Results of the projects are frequently published at various conferences, the NLP Centre also cooperates with similarly oriented institutes and companies in the Czech Republic and abroad. The key expertise of the NLP Centre related to the project is building and processing of huge corpora. The Centre has tools for fast querying, annotation and parsing of multi billion corpora.

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

NTNU is Norway's premier academic institution for technology and the natural sciences, with equally strong programmes in the social sciences, the arts and humanities. NTNU's research has an international focus with an interdisciplinary approach. The Department of Computer and Information Science (IDI) has some 180 employees including about 40 permanent research staff. Due to their central position in the European Research Consortium for Informatics and Mathematics (ERCIM) network, staff at NTNU are given easy access to the European research community and the European ICT industry. The language processing team at IDI collaborated with Masaryk University in the FP7 project PRESEMT ("Pattern REcognition-based Statistically Enhanced MT") where NTNU designed and implemented the corpus-processing modul.

2.4 Description of consortium (max. ½ page A4)

MU team

Natural Language Processing Centre is a part of Faculty of Informatics, Masaryk University in Brno and consists of K. Pala, A. Horák, P. Rychlý and Ph.D. students: V. Suchomel, V. Baisa, M. Jakubíček, and V. Kovář. The main research activities of the team include corpus linguistics and processing very large text data (which are the team's key expertise related to HaBiT), semantic web and visual lexicons, production of lexical databases and software tools for them, and machine translation. The tools created in NLP Centre are used by the research institutions all over the world, also by large publishing houses in UK (OUP, Cambridge University Press, MacMillan). In this respect the Centre cooperates with Lexical Computing Ltd. led by A. Kilgarriff from Brighton, UK.

NTNU team

The language processing team at NTNU belongs to the Artificial Intelligence division of the Department of Computer and Information Science. The Norwegian team in HaBiT will consist of Björn Gambäck (Professor of Language Technology, NTNU), Janne Bondi Johannessen (Professor at the Text Laboratory, University of Oslo), Partha Pakray (PostDoc), and a PhD student (to be appointed), together providing a strong background in language technology and knowledge representation, and in language resource building, both for Norwegian and for Ethiopian languages. Within the HaBiT project, the team will participate in and lead the research activities related to corpora building, annotation and processing for Norwegian and for the Ethiopian languages. It will also provide the liaison to the NORHED-funded collaborative project with Ethiop

2.5 Management of project (max. ½ page A4)

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

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

2.6 Communication and decision-making (max. ½ page A4)

The Board of the project will establish the communication channels and methods (organizing meetings, mailing, phoning, Skyping etc.). It will also handle matters of research progress and control the financial issues. The Board members will be the leaders of the individual teams (Karel Pala and Björn Gambäck); at specific meeting possibly represented by one of the other leading group researchers (Pavel Rychlý or Aleš Horák from the MU team and Janne Bondi Johannessen from the NTNU team) and supplemented by the WP leaders.

Day to day communication will be conducted mainly through file-sharing across the web. Regular project meetings with all partners will be scheduled bi-weekly on Google Hangouts.

An all-partner progress meeting together with a Project Board meeting will be held twice a year, but additional meetings can be arranged if needed on the request of either project partner. The Board is the primary mechanism for ensuring adequate communication and interfacing between work packages. To ensure that each partner gains direct insight into the work in progress at other sites, the meetings of the Board will be held on an alternating basis between the two partners.

Every three months the Project Promoter collect information concerning each participant's activities, the state of the work packages/tasks they participate in, and the progress on the deliverables they are contributing to. The relevant WP Leaders and the Board will assist the Project Promoter in producing a progress report combining all information. The Project Promoter will collect and collate six-monthly financial reports, giving financial forecasts and projections if required. Each year an annual report will be created, summarising the development underway in the project. Part of this report will be made public. The public annual reports will document the main results obtained and promote the objectives of the project to the broad public, and be designed for web publishing.

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

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

The Project Promoter will ensure that the risks assessment is a continuous process throughout the entire project duration, and will allocate a dedicated slot to address risk assessment in every Board meeting. The Project Promoter will work with each of the WP Leaders to establish contingency plans in the event of delays in work package deliverables, reduced quality and delays between work packages. The Project Promoter will pay particular attention to potential risks that can have a "snowball effect" for work packages that are dependent on each other. Instrumental in the contingency planning will be a simple but state-of-the-art **Risk Management Plan** which will be developed within the first six months of the project. The **Risk Management Plan** will assess the likely severity of each risk and its potential impact on the project; assess the potential probability of the risk and identify the measures that may be necessary to minimize the impact of the risk should it nevertheless occur. The accuracy of identified risks will be reviewed bi-monthly and the Risk Management Plan will be changed, improved and completed accordingly.

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

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

3. PROJECT FRAMEWORK

3.1 Description of proposed project (max. 10 A4 pages for 3.1.1 to 3.1.5)

3.1.1 Current state of art including your relevant previous work

A major bottleneck for promoting use of computers and the Internet is that many languages lack access to basic tools that would make it possible for people to access ICT in their own language. The META-NET series of White Papers[1] on the status of language processing tools for European languages (which the current EU ICT research programme[2] takes as starting point) states that only English, French and Spanish have sufficient basic tools as of today. The evolution of the Internet and of social media texts, such as Twitter and Facebook messages, has created many new opportunities for creating such tools, but also many new challenges, in particular since many of the types of texts found on the net are characterised by having a high percentage of spelling errors and containing creative spellings (*gr8 for 'great'*), phonetic typing, word play (*goooood for 'good'*), abbreviations (*OMG for 'Oh my God!'*), Meta tags (*URL, Hashtags*), etc. However, most research in this area has concentrated on English texts, whereas ¾ of the Internet users now use other languages[3] and less than half of the Twitter messages are in English (Schroeder 2010).

It is clear that even though English still is the principal language for web communication, there is a growing need to develop technologies for other languages. The primary goal of the project is to investigate and develop techniques and methods that can be used to efficiently create computational linguistic resources for new languages based on existing tools and resources. Masaryk University has together with the company Lexical Computing Ltd created tools for extracting lexical knowledge from the Internet, and has utilised these tools to build corpora for a range of languages, including BiWeC, the Big Web Corpus, aiming at 20 billion tokens (Pomikálek et al. 2009). At present, the largest annotated corpora for Norwegian are NoWaC and NorTenTen with approximately 770 million tokens, the latter created by the same team as part of a joint EU-funded project with NTNU.[4] There is also a larger unannotated Norwegian newspaper corpus continously collected by University of Bergen. The resources that will be created within HaBiT will consist of linguistically annotated text collections and tools for word- and sentence-level analysis of **Norwegian** and **Czech** (for which some such tools and text collections exist).

However, if language processing resources for Norwegian are scarce, this is at a totally different level compared to most languages of the World, including even the largest languages on the African continent. NTNU is collaborating with University of Oslo and the universities in Addis Ababa and Hawassa in Ethiopia in a project to support linguistic capacity building in Ethiopia funded by Norad through the NORHED programme[5]. There are over 80 languages spoken in Ethiopia, with four large ones and some 30 medium-sized but highly

[1] http://www.meta-net.eu/whitepapers/overview

[2] HORIZON 2020, WORK PROGRAMME 2014 – 2015: *5. Leadership in enabling and industrial technologies i. Information and Communication Technologies (European Commission Decision C (2013) 8631; Dec. 2013)*

[3] http://www.internetworldstats.com/stats7.htm

[4] "PRESEMT (**P**attern **RE**cognition-based **S**tatistically **E**nhanced **MT**)", EU Grant Agreement ICT-248307, 2010-2012.

[5] "Linguistic Capacity Building – tools for the inclusive development of Ethiopia". NORHED Partnership

disadvantaged languages. The NORHED project is targeted at those disadvantage languages, by increasing the Ethiopian universities´ knowledge and capacity to develop linguistic resources for them and to provide speakers of those languages with possibilities communicate with public authorities and to obtain education in their own languages. The HaBiT project would be able to feed into and leverage on the NORHED project, but would in contrast target the four major Ethiopian languages (for which very few such resources are available) and aim to create corpora related to those languages, namely Amharic, Tigrinya, Afaan Oromo and Somali, as further described below.

Amharic is the working language of the Ethiopian government. Following the Constitution drafted in 1993, Ethiopia is divided into nine fairly independent regions, each with its own nationality language. However, Amharic is the language for country-wide communication and was also for a long period the principal literal language and medium of instruction in primary and secondary schools of the country, while higher education is carried out in English. Amharic is spoken by about 35 million people as a first or second language, making it the second most spoken Semitic language in the world (after Arabic), probably the second largest language in Ethiopia (after Oromo), and possibly one of the five largest languages on the African continent. The actual size of the population of speakers of different languages in Ethiopia must be based on estimates: Hudson (1999) analysed the Ethiopian census from 1994 and indicated that more than 40% of the population then understood Amharic, while the current size of the Ethiopian population is about 90 million. (93.8 million according to the CIA World Factbook[6]; 76.9 million according to Ethiopian parliament projections in December 2008 based on the preliminary reports from the census of May 2007). In spite of the relatively large number of speakers, Amharic is still a language for which very few computational linguistic resources have been developed, and very little has been done in terms of making useful higher level Internet or computer-based applications available to those who only speak Amharic. The largest corpora for Amharic being a 3.5 million word untagged corpus (Gambäck & Asker 2010) and a tagged corpus of 200,000 words (Gambäck 2012). It is generally believed that applications such as information retrieval, text classification, or document filtering could benefit from the existence and availability of basic tools such as stemmers, morphological analysers or part-of-speech taggers. However, since so few language processing resources for Amharic have been available, very little is known about their effect on retrieval or classification performance for this language. Written Amharic (together with the closely related Tigrinya language) uses a unique script which has originated from the Ge'ez alphabet (the liturgical language of the Ethiopian Orthodox Church). Written Ge'ez can be traced back to at least the 4th century A.D. Unlike Arabic or Hebrew, the language is written from left to right. Like many other Semitic languages, Amharic has a rich verb morphology which is based on tri-consonantal roots with vowel variants describing modifications to, or supplementary detail and variants of the root form. The Amharic writing system uses multitudes of ways to denote compound words and there is no agreed upon spelling standard for compounds. As a result of this – and of the size of the country leading to vast dialectal dispersion – lexical variation and homophony is very common. **Tigrinya** has many similarities to Amharic, but has saved more traits of the common ancestral language Ge'ez. It is spoken as first language by about 8 million people in Ethiopia and Eritrea (2/3 of those living in Ethiopia), being the language of the Ethiopian Tigray region and the de facto national working language of Eritrea. Gathering on-line

[6] https://www.cia.gov/library/publications/the-world-factbook/geos/et.html

textual data for Tigrinya can pose some specific problems, since Eritrea is well-known for being a country with extreme control over the printed medium (Reporters Without Borders rate it as the 178th country out of 178 in terms of media freedom, just below North Korea).

Afaan Oromo is a Cushitic language and the official language of the Oromia regional state, which is the largest state in Ethiopia both in terms of area and population size and includes the nation's capital Addis Ababa. Thus, Oromo is the language with the highest number of native speakers in Ethiopia: according to both the 2007 census and the CIA World Factbook, about 34% of the Ethiopian population (or roughly 35 million) are native Oromo speakers; however, most of these use the language only for spoken communication. Afaan Oromo has a writing system (called "Qubee") that uses the Latin alphabet, but this was introduced in schools only in the 1990s and the majority of the Oromo speakers lack the capacity to use the written form of the language, hence limiting information processing, sharing and dissemination across the Afaan Oromo speaking community. **Somali** is the second largest Cushitic language in the world, with about half as many speakers as Oromo (about 18 million). (The third largest Cushitic language is Sidaama with some 3 million speakers; it is one of the languages specifically addressed in the NORHED project, but will not be a topic of HaBiT.) Somali is the working language of the Ethiopian Somali region (which covers most of the much-disputed Ogaden territory) and is the only Cushitic language accorded official national status, being one of the official languages of Somalia and also one of the recognized national languages of Djibouti. Furthermore, due to the extended conflicts in Somalia and Ogaden, Somali is a large immigration language in many countries, in particular in Norway.

Creating large textual corpora is a very important issue and the main aim of the project. However, actually using the corpora in language processing tasks is equally important, overall, for example for machine translation, for information retrieval and extraction, or for investigating and separating multiple senses of the words in the corpora, that is, for word sense induction (Lau et al. 2012). In HaBiT we will in particular investigate methods to utilize the corpora for creating models for word meaning representation and disambiguation, as well as for creating multi-sense vector spaces (Moen et al. 2013) and parallel multi-lingual vector spaces for word translation disambiguation, a theme which was initiated but not fully investigated in the above-mentioned EU-funded project PRESEMT (Lynum et al. 2012). In recent years, vector-space distributional models have been widely used to represent word meaning and to infer word similarity. Most such models represent a word type by a single vector of contextual features obtained from co-occurrence counts in large textual corpora. By assigning a single vector to each term in the corpus, the resulting model assumes that each term has a fixed semantic meaning (relative to all the other terms). However, due to homonymy and polysemy, word semantics cannot be adequately represented by a single-prototype vector, and many terms have more than one meaning, or sense. Some of these senses are static and can be listed in dictionaries and thesauri, while other senses are dynamic and determined by the contexts the terms occur in. The idea of fixed generic word senses has thus received a fair amount of criticism in the literature (Kilgarriff, 2000). Still, work in Word Sense Disambiguation often concentrate on the static word senses, making the task of distinguishing between them one of classification into a predefined set of classes (i.e., the given word senses); see, e.g., Erk et al. (2013) and Navigli (2009) for overviews of current work in the area.

Multi-prototype distributional models (Schütze 1998) in contrast employ different vectors to represent different senses of a word (Reisinger and Mooney, 2010). Multiple prototypes can be obtained by first constructing context vectors for all words and then clustering similar context vectors to create a sense vector. This may be expensive, as vectors need be to stored and clustered. Instead, we will in the HaBiT project aim to use a new method called Multi-Sense Random Indexing (MSRI; Moen et al., 2013), which is based on Random Indexing (Kanerva et al., 2000) and performs an on-the-fly (incremental) clustering. MSRI is a method for building a multiprototype/multi-sense vector space model, which attempts to capture one or more senses per unique term in an unsupervised manner, where each sense is represented as a separate vector in the model. This differs from the classical Random Indexing method which assumes a static sense inventory by restricting each term to have only one vector (sense).

3.1.2 Project objective(s)

The HaBiT project will in particular address the call topic ``large-scale data gathering´´ by doing gathering large textual corpora from the Web, primarily for Norwegian and Czech. One of the aims of the project will be to build a multi-billion word Norwegian corpus.[7] We would also investigate the possibilities to build parallel corpora, both Czech – Norwegian and between each of the languages and English. Furthermore, the present project would be able to feed into and leverage on the NORHED project, but would in contrast target the four major Ethiopian languages (Amharic, Afaan Oromo, Tigrinya, Somali) and aim to create corpora related to those. Since the Ethiopian languages are quite different both in nature and in web-exposure to most European languages, creating language processing resources for them would be a challenging and new task, thoroughly testing the previously developed technologies and hence addressing not only the large-scale data gathering topic, but also the topic on technology assessment: verification and testing, as well as the topic of ICT meeting societal challenges. In addition it would, together with the NORHED project, address the topic of reinforcing knowledge-sharing between universities and society at large.

The issues of using the Web as a corpus of its sort was the topic of a longer discussion in the area of the big data and especially of large corpora development (Kilgarriff 2003). The conclusion being that the Web as such is not suitable as a corpus because it contains boilerplate data (garbage), which prevents users from obtaining clean texts appropriate for a reasonable research. The research in this area has since been directed to the development of algorithms for obtaining clean texts serving as real corpus texts. These algorithms perform crawling of the Web, clean the obtained text, remove duplicities and recognize the character sets occurring on the Web. The outputs of these algorithms are well quantifiable and thus measurable. To be able to process very large text collections obtained from the Web software tools called corpus managers are necessary. In our previous research an extensive attention has been paid to the algorithms, which are cores of the managers and which make it possible to process quickly very large data sets. One of these tools is Manatee/Bonito (Rychlý, 2007), which will be used in the project and also further developed.

[7] Here, the legal restrictions on downloading Norwegian data from the Web also have to be addressed, in consert with the Norwegian Ministry of Culture.

Still, just building web corpora – even very large ones – rarely is useful unless they are usable in actual applications. Hence, a third aim of the project would be to build shallow processing applications, e.g., by using the Sketch Engine (Kilgarriff et al., 2004), an algorithm producing Word Sketches (a word sketch is a corpus-based summary of a word's grammatical and collocational behaviour; Horák et al., 2009). The Sketch Engine combines statistical and rule-based techniques and will be utilized at least for Czech and Norwegian, and possibly for parallel (and comparable) mappings between those two languages, as well as between each language and English. This work could also be extended to cover at least one of the Ethiopian languages (Amharic), and would minimally entail creating and/or extending present corpus tools, taggers, parsers (Kilgarriff et al., 2010) for each language, as well as building a Norwegian Word Sketch Grammar. The shallow processing grammars can then be used for investigating and separating multiple senses of the words in the corpora, that is, for word sense induction (Lau et al. 2012), as well as for creating multi-sense vector spaces (Moen et al., 2013) and parallel multi-lingual vector spaces for word translation disambiguation, a theme which was initiated, but not fully investigated in the PRESEMT project (Lynum et al., 2012). All the above-mentioned tools and resources produced in the HaBiT project will be fully usable and independent of platform.

3.1.3 Methods and approaches

1. For building the above-mentioned corpora, the techniques developed by the NLP Centre FI MU (in cooperation with Lexical Computing Ltd.) will be used. That includes tools for harvesting the web such as the following:

- Crawler Spiderling (Suchomel and Pomikálek, 2012)
- Web page boilerplate removal tool Justext (Pomikálek, 2011)
- Exact duplicate and near duplicate removal tool Onion (Pomikálek, 2011)
- Character encoding detection tool Chared (Pomikálek and Suchomel, 2011)

These methods have already been successfully used for various languages (English, German, Italian, Chinese, Arabic; Pomikálek and Suchomel, 2011). During the project attention will be paid to further development and improvement of the indicated techniques.

2. Annotation tools (taggers, parsers) will take advantage of the existing instruments based both on rule and statistical approaches, such as the Oslo-Bergen Tagger for Norwegian (Hagen et al., 2000). For Ethiopian languages new tools will have to be developed (taggers and parsers), although we will investigate the possibilities to re-use the few resources that already are available for these languages, in particular HornMorpho (Gasser 2011), which provides some morphological processing for Amharic, Tigrinya and Afaan Oromo.

3. Shallow processing techniques and tools such as the Sketch Engine will be used. These will provide detailed information about the grammatical and collocation behaviour of the individual words of language in the form of tables, and can thus be used in various lexicographical applications such as compiling electronic dictionaries, machine translation, and extracting terminology. The Word Sketch Engine is also capable of indexing and fast searching in parallel corpora.

4. As described in Section 3.1.1, most distributional models of word similarity represent a word type by a single vector of contextual features (a *context vector*), even though words commonly have more than one sense. The multiple senses can be captured by employing

several vectors per word in a multi-prototype distributional model, prototypes that can be obtained by first constructing all the context vectors for the word and then clustering similar vectors to create sense vectors. Storing and clustering context vectors can be expensive though. As an alternative, we will aim to use "Multi-Sense Random Indexing" (Moen et al., 2013), an unsupervised learning method which performs on-the-fly (incremental) clustering. Classical Random Indexing (Kanerva et al., 2000) incrementally builds a co-occurrence matrix of reduced dimensionality, by first assigning index vectors to each unique term. Context vectors are also assigned to each term, initially consisting of only zeros. When traversing a document corpus, the context vector of each term is continuously updated with information from its neighbouring terms' index vectors. Then the contextual similarity of two terms can be approximated by a vector similarity measure (such as cosine similarity). Multi-Sense Random Indexing (MSRI) extends this by allowing for several senses per term, each sense represented by different context vectors, which are updated separately. Whenever a term appears in a corpus, its present context is matched with its previously computed context vectors according to a similarity measure. If only one of the vectors is similar enough (i.e., if the similarity measure is lower than a pre-set threshold), that vector is updated in the same way as in classical Random Indexing. However, if none of the vectors is similar enough, it is assumed that a new sense of the term has been encountered and a new context vector is created. The third possibility is that two or more vectors are above the similarity threshold; then those vectors are merged, based on the assumption that they in fact represent the same sense of the term. Moen et al. (2013) carried out some small-scale experiments with limited English corpora and a few different similarity measures and threshold values. The very big corpora created in HaBiT would instead make it possible to carry out experiments at a large scale, with a wider variety of similarity measures, and for languages from four different language families: Germanic (Norwegian), Slavic (Czech), Semitic (Amharic and Tigrinya), and Cushitic (Oromo and Somali). A challenge here is that the induced dynamic word senses do not necessarily correspond to human-created senses, which makes evaluation in traditional word sense disambiguation tasks difficult. However, correlation to human word similarity judgement may provide a way of intrinsic evaluation of the models (Reisinger and Mooney, 2010).

3.1.4 Description of project plan (max ½ page A4)

The project is divided into six work packages relating to following aspects: user requirements, usability studies and evaluation (WP6), system specification, design and integration (WP1), and specific software module specification and development (WP2 – WP5). Work progress will be documented by the relevant deliverables, while a set of milestones have been defined that reflect the project's progress. The first milestone (**MS1**) marks the start of the project in the form of its kick-off meeting.

An iterative development approach will be followed, concerning both individual system modules and the system as a whole. This approach entails the creation of intermediate system prototypes, that will incorporate the results of the project's validation and evaluation activities, allowing the project consortium to effectively address any critical issues that may emerge during development. Three development phases are planned, each resulting in a system prototype. During the first development phase, the overall system specification will be completed at Month 6 (**MS2**). The first versions of the software modules will be developed in work packages WP2-WP4 and ready for integration by Month 12 (**MS3**). The **first, second** and **third** system prototypes, due at Month 18 (**MS4**), Month 24 (**MS5**) and

Month 33 (**MS6**), will include these respective versions of the software modules, and will be subsequently validated/evaluated in terms of performance and quality. The testing results will then be fed back into the module development process to support the system improvement as it proceeds towards next prototypes. The **third** (and pre-final) system prototype will be used for usability studies and final system evaluation. The usability studies will finish at the end of the project lifetime (**MS7**), when the final system will also be released.

3.2 Project outputs (max 2 A4 pages)

3.2.1 Intended short-term outcome(s)

1. Large annotated corpora will be built for Norwegian (tentatively with a size of at least 1 billion tokens, and with the aim of 5 billion tokens). For Czech, a corpus larger than 5 billion tokens will be compiled. For Amharic, Tigrinya, Oromo, and Somali, corpora of at least a few million tokens will be built (aiming at 20 million, at least for Amharic).
2. Parallel Czech-Norwegian corpus will be developed (with size up to 10 million tokens),
3. Software modules such as taggers, parsers, and Sketch Grammars will be newly built for participating languages (Norwegian, and at least Amharic among the Ethiopian languages). Improved results are planned to be obtained for Czech as well,
4. Presentations at international conferences and workshops, with corresponding papers also in the relevant journals,
5. Organization of a workshop related to the under-resourced languages (e.g., within the TSD – Text, Speech and Dialogue – conference framework).

3.2.2 Intended long-term application of outcome(s)

1. Creating a repository for the investigated languages and making them freely accessible for further research (especially in Ethiopia and Norway),
2. Presenting results obtained in the Project to the research community and disseminate the result via the HaBiT project web pages,
3. In general, the accessibility of the results will push forward the research in the area of the under-resourced language and in this way contribute to promoting our knowledge of these languages in a longer perspective.
4. The project results will make it possible to acquire information technologies in a less-developed country and contribute to its cultural development.

3.2.3 Project output(s)

Type of output	Title	Date of accomplishment (mm/yyyy)	Date of realization (mm/yyyy)
D	Evaluation of Sketch Grammar Coverage for a New Language	12/2014	08/2015
D	Building Web Corpora for Ethiopian Languages	01/2016	09/2016
D	Parallel Corpora from Web data	06/2016	02/2017
D	Part of Speech Tagging of Under-resourced Languages	07/2016	03/2017
D	Visualization of Word Sketches	11/2015	05/2016
D	Computer Aided Word Sketch Grammar Development	12/2016	07/2017
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
R	HaBiT system	04/2017	06/2017
R	Set of Ethiopian Web Corpora	08/2016	10/2016

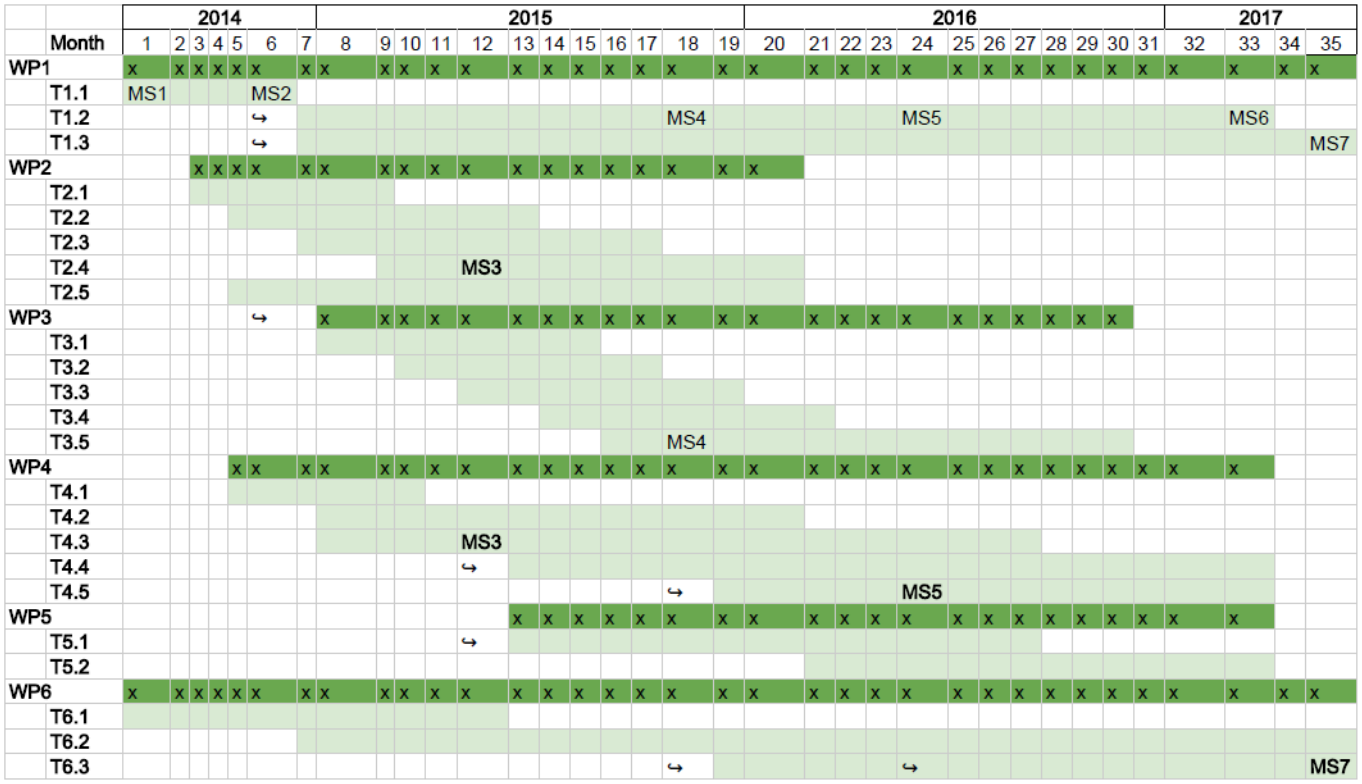
4. WORK PACKAGES (WPS) AND TASKS

4.1. Work Packages (WPs) (max. 10 A4 pages)

4.1.1 Project working packages (WP)

WP number	Title	Date of start (mm/yyyy)	Date of end (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

Gantt chart



4.1.2 WP number

WP1

4.1.3 WP title

System integration

4.1.4 WP leader

Karel Pala

4.1.5 WP start date

01/06/2014

4.1.6 WP end date

30/04/2017

4.1.7 WP objective

- To define the overall system specifications
- To integrate and test all software modules produced in WP4-WP6 in an iterative way
- To create a demo website to illustrate the project to end users and potential stakeholders
- To create the overall integrated system

4.1.8 WP task

Task T1.1: System and modules requirements specification

- Overall system specifications; definition of the functional and non-functional requirements.
- Specifications of the modules to be developed in WP4, WP5 and WP6.

Task T1.2: Module integration and testing

- Integration of the modules developed in WP4, WP5 and WP6 into the HaBiT system prototype.
- Prototype testing and evaluation in three iterations for software review and upgrade.

Task T1.3: Coordination of platform development

- To guarantee that the modules are developed in a consistent way and in line with the system specifications.
- Cross-WP validation of the software components.
- Create the final system demonstrator.

4.1.9 WP deliverable

D1.1.1 System specifications [M6]: Overall system design definitions
D1.1.2 Specification of corpora and the corpus building module [M6]
D1.1.3 Specification of word-sketch grammars and tools [M6]
D1.1.4 Specification of the semantic content matching and wordspace module [M6]
D1.2.1 The HaBiT system v1 [M18]: First integrated system prototype
D1.2.2 The HaBiT system v2 [M24]: Second integrated system prototype

D1.2.3 The HaBiT system v3 [M33]: Third and pre-final integrated system prototype
D1.3 The final HaBiT system [M35]: Tested and evaluated system demonstrator

4.1.10 WP milestone

MS2 System specification [M6]: Overall system specification completed
MS4 HaBiT system v1 [M18]: First integrated system prototype, including first versions of the software modules, ready for performance and quality validation
MS5 HaBiT system v2 [M24]: Second system prototype, including complete versions of the different software modules
MS6 HaBiT system v3 [M33]: Third, pre-final system prototype, ready for usability studies and final system evaluation
MS7 Project end [M35]: Release of the final system

4.1.11 Interdependence with other WPs

WP3 and Tasks T4.2 and T4.3 starts after Milestone MS2.
Tasks T3.5 and T6.3 starts after Milestone MS4

4.1.12 WP Human resources

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

4.1.2 WP number

WP2

4.1.3 WP title

Multi-billion word corpus building

4.1.4 WP leader

Pavel Rychlý

4.1.5 WP start date

01/08/2014

4.1.6 WP end date

31/01/2016

4.1.7 WP objective

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

4.1.8 WP task

Task T2.1: Build language dependent models required by the next step, for both languages. Texts from Wikipedia and hundreds of other web pages will be used to produce a) character trigram model for language identification, b) byte trigram model for character encoding detection, c) the most common words for the boilerplate cleaning.
Task T2.2: Run a web crawler to mine text from the internet. SpiderLing, a state of the art crawler for linguistic purpose will be used as basis. The crawler will be improved to suppress downloading web pages without fluent sentences (e.g., spam) based on Trust Rank (Gyöngyi et al. 2004). The Justext tool and models from the previous step will be used to effectively remove boilerplate, to convert data to encoding UTF-8 and to focus crawling on web domains yielding texts in the required languages.
Task T2.3: The gathered data will be deduplicated using the Onion tool (removes same as well as near duplicate paragraphs), tokenized, split to sentences and transformed in the XML format with separate structures for document, paragraph, sentence. The metadata (language, source, date, size, headings) will be stored as XML attributes of respective structures.
Task T2.4: The processed corpus data will be morphologically annotated to allow further linguistic analysis. The state of the art morphological analyzers for Norwegian and Czech will be used. This step is necessary for defining a word sketch grammar and building word sketches in WP 4.
Task T2.5: Build parallel Czech-Norwegian corpus from freely available data on Web

4.1.9 WP deliverable

D2.1: An improvement of web crawler SpiderLing to suppress downloading web pages without fluent sentences (e.g. spam) based on Trust Rank. [M 10]
D2.2: A Norwegian corpus, sized 5 billion words, cleaned (without boilerplate, de-duplicated), morphologically annotated, indexed for fast searching. [M17]
D2.3: A new Czech corpus, sized 10 billion words, cleaned (without boilerplate, de-duplicated), morphologically annotated, indexed for fast searching. [M20]
D2.4: Parallel Czech-Norwegian corpus, size 10 million tokens [M20]

4.1.10 WP milestone

MS3 The first version of the Norwegian corpus. [M12]

4.1.11 Interdependence with other WPs

The corpora created in this WP will be used in WP 4 as the data source for computing word sketches, and in WP5 for word space models. Tasks T4.4 and WP5 starts after MS3.

4.1.12 WP Human resources

Qualification level: 10 - Karel Pala (docent), Aleš Horák (docent), Pavel Rychlý (PhD), Vít Suchomel (Mgr., PhD student), Vít Baisa (Mgr., PhD student), Miloš Jakubíček (RNDr., PhD student), Vojtěch Kovář (RNDr., PhD student), Björn Gambäck (professor), Partha Pakray (PhD., postdoc researcher), TBA PhD student
Person-months: 16

4.1.2 WP number

WP3

4.1.3 WP title

Corpora for under-resourced languages

4.1.4 WP leader

Pavel Rychlý

4.1.5 WP start date

01/01/2015

4.1.6 WP end date

30/11/2016

4.1.7 WP objective

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

4.1.8 WP task

Task T3.1: Build language dependent models required by the next step, for all languages. Texts from Wikipedia and hundreds of other web pages will be used to produce a) character trigram model for language identification, b) byte trigram model for character encoding detection, c) the most common words for boilerplate cleaning. Task T3.2: Run SpiderLing, a web crawler, to mine text from the internet. Tool Justext and models from the previous step will be used to effectively remove boilerplate, to convert data to encoding UTF-8 and to focus crawling on web domains yielding texts in the required languages. Task T3.3: Since presence of texts in these languages in the internet is very scarce, an approach requiring an extra effort is needed. Methods proposed by the Brno team and successfully used in the past for the Tajik language (Dovudov, 2012) will be followed: Web pages yielding a lot of texts in these languages will be identified in data gathered in previous steps. Custom scripts for harvesting specific web domains will be created. Less data will be discarded by the boilerplate cleaning tool this way and a thus a larger corpus size (compared to standard fully automated crawling) is expected. Task T3.4: Data gathered in tasks T3.2 and T3.3 will be merged together, de-duplicated using tool Onion, tokenized, split to sentences and transformed in the XML format with separate structures for document, paragraph, and sentence. The metadata (language, source, date, size, headings) will be stored as XML attributes of respective structures. Task T3.5: The processed corpus data will be morphologically annotated to allow further linguistic analysis.

4.1.9 WP deliverable

D3.1a: An Amharic corpus, sized 20 million words, cleaned (without boilerplate, de-duplicated), indexed for fast searching. [M15]
D3.2a: An Afaan Oromo corpus, sized 3 million words, cleaned (without boilerplate, de-duplicated), indexed for fast searching. [M17]
D3.3a: A Tigrinya corpus, sized 3 million words, cleaned (without boilerplate, de-duplicated), indexed for fast searching. [M19]
D3.4a: A Somali corpus, sized 10 million words, cleaned (without boilerplate, de-duplicated), indexed for fast searching. [M21]
D3.1b: An Amharic corpus, sized 20 million words, cleaned (without boilerplate, de-duplicated), morphologically annotated, indexed for fast searching. [M18]
D3.2b: An Afaan Oromo corpus, sized 3 million words, cleaned (without boilerplate, de-duplicated), morphologically annotated, indexed for fast searching. [M22]
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. [M30]

4.1.10 WP milestone

Milestone MS4: delivery of D3.1b in month 18. An annotated corpus for Amharic

4.1.11 Interdependence with other WPs

The corpora created in this WP will be used in WP 4 as the data source for computing word sketches. Task T4.5 starts after MS4.

4.1.12 WP Human resources

Qualification level: 11 - Karel Pala (docent), Aleš Horák (docent), Pavel Rychlý (PhD), Vít Suchomel (Mgr., PhD student), Vít Baisa (Mgr., PhD student), Miloš Jakubíček (RNDr., PhD student), Vojtěch Kovář (RNDr., PhD student), Björn Gambäck (professor), Janne Bondi Johannessen (professor), Partha Pakray (PhD., postdoc researcher), TBA PhD student
Person-months: 33

4.1.2 WP number

WP4

4.1.3 WP title

Shallow processing grammars and tools

4.1.4 WP leader

Aleš Horák

4.1.5 WP start date

01/10/2014

4.1.6 WP end date

28/02/2017

4.1.7 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) will be devised for each language in the project. The morphological tagging applied in WPs 2 and 3 will be exploited. The aim of Word Sketch grammars is to enable studying grammatical and collocational behaviour of words in further research.

4.1.8 WP task

Task T4.1: Sketch Grammar evaluation procedure. We will define a procedure which can be used to evaluate a sketch grammar to measure progress in developing new sketch grammars.
Task T4.2: Tools and reports for Sketch Grammar unification and expansion. The tools and reports will help users and sketch grammar developers compare different versions of a sketch grammar or sketch grammars for different languages.
Task T4.3: Visualization of Sketch Grammar queries. A corpus query is the basic component of a sketch grammar, this task will provide 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

4.1.9 WP deliverable

D4.1: Methodology of Sketch Grammar evaluation. Month of delivery: [M10]
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. [M20]
D4.3: Visualization tool for Sketch Grammar queries [M27]
D4.4a: An improved definition of Word Sketches for Czech. [M24]
D4.4b: A new definition of Word Sketches (grammatical and semantic relations) for Norwegian. [M18]
D4.4c: A new definition of Word Sketches (grammatical and semantic relations) for Amharic. [M24]
D4.4d: A new definition of Word Sketches (grammatical and semantic relations) for Afaan Oromo, Tigrinya, and Somali [M33]
D4.5: New language Sketch Grammar module generating initial Sketch grammar for a new language from selected language features [M33]

4.1.10 WP milestone

MS3 delivery of D4.4b

4.1.11 Interdependence with other WPs

WP5 starts after MS3

4.1.12 WP Human resources

Qualification level: 11 - Karel Pala (docent), Aleš Horák (docent), Pavel Rychlý (PhD), Vít Suchomel (Mgr., PhD student), Vít Baisa (Mgr., PhD student), Miloš Jakubíček (RNDr., PhD student), Vojtěch Kovář (RNDr., PhD student), Björn Gambäck (professor), Janne Bondi Johannessen (professor), Partha Pakray (PhD., postdoc researcher), TBA PhD student
Person-months: 28

4.1.2 WP number

WP5

4.1.3 WP title

Multi-sense and multi-lingual word spaces

4.1.4 WP leader

Björn Gambäck

4.1.5 WP start date

01/06/2015

4.1.6 WP end date

28/02/2017

4.1.7 WP objective

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

4.1.8 WP task

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

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

4.1.9 WP deliverable

D5.1 Semantic search interface v1 [M12]: Search interface with rudimentary functionality
D5.2 Dynamic concept matching [M24]: Report on the information matching logic
D5.3 Semantic search interface v2 [M30]: Integrating multi-sense and multi-lingual matching

D5.4 Semantic search interface v3 [M33]: Tested and evaluated search interface

4.1.10 WP milestone

4.1.11 Interdependence with other WPs

WP5 starts after MS3 (delivery D2.2 and D4.4b)

4.1.12 WP Human resources

Qualification level: 9 - Karel Pala (docent), Aleš Horák (docent), Pavel Rychlý (PhD), Miloš Jakubíček (RNDr., PhD student), Vojtěch Kovář (RNDr., PhD student), Björn Gambäck (professor), Janne Bondi Johannessen (professor), Partha Pakray (PhD., postdoc researcher), TBA PhD student
Person-months: 40

4.1.2 WP number

WP6

4.1.3 WP title

Requirements and evaluation

4.1.4 WP leader

Karel Pala

4.1.5 WP start date

01/06/2014

4.1.6 WP end date

30/04/2017

4.1.7 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
- To define and assess the overall project success criteria

4.1.8 WP task

Task T6.1: **Project evaluation strategy**

- Identification of quantifiable evaluation and success criteria
- Selection of appropriate evaluation methods of the different tasks

Task T6.2: **Usability studies**

- Quantitative user testing: A sample of test persons will provide early indication of whether the design meets end-user requirements
- Qualitative usability study: Longitudinal study of system usage

Task T6.3: **System evaluation**

- Integratability tests: Evaluate ease of system integration at external sites
- System evaluation: Performance evaluation of the complete integrated system prototype

4.1.9 WP deliverable

D6.1 Project evaluation plan [M12]: Definition of evaluation criteria and methods
D6.2 Final Usability report [M35]: Report on the overall usability of the system
D6.3 System evaluation [M35]: Report on overall system performance

4.1.10 WP milestone

MS7 delivery of D6.2 and D6.3, final version of the system [M35]

4.1.11 Interdependence with other WPs

Task T6.3 starts after MS4

4.1.12 WP Human resources

Qualification level: 11 - Karel Pala (docent), Aleš Horák (docent), Pavel Rychlý (PhD), Vít Suchomel (Mgr., PhD student), Vít Baisa (Mgr., PhD student), Miloš Jakubíček (RNDr., PhD student), Vojtěch Kovář (RNDr., PhD student), Björn Gambäck (professor), Janne Bondi Johannessen (professor), Partha Pakray (PhD., postdoc researcher), TBA PhD student
Person-months: 13

5. PARTNERS AND TEAMS

5.1. Project Promoter

5.1.1 Project Promoter identification

5.1.1.1 Role	Project Promoter
5.1.1.2 Organization legal name (in Czech)	Masarykova univerzita
5.1.1.2.1 Legal name in English	Masaryk university
5.1.1.3 Abbreviation	MU
5.1.1.4 ID number	00216224
5.1.1.5 VAT number	**CZ00216224**
5.1.1.6 Organization legal form	public university
5.1.1.7 Registration in Commercial Register	
5.1.1.8 Status of organization by Community framework 2006/C 323/01	research organisation
5.1.1.9 Participant identification code (PIC) (if relevant)	999880657
5.1.1.10 Full legal headquarters' address	
5.1.1.10.1 Street, number	Žerotínovo nám. 617/9
5.1.1.10.2 Place/location	Brno
5.1.1.10.3 Post code	601 77
5.1.1.10.4 Country	Czech Republic
5.1.1.11 Bank details	
5.1.1.11.1 Bank full name	**Česká národní banka**
5.1.1.11.2 Bank code	0710
5.1.1.11.3 Account number	94-41924621
5.1.1.11.4 Specific symbol	
5.1.1.11.5 Variable symbol	
5.1.1.12 Contacts	
5.1.1.12.1 Telephone number	+420 549 49 1111
5.1.1.12.2 E-mail	info@muni.cz
5.1.1.12.3 Official web page(s)	www.muni.cz

5.1.2 Statutory authority of Project Promoter

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

5.1.3 Introduction of Project Promoter's team

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

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

5.1.3.2 Competence and capability of team (max. ¾ page A4)

K. Pala (20 %) and A. Horák (20 %) will be responsible for supervising the team members who are Ph.D. students and for coordinating their research work and also for the overall management. A. Horák will be as well involved in the design of annotation techniques. P. Rychlý (20 %) will function in the project as a leading researcher in the area of the building large corpora, corpus tools and their effective exploitation. As a leading programmer he will be helping other team members with the technical tasks. V. Suchomel's (40 %) role will consist in implementing tools for gathering large text data from the Web. V. Baisa (25 %) will take care of the design and implementation of the specialized user interfaces. M. Jakubíček (25 %) will work on creating the program interfaces for processing large text data. V. Kovář's task (25 %) will be to design and implement automatic annotation techniques. The capacity of the particular team member is clearly expressed in the percentages given in the brackets and it is based on our experience with similar tasks in our previous projects (PRESEMT).

5.1.3.3 Motivation factor (max. ¾ page A4)

The main MU motivation factor in the proposed project is to extend research and scientific activities, in particular to explore problems of the large data resources for under-resourced languages of different types, and in this way to obtain deeper knowledge about the concerned languages in general and also with regard to their machine processing. Based on the previous cooperation in the EU project PRESEMT we are motivated to develop parallel resources for Norwegian and Czech, possibly for Amharic as well. In this way new types of knowledge will become accessible for the world research community. The project will involve young researchers – Ph.D. students who will learn and investigate how to handle big data using information technologies. Relevant political factor motivating us is a cooperation with the less developed country (Ethiopia) through providing a high quality know-how in the ICT area.

5.1.3.3.1 Type of motivation factor

a) Increasing the size of the project or activity as a result of the support

5.1.3.3.2 Rationale of motivation factor (max. ½ page A4)

The proposed project will increase the amount of work at MU devoted to the research of obtaining and processing very large text databases in multilingual environment. With the project support, 4 PhD students will be directly involved in the research activities under the guidance of senior researchers of the MU NLP Centre. This will positively influence their research level thanks to obtaining new experience in solving complex research problems.

The project will also heavily support dissemination and international cooperation with regard to the computer processing of languages with insufficient and unbalanced language resources, especially for young researchers. The results of these activities will have significant impact to the computer processing of under-resourced languages in general.
In this respect we also consider it politically relevant that the project results can serve as a good example for development of other under-resourced languages in the world.

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

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

5.1.4 Principal investigator

5.1.4.1 Principal investigator identification

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

*Note: *Voluntary.*

5.1.4.2 Principal investigator's core activities in project (max. ¾ page A4)

Prof. K. Pala has dedicated himself for many years to computer and corpus linguistics, and to the computer processing of natural language, especially Czech. He has achieved the significant success in this area at Faculty of Informatics, Masaryk University (FI MU), the main results include work on Czech morphological analysers Lemma and Ajka, partial syntactic analyser DIS/VADIS and Czech WordNet. He created a Dictionary of Czech surface verb valencies with 15,000 items. He is currently the supervisor of the Natural Language Processing Specialization at FI MU and the head of Natural Language Processing Centre at FI MU. He is a co-chair of the international conference Text, Speech and Dialog and co-editor of the TSD Proceedings printed by Springer Verlag.
Within the project: he will participate in the project management as well as in the research activities, related to corpora building and processing. He will be coordinating the tasks concerned with annotating and evaluating issues. He also will be responsible for supervising young members of the MU team who are Ph.D. students.

5.1.4.3 Principal investigator's internationally refereed (joint) scientific publications (max. ¾ page A4)

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

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

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

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

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

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

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

5.2 Project Partner(s)

Note: Copy the whole chapter 5.2 if having more than one Project Partner.

5.2.1 Project Partner identification

5.2.1.1 Role	Project Partner
5.2.1.2 Organization legal name (in Czech/Norwegian or other)	Norges teknisk-naturvitenskapelige universitet
5.2.1.2.1 Legal name in English	Norwegian University of Science and Technology
5.2.1.3 Abbreviation	NTNU
5.2.1.4 ID number	974 767 880
5.2.1.5 VAT number	NO 974 767 880 MVA
5.2.1.6 Organization legal form	Organization section, Higher education, Public entity
5.2.1.7 Registration in Commercial Register	Brønnøysundregistrene: 974 767 880
5.2.1.8 Status of organization by Community framework 2006/C 323/01	Public body. Higher education establishment
5.2.1.9 Participant identification code (PIC) (if relevant)	999977851
5.2.1.10 Full legal headquarters' address	
5.2.1.10.1 Street, number	Høgskoleringen 1
5.2.1.10.2 Place/location	Trondheim
5.2.1.10.3 Post code	7491
5.2.1.10.4 Country	Norway
5.2.1.11 Contacts	
5.2.1.11.1 Telephone number	+47 73 59 50 00 (switchboard)
5.2.1.11.2 E-mail	postmottak@adm.ntnu.no
5.2.1.11.3 Official web page(s)	www.ntnu.no

5.2.2 Statutory authority of Project Partner

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

5.2.3 Introduction of Project Partner's team

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

The language processing team at NTNU belongs to the Artificial Intelligence division of the Department of Computer and Information Science. The team has a strong background in language technology and knowledge representation, in the application of machine learning strategies to language processing, as well as in search and retrieval techniques, and in

language resource building, both for Norwegian and for Ethiopian languages. In the EC-funded PRESEMT Machine Translation project with MU, the NTNU group lead the work on corpus modelling and word translation disambiguation through vector-space modelling. Currently, the NTNU team collaborates with several other Norwegian universities (U Bergen, U Oslo, U Tromsø) in CLARINO, a project funded by the Norwegian Research Council aiming at gathering language processing resources for Norwegian. The NTNU team also works in the NORHED-funded project with U Oslo and the universities in Addis Ababa and Hawassa on building language processing resources for disadvantaged Ethiopian languages.

The NTNU team in HaBiT will consist of Björn Gambäck (Professor of Language Technology), Partha Pakray (PostDoc research fellow), and a PhD student (to be appointed within the project). The team will be joined by Janne Bondi Johannessen (Professor at the Text Laboratory, University of Oslo) who collaborates with the NTNU team in several projects and who jointly with Gambäck will supervise the non-permanent staff. Both Prof. Gambäck and Prof. Johannessen uphold permanent, tenured university positions, and will hence not be financed through the project. Dr. Pakray is currently working at NTNU, but on an ERCIM-funded fellowship which will end in June 2014. He will thus be able to join the project at its inception, and be 100% financed through the project for a 2-year period (July 2014-June 2016). The fourth member of the NTNU team will be a PhD student who will be appointed within the project and funded by 80% through it. This position will be advertised in May, with a starting date after the summer (hence funded September 2014-April 2017).

5.2.3.2 Competence and capability of team (max. ¾ page A4)

Prof. **Björn Gambäck** (born 1964 in Skarpnäck, Sweden; tekn. dr. / Dr. Eng., Royal Institute of Technology / KTH, Stockholm, 1997, thesis: *Processing Swedish Sentences: A Unification-Based Grammar and some Applications*) was appointed professor in Language Technology at NTNU in 2007, but has lead international and national corporation projects since 1991, including coordinating projects in FP5 and FP6, and leading NTNU's work in the FP7 project PRESEMT. During the spring of 2004, Gambäck was visiting professor at Addis Ababa University, Ethiopia and has since then supervised students in Ethiopia and part time worked on using machine learning methods for rapid development of language processing tools in the project 'Language Processing Resources for Under-Resourced Languages' sponsored by the Swedish international development agency, Sida.
Prof. **Janne Bondi Johannessen** (born 1960 in Asker, Norway; Dr. Philos. University of Oslo, 1994, thesis: *Coordination. A minimalist approach*) is since 1999 Professor at the Text Laboratory, ILN, University of Oslo. The professorship entails being director of the Text Laboratory, which is a lab responsible for developing and teaching language technology. Johannessen's expertise and research interests in both theoretical linguistics and language technology make her able to plan the development of tools (databases, corpora and specialised software) that are useful for a wide research community. She was President of the North European Association of Language Technology in 2010-2011. Prof. Johannessen has been in charge of several research and technology projects, in particular on part-of-speech tagging of Norwegian and on Norwegian and dialectal corpora, incl. NorDiaSyn: Nordic Dialect Corpus and Nordic Syntax Database (Norwegian Research Council 2009-2013). She leads the Norwegian side of the NORHED collaboration with Ethiopia and jointly supervises PhD students at NTNU with Prof. Gambäck.

Dr. **Partha Pakray** (born 1982 in Burdwan, West Bengal, India; PhD Jadavpur University, Kolkata, 2013, thesis: *Answer Validation through Textual Entailment*) is currently a Post-Doctoral Fellow at NTNU funded through the ERCIM/EC "Alain Bensoussan Fellowship Programme", 2012-2013. Dr. Pakray has a wide competence in language technology and artificial intelligence and has published over 30 papers on Textual Entailment, Question Answering, Answer Validation, Machine Translation, and Information Retrieval. Three years in a row (2011-2013), Dr. Pakray has achieved the first place of all participating systems for the task in the QA4MRE@Conference on Multilingual and Multimodal Information Access Evaluation (CLEF), World Wide Research Evaluation Track.

5.2.3.3 Motivation factor (max. ¾ page A4)

The HaBiT project will employ innovative data mining and semantic technologies to assist in the efficient and effective processing of large textual data sources. Linguistic and cultural developments are amongst the major aspects of the development of a nation. One cannot talk about development in general without respecting the linguistic and cultural rights and also developing the linguistic and cultural dimensions in nations like Ethiopia. NTNU will carry out dissemination and exploitation of the project's results through many channels, such as contracted education, researcher mobility, scientific publications (including electronic), licensing of patents, and distribution of open software on the Internet. The project will strengthen existing MA and PhD programmes. This way the new language resources developed in the project wil be used. An important aspect of the HaBiT consortium is that it comprises partners that have in the past carried out the successful implementation of a joint international project.

5.2.3.3.1 Type of motivation factor

The project will help to increase the size, scope and intensity of the research works (and results).

5.2.3.3.2 Rationale of motivation factor (max. ½ page A4)

NTNU is a public university, that is, a non-profit research and teaching organization. NTNU has developed several strategies to exploit its research results. It works in close collaboration with Norwegian companies in the area, and help in this way to transfer the results to the market. Most importantly, it has been using for a long time suitable open source licenses, which allows for efficient co-operation with industry. NTNU expects to benefit academically and economically from achieving its objectives in the HaBiT project, in particular by developing novel models to leverage machine learning and advanced linguistic representations. This will open concrete opportunities in many areas/problems, such as knowledge representation for Artificial Intelligence solutions and multilingual search and retrieval. NTNU will use the advances achieved during the project to strengthen its relationships with industry, for scientific publications, and – the foremost objective of a university – by increasing the quality of its offered education in Computer Science, in particular through the supervision of student projects related to HaBiT, both at the PhD and MSc level. The non-permanent staff in the project (PostDoc and PhD student) are expected to move on to new positions, both within and outside the research community. Such mobility is one of the most powerful instruments for dissemination of new knowledge and technology. Last but not least, research results such as the ones expected in HaBiT, with

highest applicability, usability, and commercial prospects, has the potential to lead to forming spinoff companies.

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

First name	Surname	Position in project	*PhD/ post doc	Female researcher (Y/N)**
Björn	Gambäck	Norwegian team leader	-	N
Janne	Bondi Johannessen	Researcher	-	Y
Partha	Pakray	Researcher	Postdoc	N
TBD	TBD	Researcher	PhD	(Y/N)

5.2.4 Leader of Norwegian Project Partner(s)

5.2.4.1 Norwegian leader identification

Role in the project	Norwegian team leader
Degree(s)	PhD (tekn. dr.)
First name	Björn
Surname	Gambäck
Citizenship	Swedish
Position in organization	Professor of Language Technology
Work load in project (0-1.0)	**0.2**
Telephone	+47 735 933 54
E-mail	gamback@idi.ntnu.no
Personal web page*	www.idi.ntnu.no/people/gamback

*Note: *Voluntary.*

5.2.4.2 Norwegian leader's core activities in project (max. ¾ page A4)

The Norwegian team will be lead by Björn Gambäck who has been a professor at NTNU since 2008 and 2010-2012 was NTNU's principal investigator in the FP7 project PRESEMT (Pattern REcognition-based Statistically Enhanced Machine Translation, ICT-248307; Language Technology), and currently leads NTNU's activities in the NORHED collaboration with Ethiopia and NTNU's involvement in the Norwegian national language resource project CLARINO. Previously, Prof. Gambäck was a project leader at SICS for over 20 years and there coordinated a number of national and international projects, including projects both in FP5 (DUMAS: Dynamic Universal Mobility for Adaptive Speech Interfaces, IST-2000-29452; Language Technology) and FP6 (the IP EVERGROW: Ever-growing global scale-free networks: their provisioning, repair and unique functions, IST-2004-001935; Complex Systems). He has also been the principal investigator at SICS in projects such as the FP6 IP COMPANIONS (Intelligent, Persistent, Personalised Multimodal Interfaces to the Internet, IST-034434; Language Technology) and FP5 Research Infrastructure project SCHOLNET (A Digital Library Testbed to Support Networked Scholarly Communities, IST-1999-20664; Digital Libraries).

Within the HaBiT project, Gambäck will participate in the research activities related to Norwegian corpora building, annotation and processing. He will also lead this work for the Ethiopian languages and provide the liason to the NORHED and CLARINO projects.

Furthermore, Gambäck will lead the tasks concerned with word space modelling and semantic term similarity. He will also, together with Janne Bondi Johannessen, act as supervisor for the non-tenured NTNU team members, the PostDoc and the PhD student.

5.2.4.3 Norwegian leader's internationally refereed (joint) scientific publications (max. ¾ page A4)

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

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

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

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

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

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

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

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

6. PROJECT PROMOTION

6.1 Project promotion and information activities about project (max. 1 page A4)

The importance of promotion activities has been underlined by appointing a dissemination manager, Lucia Kocincová from MU, she will be responsible for coordinating all relevant activities, and especially for reassessing promotion strategies during the project lifecycle. During the first two month of the project, a detailed dissemination plan will be created, with an analytical description of the foreseen dissemination activities. The dissemination strategy of HaBiT is intended to actively explore the following main axes: dissemination channels (e.g., internet, mass-media, scientific publications), dissemination events (e.g., conferences, workshops, symposia, lectures, releases, and interviews by press agencies and media) and open-access strategies (e.g., public releases of the developed tools and resources).

The official website of HaBiT will be set-up as soon as the project formally starts and will function as the heart of the project, serving as a means of communication between the partners and containing information about HaBiT and work progress, the project partners' profile, similar projects and links to related web pages, public deliverables etc. The website is intended to be hosted on a ".eu" domain. The HaBiT website is intended to be maintained for the foreseeable future after the formal completion of the project. It is expected that near the end of the HaBiT project, the website license will be renewed for a period of 10 years, to ensure its continuous existence in order to provide a continuous link between HaBiT and any related future projects and developments.

To maximize availability of HaBiT outcomes relevant resources will be submitted into European or international resource providers like META-NET (www.meta-net.eu), CLARIN (clarin.eu), in which both the Czech and the Norwegian (in CLARINO) teams participate, or the Linguistic Data Consortium, LDC (www.ldc.upenn.edu).

Within the project lifetime it is intended to publicise as extensively as possible the results achieved via publication in both established and widely-attended specialist international conferences as well as scientific journals with high impact factors. The aim of raising the awareness of the research community for the project's outcomes will also be reinforced by granting free access to any of the aforementioned publications or journal articles, by depositing them (possibly after an embargo period) in a publicly available repository (open archive), depending naturally on the approval of the respective publishing house(s), who will be the copyright holder(s) of the publications.

7. PROJECT BUDGET

7.1 Budget rates

Rate of requested grant in the total project budget (in %)	100%
Project Promoter share of total requested grant (in %)	Masaryk university, 38.77%
Project Partner share of total requested grant (in %)	Norwegian University of Science and Technology, 61.23 %
Used indirect cost model (overheads) in the project (in %) – Project Promoter	Masaryk university, 60%
Used indirect cost model (overheads) in the project (in %) – Project Partner	Norwegian University of Science and Technology, 60%

7.2 Project budget and requested funding justification (max. 3 pages A4)

The requested funding corresponds in case of both the Project Promoter and the Project Partner to 100% of the planned budget as they are public research organizations (universities) and the proposed project does not include commercial activities.
The overall budget justification is divided into the part describing the MU (Project Promoter) budget and the NTNU (Project Partner) budget.

MU Budget

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

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

Personnel cost

The MU Personnel costs consists of actual salaries (including usual remuneration) staff assigned to the project. The amounts are computed from average income tables for the corresponding positions (associate professor, assistant professor, researcher) at the Faculty of Informatics MU in 2012 plus social security charges and other statutory costs. All the proposed staff members are MU employees at the moment. The particular personnel costs are based on the following roles and working loads of the project staff members:

- doc.PhDr. Karel Pala, CSc., 20%, associate professor (16000 CZK/month) - Principal Investigator, project team leader
- doc.RNDr. Aleš Horák, Ph.D., 20%, associate professor (16000 CZK/month) - design and analysis of annotation techniques and structures
- Pavel Rychlý, Ph.D., 20%, assistant professor (12000 CZK/month) - design and analysis of effective processing of large textual data
- Mgr. Vít Suchomel, 40%, researcher (18000 CZK/month) - implementation of tools for obtaining language data from Internet
- Mgr. Vít Baisa, 25%, researcher (11250 CZK/month) - design and implementation of specialized user interfaces
- RNDr. Miloš Jakubíček,25%, researcher (13250 CZK/month) - creation of program interfaces for big data processing
- RNDr. Vojtěch Kovář, 25%, researcher (13250 CZK/month) - design and implementation of automatic annotation techniques

- Bc. Lucia Kocincová, 30%, project administrator (11400 CZK/month) – dissemination manager, administrative processing and checking of project clerical papers

The personnel costs also include the amount of 80000 CZK for part-time job agreements for text annotation required in WP3, WP4 and WP6. This amount comprises 20000 CZK in 2014, and 30000 CZK in both 2015 and 2016. The annotation work will be presumably performed by skilled students of the Faculty of Arts and Faculty of Informatics, MU. The planned price per hour for the annotation is 150-160 CZK, which corresponds to 125 hours in 2014 and 190 hours in 2015 and 2016. The planned costs include also the social security charges and other statutory costs related to the agreements.

Travel and subsistence

The travel and subsistence costs include:

- active participation of MU research team members at scientific international conferences designated to the computer processing of natural languages, such as ACL, COLING, TSD, CICLING, LTC, or Grammar and Corpora. We plan in average 3 active participations per year with the cost of 30000 CZK per each participation. This cost includes also the respective conference fees.
- visits of research team members to the Project Partner institution. We suppose one visit of 2 researchers for 3-5 days with the overall cost of these visits of 80000 CZK per year.

New or second hand equipment

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

Consumables and supplies

We plan standard consumables for computer processing project work such as disk storage for the project data, or printer toner for the project materials printing in the amount of 20000 CZK per year.

Other costs

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

Indirect cost (overheads)

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

NTNU Budget

The budget for NTNU (see Annex IV) primarily consists of direct salary costs and the related indirect costs. In terms of the cost structure given in the Guide for Applicants:

Personnel cost: Given in terms of actual salaries (including a 4.5% average annual increase) of the non-permanent staff assigned to the project. The salaries of the permanent staff is covered by their regular appointments and hence not included in the project budget. That is why in the Annex XI for NTNU only one contact agreement is attached for Partha Pakray at the PostDoc position. The non-permanent staff salaries are based on the standard salary levels for PostDocs and PhD student positions at the Department of Computer and Information Science, NTNU, and follow the Norwegian government's salary scale (lønnstrinn, Ltr), with a PostDoc currently being at Ltr 57 (473,400 NOK/year in gross pay, including social

costs and compulsory governmental pension fund deduction) and a PhD student at Ltr 51 (427,900 NOK).

Travel and subsistence: Have been calculated in accordance to the NTNU standard for travelling in FP7 - namely with €900 per trip, plus a daily allowance of €250 for 5 days per trip - and with one trip each per year for the four researchers involved (or 35/12 trips per person, to be exact, given the length of the project, which is 35 months).

Indirect costs: Computed with a flat rate of 60% of the total direct eligible costs.

No other project-related costs are forseen at this time (this then includes preparatory costs, equipment, consumables, subcontracting, etc. - none of which is foreseen). However, one financial audit will be carried out, for a projected cost of €2,000.

8. MANDATORY ANNEXES

8.1 Overview of annexes required to project proposal

No.	Annex	Mandatory online submission format
I.	Documentation evidencing trade licence or any other requested authorization, for SMEs Research and Development activities must be listed in the trade licence (Except public universities or colleges or higher education institutions, public research organizations)	pdf
II.	Authorization of research organisation status (only Norwegian Project Partners)	pdf
III.	Draft of Partnership agreement	pdf
IV.	Project budget	xlsx
V.	Statutory declarations (Project Promoter and each Czech Project Partner)	pdf
VI.	Consent to processing of personal data (Project Promoter and each Project Partner)	pdf
VII.	CVs of key project personnel (Europass format)	pdf
VIII.	Project abstract for evaluators (you may copy 1.14)	docx
IX.	Statutory declaration on category of enterprise (only for SMEs)	pdf
X.	Authorisation document authorising another official authorised to sign on behalf of the responsible official (if applicable) – no template	pdf
XI	Draft of labour contract (if applicable) for newly hired persons – no template	docx

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

9. OTHER

References

Dovudov, G., Suchomel, V., Šmerk, P. Towards 100M Morphologically Annotated Corpus of Tajik. Aleš Horák, Pavel Rychlý (Eds.). Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN, 91-94., 2012.

Erk, Katrin, Diana McCarthy, and Nicholas Gaylord. 2013. Measuring word meaning in context. *Computational Linguistics*, **39**(3): 501–544.

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

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

Gyöngyi, Zoltán, Hector Garcia-Molina, and Jan Pedersen. 2004. Combating Web Spam with TrustRank. Proceedings of the International Conference on Very Large Data Bases 30: 576.

Hudson, Grover. 1999. Linguistic analysis of the 1994 Ethiopian census. *Northeast African Studies,* **6**(3):89-107.

Kanerva, Pentti, Jan Kristoferson, and Anders Holst. 2000. Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, page 1036, Philadelphia, Pennsylvania. Erlbaum.

Kilgarriff, Adam. 2000. I don't believe in word senses. *Computers and the Humanities*, **31**(2):91–113.

Kilgarriff, Adam and Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Comput. Linguist.* 29, 3 (September 2003), 333-347.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. The Sketch Engine *Proc. Euralex.* Lorient, France, pages 105-116. Reprinted in *Lexicology: Critical concepts in Linguistics* Hanks, editor. Routledge, 2007.

Lau, Jey Han, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 591-601. Association for Computational Linguistics.

Lynum, André, Erwin Marsi, Lars Bungum, and Björn Gambäck. 2012. Disambiguating Word Translations with Target Language Models. In: *Text, Speech and Dialogue: 15th International*

Conference, TSD 2012, Brno, Czech Republic, September 3-7, 2012, Proceedings. Springer Science+Business Media B.V.

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

Pomikalek, Jan, Rychly, Pavel, and Kilgarriff, Adam. 2009. Scaling to Billion-plus Word Corpora. Advances in Computational Linguistics. Special Issue of *Research in Computing Science* Vol 41, Mexico City.

Navigli, Roberto. 2009. Word Sense Disambiguation: a survey. *ACM Computing Surveys*, **41**(2):1–69.

Reisinger, Joseph and Raymond J. Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117, Los Angeles, California.

Rychlý, P. 2007. Manatee/Bonito - a Modular Corpus Manager, Proceeding of the RASLAN Workshop, Masaryk University, pp. 65-70.

Schroeder, Stan. 2010. Half of Messages on Twitter Aren't in English [STATS], Feb. 24, 2010. http://mashable.com/2010/02/24/half-messages-twitter-english/

Schütze, Hinrich. 1998. Automatic word sense discrimination. Computational Linguistics, **24**(1):97–123

Suchomel, Vít and Pomikálek, Jan. Efficient Web Crawling for Large Text Corpora. In *Proceedings of the seventh Web as Corpus Workshop* (*WAC7*). Lyon, 2012.

Gambäck, Björn, Fredrik Olsson, Atelach Alemu Argaw, and Lars Asker. 2009. Methods for Amharic Part-of-Speech Tagging. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 104-111, Athens, Greece. First Workshop on Language Technologies for African Languages.

Karlgren, Jussi, Björn Gambäck, and Pentti Kanerva. 2002. Acquiring (and Using) Linguistic (and World) Knowledge for Information Access. *AI magazine*, **23**(4):101.

Lynum, André, Erwin Marsi, Lars Bungum, and Björn Gambäck. 2012. Disambiguating Word Translations with Target Language Models. In: *Text, Speech and Dialogue: 15th International Conference, Brno, Czech Republic*. Springer Science+Business Media B.V.

Eyassu, Samuel and Björn Gambäck. 2005. Classifying Amharic News Text Using Self-Organizing Maps. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, Michigan. Workshop on Computational Approaches to Semitic Languages.