LetsMT!

Platform for Online Sharing of Training Data and Building User Tailored MT

www.letsmt.eu

Project no. 250456

Annual Public Report

Version no. 1.0

15/11/2011

Contents

1. CHALLENGES ADDRESSED

In recent years, statistical machine translation (SMT) has become the leading paradigm for machine translation. SMT systems are built by analysing huge volumes of parallel corpus and learning translation models from this data.

The quality of SMT systems largely depends on the size of training data. Since the majority of parallel data used for training systems is in major languages, SMT systems for larger languages are of much better quality compared to systems for smaller languages. This quality gap is further deepened due to the complex linguistic structure of many smaller languages. Languages like Latvian, Lithuanian, and Croatian (to name just a few) have a complex morphological structure and free word order. To learn this complexity from corpus data, much larger volumes of training data are needed. Current systems are built on the data accessible on the web, but this is just a fraction of all parallel texts. The first challenge of the LetsMT! project is to support smaller languages with access to parallel texts. Most parallel texts still reside in the local systems of different corporations, public and private institutions, and on the desktops of individual users.

Another challenge which is addressed in the LetsMT! project is the complexity of training SMT systems. Currently users who would like to train SMT on their parallel texts can either contract a service provider specializing in custom SMT solutions, or install the Giza++ and Moses SMT toolkit and build a SMT system themselves. Both solutions are expensive and deter users from trying and using SMT.

2. PROJECT OBJECTIVES

To fully exploit the huge potential of existing open SMT technologies, the main objective of the LetsMT! project is to build a user-friendly, innovative online collaborative platform for data sharing and building custom machine translation (MT) systems. LetsMT! is a collaborative platform that thrives on resources contributed by its users. It will be a major breakthrough regarding the availability of parallel language resources. Consequently, this will also be a major breakthrough for MT services of good and acceptable quality for less-covered languages where the current MT systems perform poorly due to limited availability of training data.

This platform will support the upload of public as well as proprietary MT training data and the building of multiple MT systems, public or proprietary. LetsMT! will build on existing state-of-the-art SMT methods which will be applied to data supplied by users, increasing the quality, scope, and language coverage of machine translation.

LetsMT! will provide a platform that supports the following features:

- Uploading of parallel texts for users that will contribute their own content
- Access to a directory of web and offline resources gathered by LetsMT! as well as links provided by users to other content sources not included in the LetsMT! repository
- Automated training of SMT systems from specified collections of training data
- Building of custom MT engines from a selected pool of training data, for larger donors or paying customers
- Building of custom MT engines from proprietary non-public data
- MT evaluation facilities

The solution will deliver the following core functions:

- **A website for uploading** parallel corpora and building specific MT solutions
- **A website for translation,** where source text can be typed and translated
- **A translation widget** provided for inclusion into websites to translate their content
- **Browser plug-ins** that will provide the quickest access to translation
- **Integration in CAT tools** and other applications

3. SUMMARY OF ACTIVITIES

Project LetsMT! is divided between 6 major activities with supplementing subtasks. Each activity is dedicated to a specific element of the LetsMT! platform, such as infrastructure building, data collecting, training, etc.

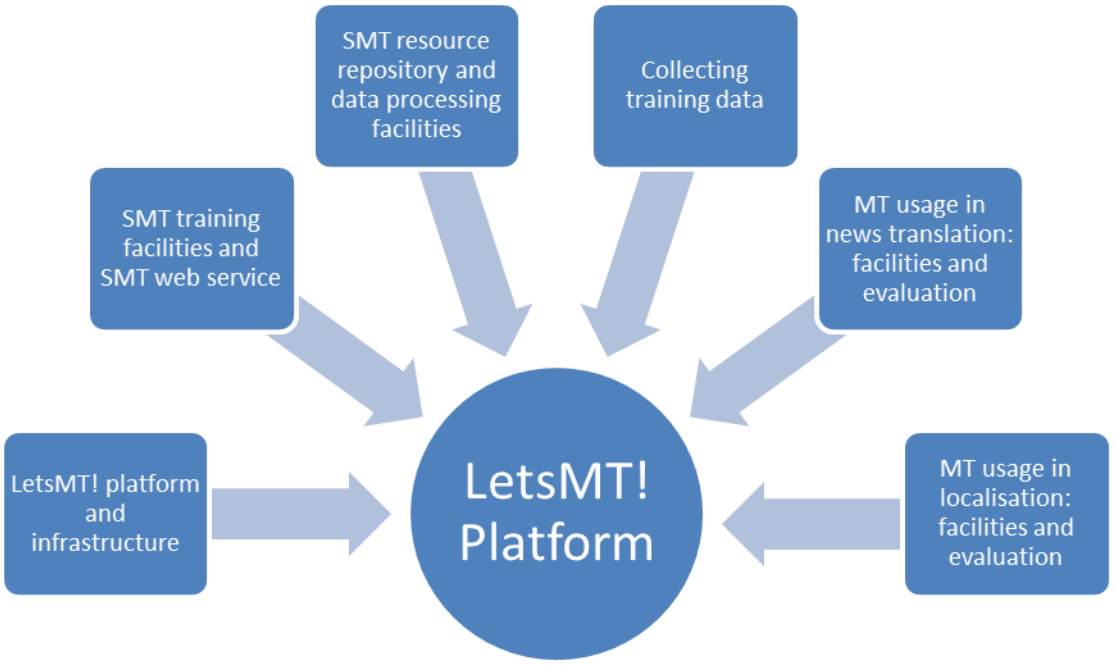


Figure 1: LetsMT! structure

LetsMT! platform and infrastructure. Work on the LetsMT! platform and infrastructure is the core activity within the LetsMT! project. The LetsMT! platform includes modules for sharing of SMT training data, SMT training and running, use in a news translation scenario, and use in a localisation usage scenario.

The beta versions of all the main modules have been built. The project Consortium has developed a common platform and supporting software infrastructure that provides the core functions necessary to integrate the modules of the LetsMT! platform. The supporting software infrastructure includes: the LetsMT! website, an API for external systems, User Management and Access Rights Control, Application Logic, an MT web page where users can try trained MT systems, etc.

It is obvious that hosting the LetsMT! platform requires a lot of computing capacity. The Project Consortium, instead of buying servers, intends to lease capacity. It is economically efficient and

will provide flexibility in adding new resources as necessary. During the analysis of detailed requirements, it was discovered that operating the LetsMT! platform on AWS (Amazon Web Services) was the most economically efficient option. It is planned to deploy the LetsMT! platform completely within the AWS, as this is a well established solution. The AWS cloud provides a reliable and scalable infrastructure for deploying web-scale solutions. Alternative cloud computing suppliers may be selected if AWS fails to meet the requirements of the LetsMT! platform. The LetsMT! platform also can be deployed on a local server infrastructure.

So far all LetsMT! platform implementation and integration tasks have progressed well and according to the schedule. The project consortium has successfully achieved all the important milestones and resolved technical and scientific challenges. As a result, the Beta version of the LetsMT! platform has been publicly accessible since August 2011. The newest public LetsMT! Beta version can be accessed and evaluated at https://letsmt.eu.

System development, testing, deployment, and evaluation are ongoing processes, and we are expecting several more releases before the final release in February 2012.

SMT resource repository and data processing facilities. The backend of the LetsMT! platform includes a modular resource repository. Figure 2 illustrates the general architecture of the software. Its design emphasizes possibilities of running the system in a distributed environment which makes the system suitable for scalable cloud-based solutions. Communication between the web-frontend and the individual modules is handled by secure web service connections. A central database handles metadata information in a flexible key-value store that supports schema-less expandable information collections. The physical data storage can be distributed over several servers to reduce bottlenecks when transferring large data collections. Data collections can be stored using a version-controlled file system that supports data recovery and history management in a multi-user environment. The repository provides essential features for importing documents to the LetsMT! platform. Documents are converted, and sentences in translated documents are aligned automatically. The software is connected to a high-performance cluster that can execute various jobs with connection to the data stored in the repository, for example, the import and alignment of jobs. A cloud-based cluster enables scalability of the system according to the needs of the platform. The repository software is fully integrated in the current LetsMT! platform and can easily be extended with additional modules.

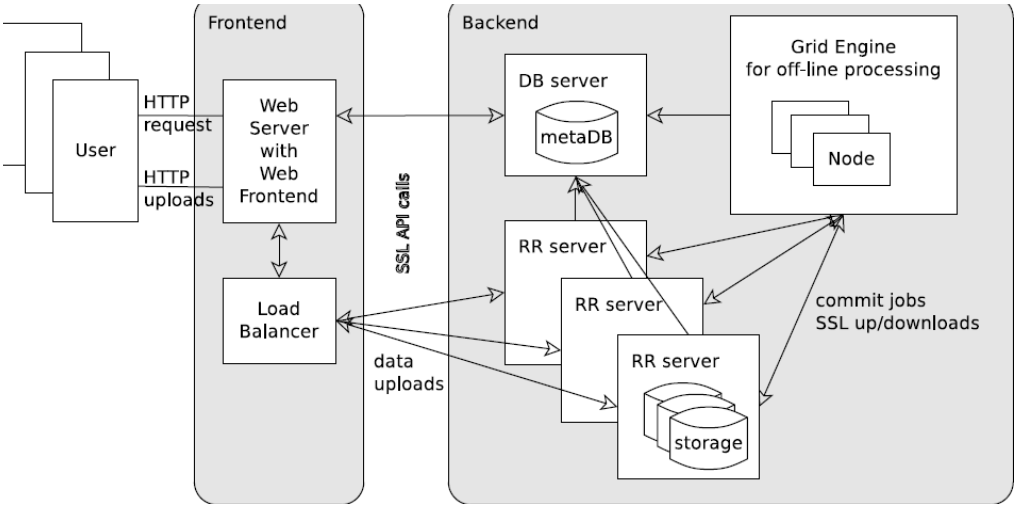


Figure 2: General architecture of a Resource Repository

SMT training facilities and the SMT web service. Users of the LetsMT! platform may select training resources from the SMT Resource Repository and train tailored SMT systems using the selected training resources. SMT training facilities include the following features: a user interface for resource selection and system training, integration with user authentication and access rights module, integration with SMT Resource Repository, simultaneous and effective execution of resources consuming training tasks, an interface providing information about running training tasks, progress, status, etc.

The SMT training facility and web service is built on top of the Moses machine translation toolkit. Originally developed at the University of Edinburgh in 2007, Moses has since undergone a great deal of evolution. Many new features have been added, improving translation quality and keeping Moses up to date with the cutting edge of MT research. While of great importance, translation quality, however, is not the only aspect to have been worked on. SMT is extremely computationally demanding. Literally millions of options must be searched through in order to translate a single sentence, and the amount of data required to do so far outstrips the resources of and average desktop computer. Therefore, much research has been conducted on how to speed this process up and reduce the computational resources needed for translation.

Translation is only a part of what the Moses SMT Toolkit can do, though also included with it are the tools to train new translation systems. As with the actual method of translating, huge amounts of work have gone into training systems to yield better translations, as well as making the training process itself less resource intensive. The process of training a translation system is very in depth and intricate, but that too is handled by the toolkit.

Despite all the work that has gone into developing Moses, there are a few features required by the LetsMT! platform that Moses did not have. Having been conceived in academia, the focus of Moses has generally been towards features required by researchers and academics., However, the environment in which it will operate in the LetsMT! platform is very different. Developing Moses to support these new requirements is the main focuses of project activities and the work done in doing so is detailed below.

End users will want a service that delivers translations in a fast, interactive manner. Translating sentences requires a large amount of data, and waiting for this to be loaded each time would make the interactive user experience impossible to deliver. This has been addressed by the implementation of a version of Moses which runs on a background server, can be given sentences to translate interactively, and returns the translations quickly — without having to wait for the whole system to load up.

Users of the LetsMT! platform will also be translating between many different pairs of languages, and therefore, separate background processes for each pair would be impractical. This has been countered by allowing Moses to simultaneously have multiple translation systems in memory and by providing the language to translate into along with the sentence.

Modern computers are increasingly geared towards doing many things in parallel, instead of doing them sequentially. In order to make the best use of available resources, Moses must be able to translate sentences in parallel. This feature, called ‘multi-threading’ has been integrated into Moses and enables it to deliver many translations in a fraction of the time compared to doing them one after another.

Other features such as being able to leverage new data without having to retrain the entire system have also been implemented and are in the process of being integrated with the rest of the platform. Methods for improving the fluency of translations using many billions of words of text are also in active development.

The LetsMT! platform is a great example of an EU project putting cutting edge technology to great use for the wider public, and as it does so, feeding back improvements to the academic community from where its ideas originate.

Collecting the training data. The right training data for the LetsMT! platform is essential for good translation results. Therefore, the collection of training data is an important task.

The aim of the LetsMT! project is to collect data from both general language and from different subject domains. A special effort is being made by two of the project partners to collect business and finance news and localisation texts, mostly from the IT domain. Other subject domains are interesting for the project, so the partners focus on finding text providers with general language texts, in addition to domain specific texts.

The initial training corpora focused on Croatian, Czech, Danish, Dutch, Latvian, Lithuanian, Polish, Slovak, and Swedish, but other languages were also presented. We still focus on these original languages, though other languages might occur as one of the languages of a parallel corpus.

During the first year, lots of publicly available data was collected for a wide range of languages with focus on the project partner languages. The established goals were by far achieved for all languages except for Croatian, a situation that was foreseen and is now being mitigated.

This year the Project Consortium has concentrated on identifying new text providers and potential future users of the LetsMT! system.

For business and finance news, the Project Consortium uses a list of the largest companies from the involved countries to automatically harvest the newest parallel texts from these companies, and therefore, the collection is steadily growing.

The collection of parallel texts from the general language and from other subject domains is being advanced by making contacts at different levels. At the international level, the Project Consortium is in contact with TAUS (which has one of the largest repositories of parallel corpora) and with various EU institutions and projects. At the national level, project partner - Tilde has made a cooperation agreement with the National Library of Latvia. The partner University of Zagreb has made contact with several translation and localisation companies that are interested in the project and two of these have committed themselves to become text providers. The project partner- Moravia has made contact with the Slovak national corpus, but due to IPR problems their corpora cannot be used outside the institute. The project partner- Uppsala University contacted two institutes at Stockholm University who might be interested in using LetsMT!. The project partner — Kobenhavns Universitet contacted several potential text providers and has received acceptance from a company that write press releases in the EU languages and from at least 12 companies with annual reports. Furthermore, Kobenhavns Universitet has started co-operation with a translation centre connected to Kobenhavns Universitet about texts in the domain of university administration.

The LetsMT! project has been presented at various events both at the national and international levels in order to spread knowledge and create awareness of the project in

general and of the need for data in particular. Generally, the responses to the presentations are very positive, but IPR constitutes a challenge to the project. It turns out that some of the texts originally identified cannot be used outside the company/institution and thus cannot be uploaded to an external server like LetsMT!. Others can only be uploaded for private use and will not become publicly available on the LetsMT! platform. However, some of the contacts mentioned above are ready to sign a text provider agreement and others will follow soon.

Usage Scenarios

In particular, two specialized usage scenarios are supported by the LetsMT! platform: 1) machine translation of financial news, and 2) translation process in localization industry companies.

1. MT usage in news translation: facilities and evaluation. During the reporting period, the Project Consortium has implemented the widget and browser (Mozilla, Internet Explorer) plug-ins of the LetsMT! platform.

The business scenario was developed in which the use of the widget is described. The aim for the business scenario is to provide translated business and financial information through several facilities. There are two scenarios which are currently being investigated, a free and a paid, professional service. Free services will attract a broad audience of users with an interest in business related news and financial background information. The content will be information with a high latency, background information of local stock markets, local listed company information and comments. For low latency and emerging news, users can subscribe to a paid service. The targeted users are professionals and individuals that are interested in local and international breaking news and financial information. At the moment, the LetsMT! widget is integrated into SemLab's business and financial news website www.newssentiment.eu for trial and evaluation purposes. The system is being tested on the website to ensure positive results in dissemination and exploitation activities through other (financial news) websites.

2. MT usage in localisation: facilities and evaluation. Professional users need MT services integrated in their working environment. Translators use CAT (Computer Aided Translation) tools (such as SDL Trados and MemoQ) in everyday activities. One of the prerequisites conditioning successful localisation scenario implementation is, without any doubt, the integration with CAT tools. In order to fulfil these requirements, the Project Consortium has developed a LetsMT! platform plug-in for SDL Trados Studio 2009 which allows for the use of the LetsMT! platform during translation process and experimentation on the evaluation of an English-Latvian SMT system applied to an actual localisation assignment. Results of the evaluation have been described in a paper which has been presented at the EAMT 2011 conference and JEC 2011 workshop. The paper shows that such an integrated localisation environment can increase the productivity of localisation by 32.9% without critical reduction in quality.

4. DISSEMINATION

4.1. Dissemination Plan and Visual identity

The dissemination plan was created on May 2010 defining all necessary means of dissemination of information about LetsMT!: target groups, dissemination channels, general visual identity, public web-site design, posters, flyers, t-shirts, participation on key conferences,

LetsMT! events, public showcases, web presence, appearance in scientific journals and conferences, issuing announcements, social-networks presence, etc.

The visibility of the LetsMT! project has a unique visual identity (logo) that helps recognise the project among similar projects. The visual identity was designed and applied to all possible and even non-conventional channels of dissemination such as the public web site, presentation template, leaflets, posters, t-shirts, video lectures, Wikipedia articles, and social networks.

4.2. Web site

The LetsMT! website (www.project.letsmt.eu/) is one of the main project communication tools. The LetsMT! website has undergone a major change in 2011 since the primary LetsMT! address (http://www.letsmt.eu/) has been reserved for access to the LetsMT! services and tools website. The LetsMT! project website has moved to address http://project.letsmt.eu. The alternative address http://www.letsmt.org points to the project website, while http://www.letsmt.com points to the LetsMT! services and tools website.

4.3. Dissemination to the scientific community and the industry

It is foreseen to promote the project's innovative technologies in the framework of national and international conferences, exhibitions and events most attractive to the language industry (such as, localisation and/or translation industry events) and scientific community.

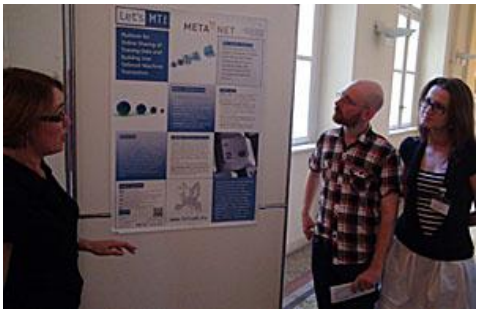
LetsMT! participation in conferences and workshops (with short reports from events in reverse chronological order):

TAUS user conference 2011. Andrejs Vasiļjevs gave a presentation titled "Technologies for smaller languages" at the Call for Ideas session where he demonstrated how LetsMT can train a MT system from user provided data.

Localization World Silicon Valley 2011. Indra Sāmīte presented LetsMT! at the system demonstration session of the conference. Andrejs Vasiļjevs presented work on LetsMT! at the panel session "MT: Free ride is over". Tilde provided continuous demonstrations of the LetsMT! platform at the conference exhibition booth.

MT Summit XIII. The LetsMT! platform was demonstrated at the MT Summit XIII, the largest event in the MT community. It was held in Xiamen, China, from 19 to 23 September 2011. The LetsMT! platform was presented in the system presentation session by Raivis Skadiņš under the title "LetsMT!: Cloud-Based Platform for Building User Tailored Machine Translation Engines". The LetsMT! platform has received great interest from conference participants, especially practitioners in this field.

CLARA Career Course. CLARA is the Initial Training Network in the Marie Curie Actions. Its Career Course on Product Planning for Next Generation Information Access Technology Solutions was held in the Centre for Advance Academic Studies, University of Zagreb in Dubrovnik, Croatia from 20 to 23 September. The whole course was targeted at early stage and experienced researchers in Language Resources and Technology. Within this



course, different applications of LRT were presented starting with whole life cycle and finishing with scientific results. The LetsMT! project was explained and presented by Inguna Skadiņa.

SlaviCorp2011 conference. The 2nd Slavic Corpora Conference, SlaviCorp2011 was held in the Centre for Advance Academic Studies, University of Zagreb in Dubrovnik, Croatia from 2011-09-12 until 2011-09-14. Between different projects that deal with the usage of corpora of Slavic languages, the research within LetsMT! project was presented.

META-FORUM 2011. The central disseminating event of the META-NET community in 2011, META-FORUM 2011 took place in Budapest on 27/06/2011 and 28/06/2011. During these two days of the conference, densely packed with a number of parallel activities, — oral presentations, poster presentations, software demonstrations, an exhibition of European projects was organized. The LetsMT! project was presented and it attracted quite an interest since the first on-line system for building tailor-made SMT systems was demonstrated.

NooJ2011 conference. The NooJ community organizes its conferences every year in May or June. This year the NooJ2011 conference took place in Centre for Advanced Academic Studies in Dubrovnik, Croatia. Within this three day conference that started on 12/06/2011 an exhibition of European projects was organized where LetsMT! project was presented.

Localization Word 2011 Conference & Exhibition. The event took place in Barcelona, Spain, on 14-16 June 2011. It is a conference and networking organization dedicated to the language and localization industries. This year, it gathered more than 550 industry professionals from all over the world. In this event LetsMT! was presented and attracted the conference participants' interest. One of the main focus points at the Conference was also the recent development in the MT field as well as MT experiences at the leading IT companies.

EU projects exhibition at EAMT2011. The European Association for Machine Translation (EAMT) organizes its yearly conferences regularly in May. This year the venue was the Faculty of Arts, Katholieke Universitet Leuven in Belgium. This two day conference started on 30/05/2011, and there was an exhibition of European projects related to machine translation on 31/05/2011. Since this audience is considered to be a natural lieu for LetsMT! services, project presentation was held by Raivis Skadiņš.

FLaReNet Forum 2011. The FLaReNet Forum 2011 took place in Venice from 26/05/2011 to 27/05/2011. It gathered numerous representatives from the LRT community from all over Europe and it can be considered the largest event in Europe in this year so far. LetsMT! was

presented there and attracted considerable interest. Andrejs Vasiļjevs also had a presentation *How to get more data for under-resourced languages and domains?* where LetsMT! project was presented.

ICT-PSP info week in Zagreb. The Central State Office for eCroatia organized the ICT-PSP info week where ICT-PSP projects with Croatian partners were presented as examples of successful on-going activity. Among seven projects, LetsMT! was presented as one of the most interesting and dealing with much needed topic that would lead to a solution for machine translation regarding under-resourced languages and domains. The video clip from TV broadcast eHrvatska no. 78 reporting on this presentation can be seen at our video lectures page or directly from YouTube.

4.4. Publications of the project team, November 2010—November 2011

- Skadiņš R., Puriņš M., Skadiņa I., Vasiļjevs A., Evaluation of SMT in localization to under-resourced inflected language, in Proceedings of the 15th International Conference of the European Association for Machine Translation EAMT 2011, p. 35-40, 30-31 May 2011, Leuven, Belgium.

- Vasiļjevs, A., Skadiņš, R., & Tiedemann, J. (2011). LetsMT!: Cloud-Based Platform for Building User Tailored Machine Translation Engines. In Proceedings of the 13th Machine Translation Summit (pp. 507-511). Xiamen, China.

- Andrejs Vasiļjevs, Raivis Skadiņš and Inguna Skadiņa. Towards Application of User-Tailored Machine Translation. Proceedings of the Third Joint EM+/CNGL Workshop "Bringing MT to the User: Research Meets Translators", pp. 23-31.

5. LetsMT! CONSORTIUM AND CONTACT PERSONS

TILDE SIA
Vienības gatve 75a, Riga, LV1004
Latvia
Project Coordinator: Dr. Andrejs Vasiļjevs
e-mail: andrejs@tilde.lv
URL: http://www.tilde.eu

UNIVERSITY OF EDINBURGH
Old College, South Bridge Edinburgh EH8 9YL
UK
Contact person: Dr. Philipp Koehn
e-mail: philipp.koehn@ed.ac.uk
URL: http://www.statmt.org/ued/

UNIVERSITY OF ZAGREB
FACULTY OF HUMANITIES AND SOCIAL SCIENCES
Ivana Lučića 3, ZAGREB, HR-10000
Croatia
Contact person: Prof. Marko Tadić
e-mail: marko.tadic@ffzg.hr
URL: http://hnk.ffzg.hr/default_en.htm

KØBENHAVNS UNIVERSITET
Njalsgade 80, DK-2300 , Copenhagen S
Contact person: Lene Offersgaard
e-mail: leneo@hum.ku.dk
URL: http://www.humanities.ku.dk/

UPPSALA UNIVERSITY
P.O. Box 635, S-751 26 Uppsala,
Sweden
Contact person: Dr. Jörg Tiedemann
e-mail: jorg.tiedemann@lingfil.uu.se
URL: http://www.lingfil.uu.se/lingfil_eng/

ZOOROBOTICS – (TRADING NAME SEMLAB)
Zuidpoolsingel 14a, 2408ZE Alphen a/d Rijn
The Netherlands
Contact person: Bram Stalknecht
e-mail: stalknecht@semlab.nl
URL: http://www.semlab.nl/pages/index.jsp?id=170

MORAVIA IT A.S.
Hilleho 4, 602 00 Brno
Czech Republic
Contact person: Marta Buchtova
e-mail: MBuchtova@MoraviaWorldWide.com
URL: http://www.moraviaworldwide.com/