# LetsMT!

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

**www.letsmt.eu/**

**Project no. 250456**

## Deliverable D1.6
## Revised Functional Specification

**Version No. 1.0**

**29/02/2012**

## Document Information

| | |
|---|---|
| Deliverable number: | D1.6 |
| Deliverable title: | Revised Functional Specification |
| Due date of deliverable according to DoW: | M24 – February, 2012 |
| Actual submission date of deliverable: | February 29, 2012 |
| Main Author(s): | Raivis Skadiņš (TILDE), Kārlis Goba (TILDE), Māris Puriņš (TILDE), Valters Šics (TILDE), Normunds Lauva (TILDE), Jörg Tiedemann (UUP), |
| Participants: | David Filip (MOR) , Željko Agić (FFZG) , Thomas Dohmen (SEM), Lene Offersgaard, Tomas Hudik (MOR) |
| Reviewer | UEDIN |
| Workpackage: | WP1 |
| Workpackage title: | LetsMT! platform and infrastructure |
| Workpackage leader: | TILDE |
| Dissemination Level: | PU |
| Version: | 1.0 |
| Keywords: | Machine translation, functional specification, user story, use case, LetsMT!, LetsMT |

## Executive summary

This document describes the functional requirements, logical and infrastructure design of the elaborated LetsMT! system.

# Table of Contents

# 1 Introduction

This document describes the functional requirements, logical and infrastructure design of the LetsMT! system.

## 1.1 *Preface*

This document is based on the results of the previous project activities, e.g. deliverable D1.1 “Report on Requirements Analysis”, D2.1 “Specification of data formats allowed”, “D2.2 SMT resource repository and data processing facilities ready for integration”, “D1.3 Hardware infrastructure”, “D3.1 Adapted Moses toolkit”, “D3.2 Training Methodology”, “D3.3 SMT training facilities ready for integration”, “D3.4 SMT web service ready for integration”, “D3.5 SMT Multi-Model Repository ready for integration”, “D5.1 Widget for web page translation / Browser plug-in for web page translation”, “D6.1 Integration in CAT tools” and the project’s Annex I - “Description of Work” document.

## 1.2 *Objectives*

The aim of the Functional Specification document is to describe the function of the LetsMT! system and the high-level system design necessary to support further development, implementation and testing of the software.

The document describes the functional and non-functional requirements for the LetsMT! system in detail sufficient for system architects, developers, and testers to proceed with design, development and testing activities.

This document consists of two major parts. The first part (Chapter 2) contains the conceptual design and user stories/requirements for the system. The second part (Chapter 3) describes the logical and infrastructure design of the system.

# 2 Conceptual Design

This chapter is based on current development system version developed according to the project deliverable D1.1 “Requirements analysis” findings, deliverable “D1.2 Functional Specification”, feedback from LetsMT! platform beta testers and the experience of the LetsMT! Consortium members.

## *2.1 System overview*

In recent years statistical machine translation (SMT) has become a major breakthrough in machine translation (MT) development providing a cost effective and fast way to build MT systems. This development was particularly facilitated by the MT training and decoding tool Moses. Another factor for facilitating the development of MT for many languages was the EU translation corpus and other parallel data available on the Internet. The EuroMatrix project has demonstrated how open source tools and publicly available data can be used to generate SMT systems for all language pairs of the official EU languages.

However, these achievements do not fulfill all expectations regarding the application of available SMT methods. The quality of an SMT system largely depends on the size of training data. Obviously, the majority of parallel data is in widely-used languages (e.g. English, German and some others). As a result, SMT systems for these languages are of much better quality compared to systems for under-resourced languages, i.e. languages with scarce linguistic resources. Current systems are built on the data accessible on the web, but it is just a fraction of all parallel texts. The majority of valuable parallel texts still reside in the local systems of different corporations, public and private institutions, and desktops of individual users.

Another obstacle preventing wider use of MT is its general nature. Although free web translators provide reasonable quality for many language pairs, they perform poorly for domain and user-specific texts. Current free systems cannot be adjusted for particular terminology and style requirements. Large international corporations contract MT companies like Language Weaver to adapt translation systems for their particular needs. However, this costly process is not accessible to smaller companies or the majority of public institutions. This prevents a large segment of the EU population from using existing MT solutions to get access to online information.

Specifically regarding application in the localization and translation industry, a huge number of parallel texts in a variety of industry formats have been accumulated, but the application of this data does not fully utilize the benefits of the modern MT technology. At the same time, this industry is experiencing growing pressure to increase efficiency and performance, especially due to the fact that the volume of texts to be translated grows at a higher rate than the availability of human translation, and translation results are expected in real-time. At present, the integration of MT in localization services is in its early stages, and the cost of developing specialized MT solutions is prohibitive to most players in the localization and translation industry. The quality of the generic MT offerings provided for free is too low to reap any efficiency gains in the professional localization industry setting. The same problem is faced by online information providers. They provide information mostly in the larger languages because the cost of human translation into smaller languages is prohibitively high and the quality of existing MT solutions for smaller languages is unsatisfactory.

There are a number of freely available translation systems on the Internet; however majority of those allow users the options to choose between domain–specific systems. Also users cannot directly influence SMT systems by uploading corpora in user's possession for achieving better quality of translations for their needs.

To fully exploit the huge potential of existing open SMT technologies and the huge potential of user-provided content, we have built **an innovative online platform for data sharing and MT building**. The LetsMT! project extends the use of existing state-of-the-art SMT methods enabling users to participate in data collection and MT customization to increase the quality, scope, and language coverage of MT.

Localization and translation industry business and translation professionals can use the LetsMT! platform for uploading their parallel corpora in the LetsMT! website, building custom SMT solutions from the specified collections of training data, and accessing these solutions in their productivity environments (typically, various CAT tools).

### 2.1.1 System concept

LetsMT! is a cloud-based platform that gathers public and user-provided MT training data and generates multiple MT systems by combining and prioritizing this data. Authenticated users with appropriate permissions upload corpora using a simple web interface. User creates SMT system, choose corpora to use for the particular system training and initiate the training. When training is successfully completed, trained SMT systems can be used for translation in several ways:

- through the web portal,
- through integration in computer-assisted translation (CAT) tools,
- through an API,
- through a widget provided for free inclusion in a web-page,
- through browser plug-ins.

### 2.1.2 User Groups

LetsMT! system is targeted to support specific organizations and user groups needs for machine translation.

**Table 1 Main needs of LetsMT! user groups.**

| User group | Main needs |
|---|---|
| Individual Internet users | • Free MT for under-resourced languages or specific domain<br>• Access to multilingual news in business and finance |
| Localization and translation industry companies | • Increase in productivity of translation work through application of translation tools and MT<br>• MT adapted to terminology and stylistic requirements of the particular project |

| User group | Main needs |
|---|---|
| Web developers | - Provide access to websites for multilingual user community using widgets |
| University education<br>and research community | - Easy-to-use infrastructure for training and research on SMT<br>- Parallel corpus data, especially for under-resourced languages and domains<br>- Easy to use infrastructure for experiments in SMT training on different data<br>- Parallel corpus sharing platform |
| Translation automation solution developers and providers | - Integration of new MT directions into translation automation tools and solutions |

### 2.1.3 User Characteristics

During interviews, conducted in order to gather user base requirements, it was discovered that different types of users have different requirements for the LetsMT! platform.

**Table 2. LetsMT! user characteristics.**

| User type | Description |
|---|---|
| Individual Internet users | |
| Anonymous Internet User | A person browsing the Web; who wants to test SMT or hopes to receive better quality machine translation than provided by other publicly available MT systems like Google Translate.<br>This user does not want to make an effort, just use the service to either translate a desired text or to compare with other SMT engines.<br>This user is not expected to contribute corpora. |
| Business news reader | Wants to read international financial news in “smaller” native EU languages. Foreign language skill is limited; therefore LetsMT! is used to gist English news in local language and vice versa. Requirements for translation quality are not very high, however it is expected that terminology and the essence of news will be translated correctly and understandably.<br>This user is not expected to contribute corpora. |

| User type | Description |
|---|---|
| MT enthusiast | A person who is aware of various available MT tools and technologies.<br>Attraction to LetsMT! is based on the possibility to directly influence the SMT engine and review the SMT system training and translation logs.<br>This MT enthusiast is ready to contribute corpora in order to achieve better SMT results. Receiving praise for contributing or belonging to some elite group would be considered a bonus.<br>Might submit poor quality comparable corpora. |
| Localization and translation industry company | |
| Translator | Needs to use SMT because of organization's workflow. May be very skeptical to MT results. Resistant to change.<br>Wants to use SMT seamlessly integrated in daily routines. CAT tool integration is preferred. Very simple and quick on-line tool could be acceptable that supports file formats used in translation projects. |
| Translation Project Manager | Wants to reduce translation project cost by reducing time translators spend on translation. Would be ready to contribute high-quality corpora/translation memories in order to improve SMT system quality.<br>Needs to have good control and quality measurements of trained SMT systems.<br>Different clients have different domains, vocabularies, tools, translation styles. Thus a single SMT system does not provide the necessary quality of translation and post-editing of machine translated text takes the same or even longer time. Will need lots of specific SMT systems for each project type or customer.<br>Would use only a few SMT systems simultaneously, but needs lots of historical data to build SMT systems on-demand whenever specific project starts. |
| Localization Company Manager | Have high quality corpora which are most likely protected by IPR or confidentiality agreements.<br>Highly aware of IPR and the need to protect organization-specific knowledge (competitive advantage).<br>SMT usage in current translation workflow and tools must be cost and time efficient.<br>Currently maintenance and even evaluation of SMT feasibility is very expensive due to lack of knowledge of technologies involved and infrastructure requirements. |

| User type | Description |
|---|---|
| Web developers | |
| Web developer | Web developer who needs to create multi-lingual sites, but does not have necessary resources to provide high quality translation. Commonly will use LetsMT! for prototyping of web sites in different languages. If the quality of the translation is accepted by the client in general, the web developer could provide minor fixes and improvements of translation. |
| University education and research community | |
| Researcher | Member of educational and research organization or translation and localization industry organization investigating options available in SMT field.<br>Most interested in quality of translation and possibilities to control SMT system training. Has some parallel and monolingual corpora available and is interested in improvements in translation quality after corpora submission.<br>This user will want to tweak every possible SMT system option and compare results of these tweaks.<br>Could be a decision maker or contribute to decision making about use of LetsMT! services. |
| Research organization leader | Owns large amounts of mono and bi-lingual corpora. Is interested in theoretical aspects of results of SMT systems and how different input data influence quality of translations.<br>Research organizations most likely will have many different SMT systems and will re-train SMT systems frequently.<br>Would benefit from using LetsMT! platform as no investment in infrastructure is required. Funding possibilities are limited. |
| Translation automation solution developers and providers | |
| Translation Solution Product manager | As user community constantly requires MT solution integration solution developer would benefit from integration of LetsMT! platform into their products/solutions. Also many competitors have introduced MT modules in their solutions. Installation and maintenance of MT requires infrastructure and skilled specialists in MT are scarce.<br>Even access to public SMT systems through products could be seen beneficial to customers as trial of MT integration. More user-tailored solutions could be sold as a separate service. |

Major requirements and challenges encountered in currently available MT systems that are addressed by LetsMT! are listed in table below for each user group.

**Table 3. LetsMT! requirements summary by user types**

| User type | LetsMT! requirement | Challenges |
|---|---|---|
| Anonymous Internet user | • Free!<br>• Provide good enough translation to understand essence of foreign language texts | • Current SMT results are hard to comprehend due to poor quality and broad domain of translation text |
| Business news reader | • Good quality (better than Google) of multi-lingual financial news translations | • Machine translated financial information is hard to understand in under-resourced languages |
| MT enthusiast | • Can upload mono or bilingual corpora for improvement of SMT quality thus influencing quality of translation<br>• Uploaded corpora quality should be assessed | • Very limited influence on quality of publicly available SMT systems |
| Translator | • Easy to use<br>• Incorporated in workflow or CAT tool<br>• Fast to translate<br>• Quality of translation is critical<br>• Utilization of already available high quality translation memory | • Poor translation quality of SMT systems<br>• Hard to distinguish translations from TM (trusted) and MT (not trusted) in current translation workflows |

| User type | LetsMT! requirement | Challenges |
|---|---|---|
| Translation project manager | • Integration in current workflows/practices/tools<br>• Control over corpora included in SMT system training<br>• Single point of corpora/TM storage to simplify management<br>• On-demand SMT system training<br>• Frequent (incremental) re-training of SMT systems | • Resistance of Translators to use MT<br>• Complicated maintenance of different MT systems<br>• Need to establish new process to manage use of MT |
| Localization Company Manager | • Safeguard IPR and competitive advantage<br>• Integration in current localization and translation practices/tools<br>• Reduction of overall cost and time of project execution | • Maintenance overhead of storing various translation memory and corpora artifacts<br>• Project ramp-up time and context switching for translators too long<br>• Creation of MT infrastructure is expensive |
| Web developer | • Widget or web service for automatic web-page translation | |
| Researcher | • Support for many different SMT systems<br>• Possibility to re-train models often<br>• Possibility to tweak every possible training option<br>• Detailed reporting | |
| Research Organization Leader | • Possibility to upload large amounts of corpora<br>• Support for many SMT systems | • Hard to maintain infrastructure for corpora storage and many SMT system application in research<br>• Creation of MT infrastructure is expensive |

| User type | LetsMT! requirement | Challenges |
|---|---|---|
| Translation Solution Product Manager | • Open API<br>• Stability / availability<br>• Available trained SMT systems “out of the box” | • No need to develop and maintain SMT infrastructure for integration into tools/services provided<br>• Insufficient corpora available for SMT system training |

## 2.2 *Typical Usage Scenarios*

The main usage scenarios which are relevant to define the functionality of the LetsMT! system are described in this section.

### 2.2.1 General Use Scenario

This scenario describes general use of the LetsMT! system by various user types to acquire translation from user tailored SMT system.

There are a number of freely available translation systems on the Internet; however none of those allow users the options to choose between domain–specific systems. Also users cannot directly influence SMT systems by uploading corpora in user’s possession for achieving better quality of translations for their needs. LetsMT! addresses these shortcomings.

Users are able to upload corpora into the system, using a widely accessible client application, for example, by using a web page interface. Users are provided with an effective way of searching, navigating and selecting a trained SMT system to use. And, of course, translate texts using one of available LetsMT! trained SMT systems.

### 2.2.2 LetsMT! and Common Localization Process

MT in localization can only be successful if highly specific engines are developed. The MT training capability should be exposed to a technically skilled end user through a simple GUI or API. Advanced meta-data management, including legal is essential in order to select appropriate training data for models.

Localization data typically contain significant amounts of mark-up. The mark-up is of two varieties - meta-segment and in-line. Whereas meta-segment mark-up is easily filtered out the in-line elements constitute a cleaning challenge – placeholders can’t be simply filtered out. In case no working substitution algorithm is found these segments must be thrown away not to pollute the training corpus.

General localization process is summarized in the following diagram:

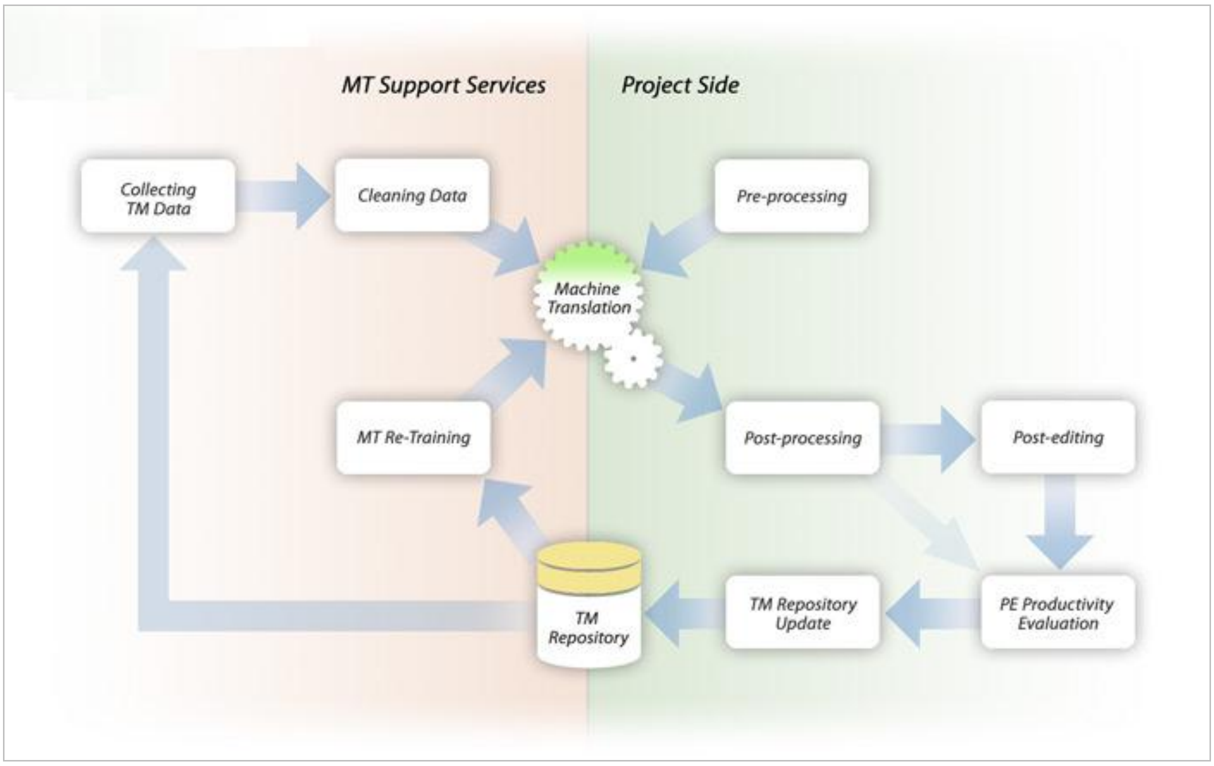


**Figure 1. Localization process diagram.**

The LetsMT! system provides the management of trained engines and training data and the organization of  these by means of various meta-data.

Production pre-processing is able to recognize factoids of source formats and processes them according to strict rules. It strips or replaces all mark-up but stores it to attempt reapplication during post-processing.

Major automated evaluation metrics – BLEU, TER (including language specific where available), METEOR (including language specific where available) are built in. The end-user MT trainer is able to rapidly assess the efficiency of added training data with respect to the test set.

It is possible to achieve good enough quality in raw-output-publishing scenarios through human feedback, i.e. post-editing and incremental retraining. In human quality publishing scenarios integration with CAT tools is a must, because only segments that do not have good TM matches (based on a configurable threshold) are typically sent to MT. Translator/Post-editor typically edits the TM and MT suggestions at the same time. Therefore it is critical for post-editing scenarios to integrate MT suggestions in major CAT tools, such as SDL Trados and Kilgray MemoQ. Post-edited translations should be automatically stored in a TM repository (ideally as increment to the original training data) to be used from time to time for MT retraining. Ideally, the end-quality post-edited strings should feedback directly into the MT's retraining capability.

LetsMT! integrates in translation process through an API or files fed directly into the system. LetsMT! currently doesn't support incremental training but research goes on and it is planned to be implemented in LetsMT! platform as soon as possible.

### 2.2.3 Financial News Use Scenario

The LetsMT! consortium has identified that press releases are an ideal source for LetsMT! data, specifically for the online MT service of business and financial news. International press releases are almost invariably distributed in English to the international business community. However, as many companies registered in non-native English countries are required to release important business news in their native language, a large percentage of business press releases are available in two languages.

For the majority of cases, the translation was done by the company themselves, before submitting the press release to the news provider. In this case either the company dissemination (their website) or the local national press release agency provides the original language press release. Listed companies normally provide an overview of their press releases from their corporate websites. National press agencies also support access to the releases made through their service. Full texts are available only for a fee to registered users. In addition also archived data is commercially available. As timing is critical for press release data, the time stamp of the message is a reliable indicator by which the original language document can be matched to the English translation.

### 2.2.4 Academic use scenario

Researches, students and other academic users can have great benefit of the LetsMT! platform. Especially researchers teaching in Natural Language processing, translation and corpus linguistics can use the LetsMT! platform. Easy access to try out SMT systems covering different domains and based on different training data will be very usable in teaching in a number of courses at universities in Europe.

For researchers it is important that large corpora are available for training without further conversion, filtering and data storage needs. Another important facility is uploading of user-defined corpora and availability of those for training new systems.

The researchers have more special needs for exact documenting of the training parameters and the performance of the trained systems, than other users. Therefore upload of specific tuning and evaluation corpora are demands for the academic user. Also display of used training parameters are very important, and should be available in the user interface. Concerning the data upload and availability of corpora, it is very important to be able to specify metadata for each corpus, not only covering subject domain, but also text type and other information, also including a small description of the corpora. This make the corpora widely documented in a formal manner, and hopefully can contribute to build up a best practice for documenting stored corpora. The researchers will also in the LetsMT! platform have a common reference platform for SMT systems, having a public and open platform for students to access, and for technical discussions.

Some quota for training research systems should be available also in the future if EU-funding runs out, as it is much more feasible and cost-effective to have a common available SMT-platform for researchers, than letting each researcher install SMT software and corpora by themselves.

## 2.3 *User Stories*

The next chapters cover most common LetsMT! usage scenarios. They are grouped by three main LetsMT! features – (i) **translate**, (ii) **build SMT system** and (iii) **store and share training data**.

### 2.3.1 Translating with LetsMT!

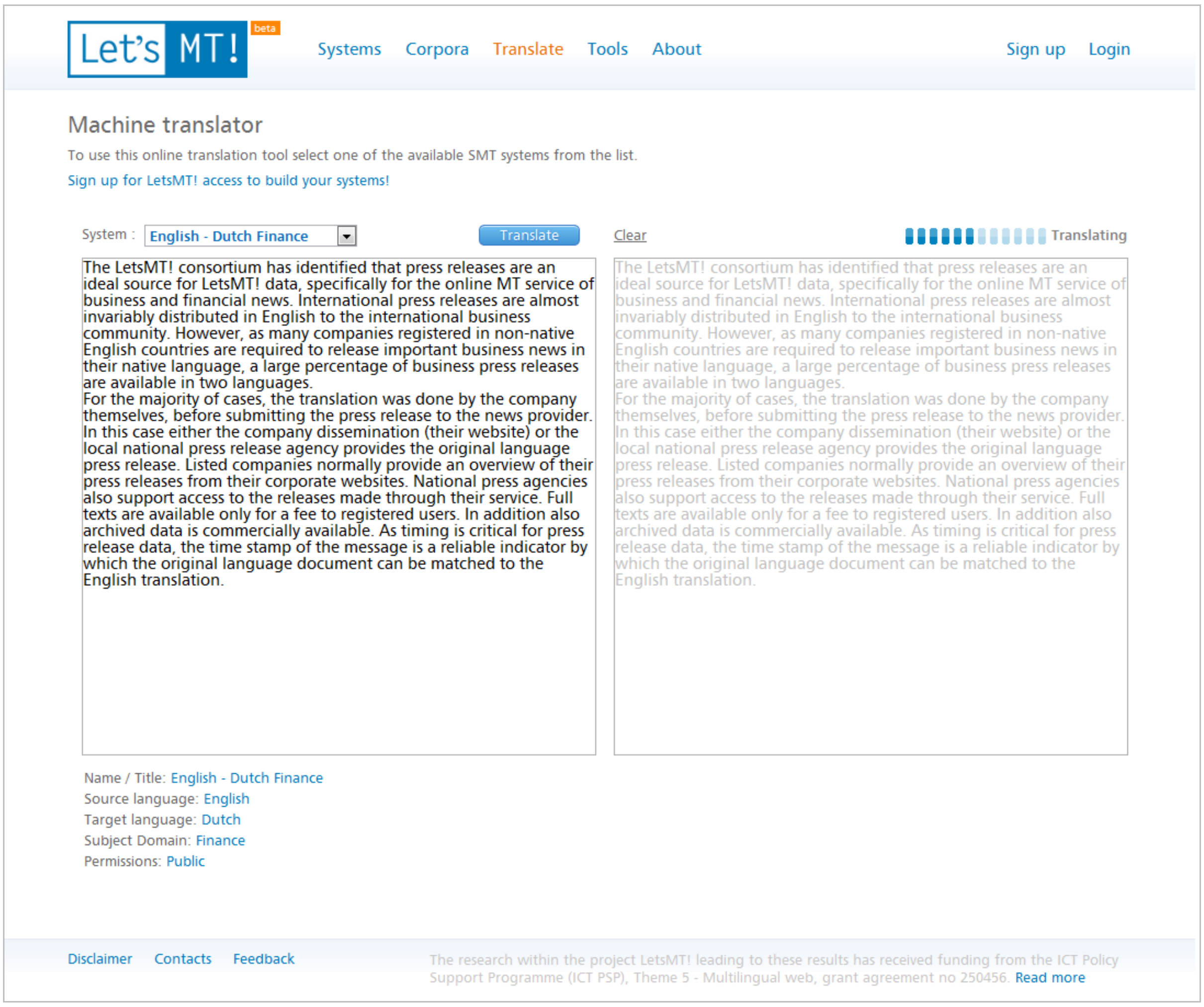
This section describes different LetsMT! options to perform its main task – translate.

### 2.3.1.1 Translating in LetsMT! website

The LetsMT! platform provides a possibility to try different SMT systems in order to evaluate them for further possible usage through API or CAT tools. An anonymous user can try only public MT systems; authenticated user can try also private systems which are allowed to him.

User navigates to "Translate" section (see Figure 2), selects one of the available LetsMT! systems and enters the text for translation in the left text area. Upon paste, enter, pause in typing (few seconds) or after pressing the *Translate* button, the text is translated and the translation shown in the text area on the right. The LetsMT! supports progressive translation – the text is translated and updated paragraph by paragraph as soon as translation arrives from the SMT system. Progress bar shows translation progress.

To help focus on matching source and translation fragments (paragraphs) Synchronous scrolling is performed and matching original and translated text sentences are highlighted.



Machine translator

To use this online translation tool select one of the available SMT systems from the list.

Sign up for LetsMT! access to build your systems!

System : English - Dutch Finance

The LetsMT! consortium has identified that press releases are an ideal source for LetsMT! data, specifically for the online MT service of business and financial news. International press releases are almost invariably distributed in English to the international business community. However, as many companies registered in non-native English countries are required to release important business news in their native language, a large percentage of business press releases are available in two languages.
For the majority of cases, the translation was done by the company themselves, before submitting the press release to the news provider. In this case either the company dissemination (their website) or the local national press release agency provides the original language press release. Listed companies normally provide an overview of their press releases from their corporate websites. National press agencies also support access to the releases made through their service. Full texts are available only for a fee to registered users. In addition also archived data is commercially available. As timing is critical for press release data, the time stamp of the message is a reliable indicator by which the original language document can be matched to the English translation.

The LetsMT! consortium has identified that press releases are an ideal source for LetsMT! data, specifically for the online MT service of business and financial news. International press releases are almost invariably distributed in English to the international business community. However, as many companies registered in non-native English countries are required to release important business news in their native language, a large percentage of business press releases are available in two languages.
For the majority of cases, the translation was done by the company themselves, before submitting the press release to the news provider. In this case either the company dissemination (their website) or the local national press release agency provides the original language press release. Listed companies normally provide an overview of their press releases from their corporate websites. National press agencies also support access to the releases made through their service. Full texts are available only for a fee to registered users. In addition also archived data is commercially available. As timing is critical for press release data, the time stamp of the message is a reliable indicator by which the original language document can be matched to the English translation.

Name / Title: English - Dutch Finance
Source language: English
Target language: Dutch
Subject Domain: Finance
Permissions: Public

The research within the project LetsMT! leading to these results has received funding from the ICT Policy Support Programme (ICT PSP), Theme 5 - Multilingual web, grant agreement no 250456. Read more

**Figure 2. Public translation web page**

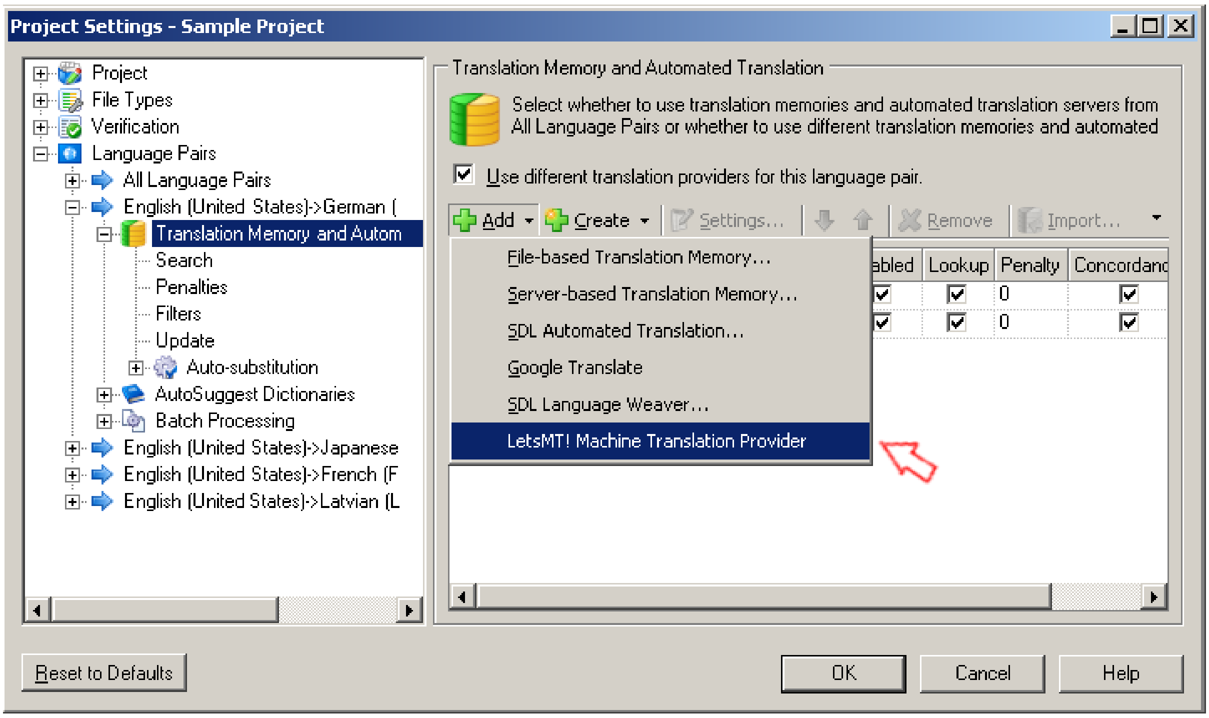
### 2.3.1.2 Translating with LetsMT! SMT systems in SDL Trados Studio 2009

One of the LetsMT! features is to provide integration with CAT tools to support “on the fly” translation requests over the network. This scenario describes usage of LetsMT! within SDL Trados Studio 2009.

To use LetsMT! in SDL Trados “LetsMT! Machine Translation Provider” must be downloaded from the LetsMT! site, installed and associated with a project (see Figure 3). User can choose to use LetsMT! as an authenticated user or as an anonymous user (see Figure 4). Anonymous users are able to use only public SMT systems. Authenticated users are able to use public systems and SMT systems running for the respective LetsMT! service subscriber. The configuration dialog appears where the user is asked to choose the SMT system if several systems are available (see Figure 5). “LetsMT! Machine Translation Provider” appears next to translation memories (see Figure 6 on page 19).

When user initiates translating a segment of a document (user clicks on particular segment), the translation result window is populated with suggested translations. These result are suggestions from translation memories (if any) that are supplemented with results from the LetsMT! Machine Translation Provider (see Figure 7).

One of important LetsMT! features is **correct handling of mark-up** in translation results. For example, if source text contains some formatting markup like bolded text, it is preserved in translated text.



Project Settings - Sample Project

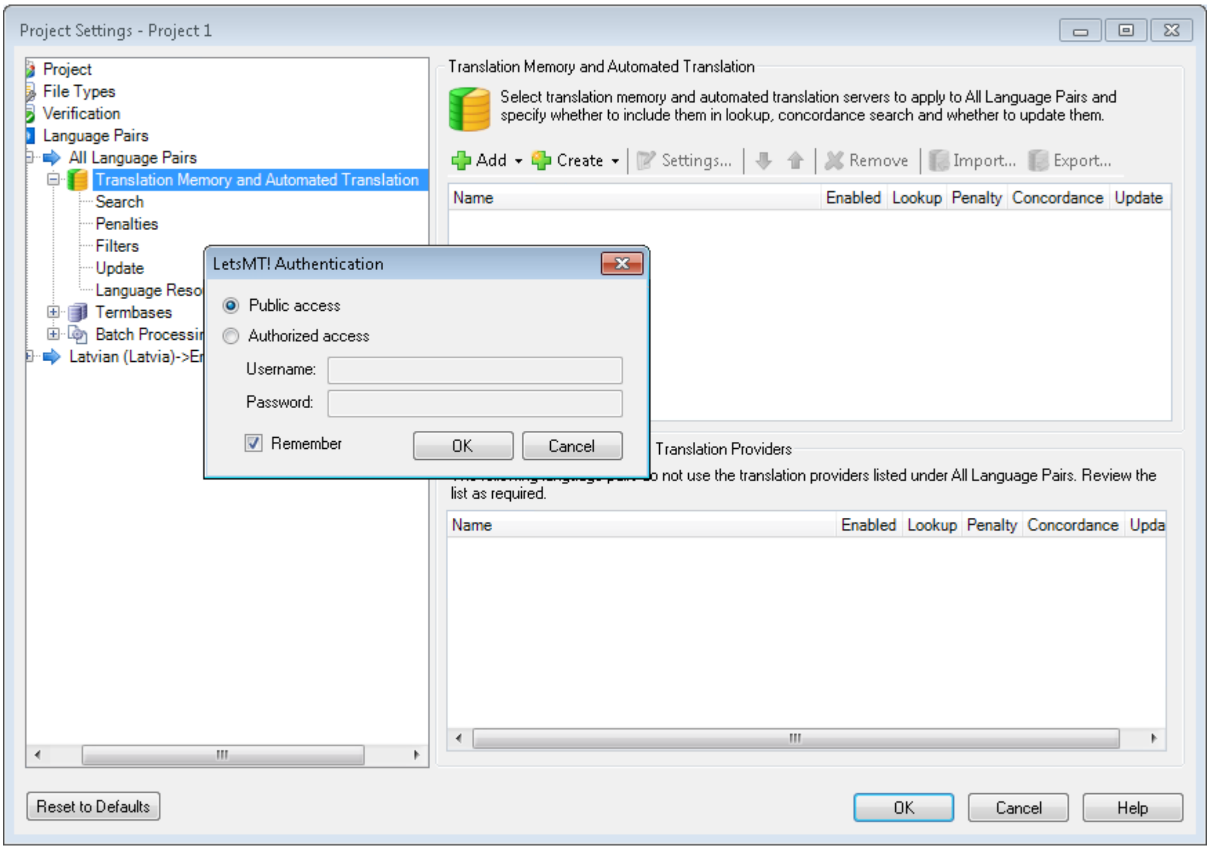
Translation Memory and Automated Translation

Select whether to use translation memories and automated translation servers from All Language Pairs or whether to use different translation memories and automated

☑ Use different translation providers for this language pair.

Project
File Types
Verification
Language Pairs
All Language Pairs
English (United States)->German (
Translation Memory and Autom
Search
Penalties
Filters
Update
Auto-substitution
AutoSuggest Dictionaries
Batch Processing
English (United States)->Japanese
English (United States)->French (F
English (United States)->Latvian (L
Add
Create
Settings...
Remove
Import...
File-based Translation Memory...
Server-based Translation Memory...
SDL Automated Translation...
Google Translate
SDL Language Weaver...
LetsMT! Machine Translation Provider
abled
Lookup
Penalty
Concordanc
0
0
Reset to Defaults
OK
Cancel
Help

**Figure 3. Adding LetsMT! translation provider to project.**



Project Settings - Project 1

Translation Memory and Automated Translation

Select translation memory and automated translation servers to apply to All Language Pairs and specify whether to include them in lookup, concordance search and whether to update them.

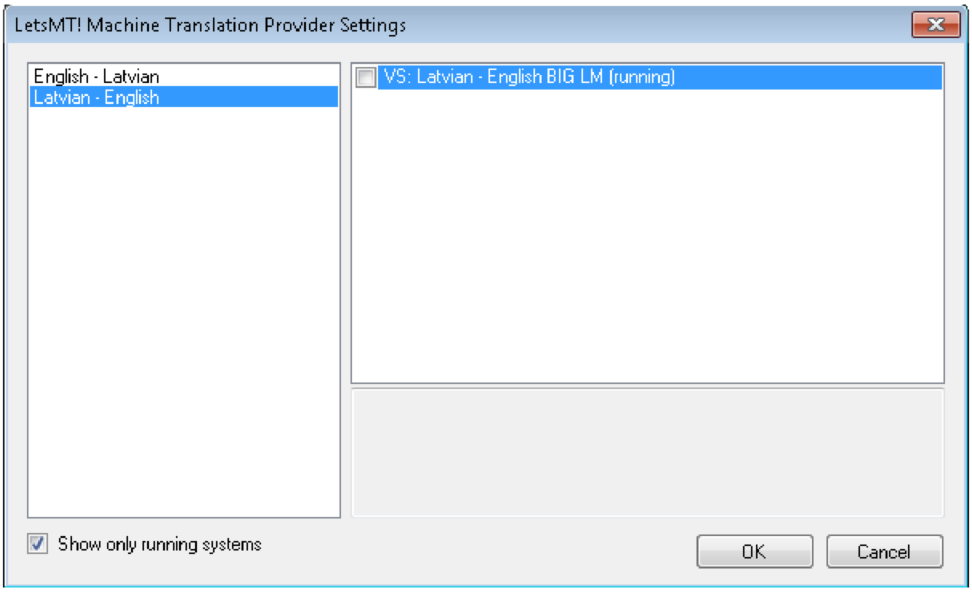
LetsMT! Authentication

● Public access
○ Authorized access
Username:
Password:
☑ Remember

Translation Providers

…o not use the translation providers listed under All Language Pairs. Review the list as required.

**Figure 4. User authentication in LetsMT! Machine Translation Provider**



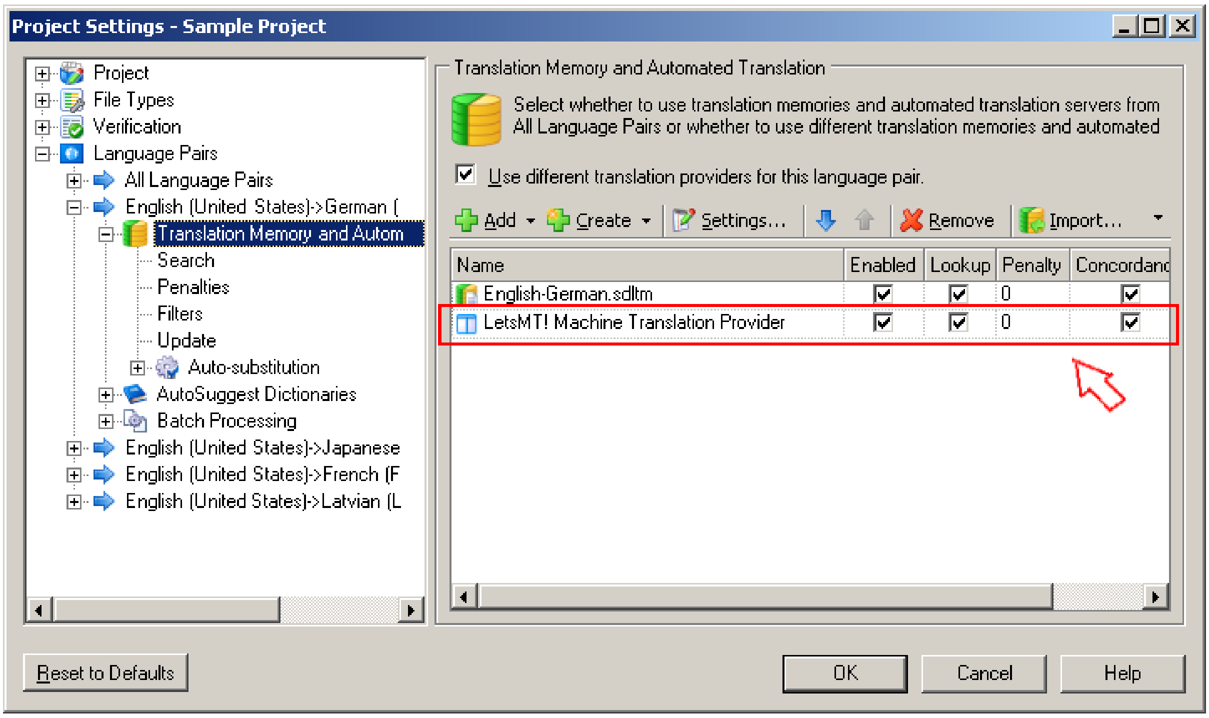
LetsMT! Machine Translation Provider Settings

- English - Latvian
- Latvian - English

□ VS: Latvian - English BIG LM (running)

☑ Show only running systems

**Figure 5. MT system selection**



Project Settings - Sample Project

Project
File Types
Verification
Language Pairs
All Language Pairs
English (United States)->German (
Translation Memory and Autom
Search
Penalties
Filters
Update
Auto-substitution
AutoSuggest Dictionaries
Batch Processing
English (United States)->Japanese
English (United States)->French (F
English (United States)->Latvian (L

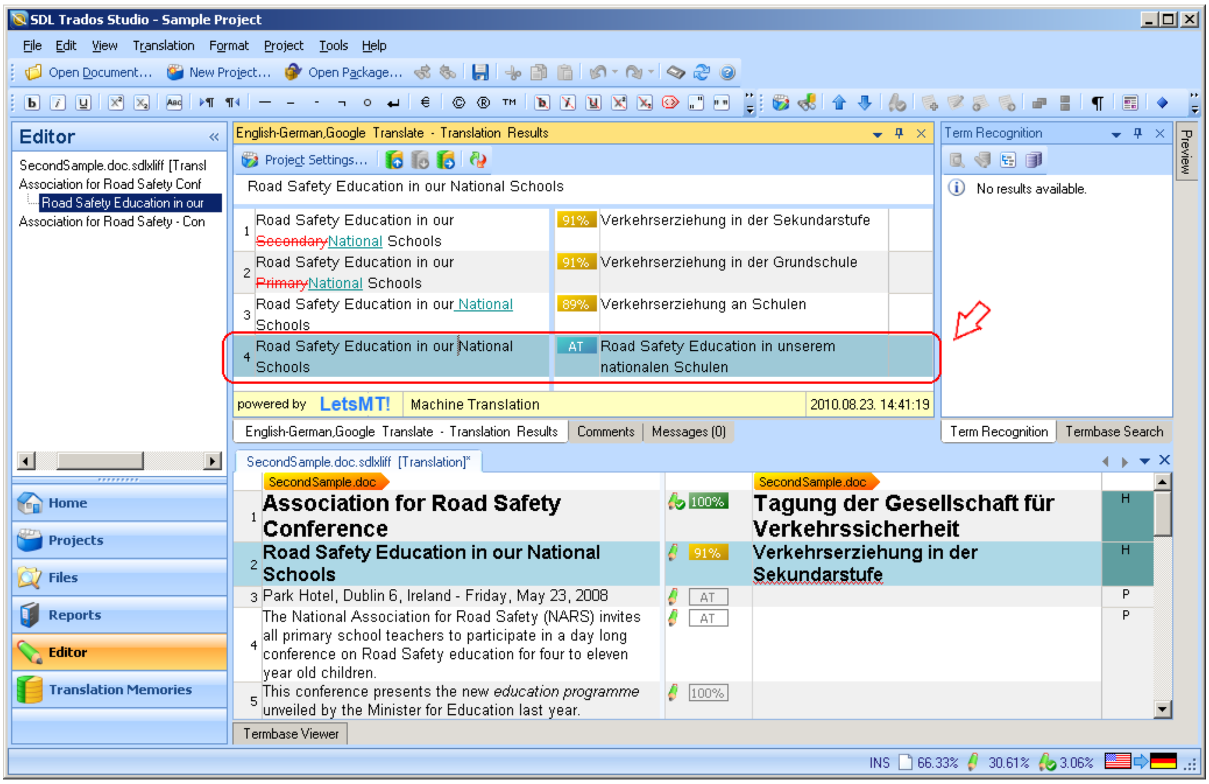
Translation Memory and Automated Translation

Select whether to use translation memories and automated translation servers from All Language Pairs or whether to use different translation memories and automated

☑ Use different translation providers for this language pair.

| Name | Enabled | Lookup | Penalty | Concordanc |
|---|---|---|---|---|
| English-German.sdltm | ☑ | ☑ | 0 | ☑ |
| LetsMT! Machine Translation Provider | ☑ | ☑ | 0 | ☑ |

**Figure 6. Successful LetsMT! translation provider configuration.**


SDL Trados Studio - Sample Project
File Edit View Translation Format Project Tools Help
Open Document... New Project... Open Package...
Editor
SecondSample.doc.sdlxliff [Transl
Association for Road Safety Conf
Road Safety Education in our
Association for Road Safety - Con
Home
Projects
Files
Reports
Editor
Translation Memories
English-German,Google Translate - Translation Results
Project Settings...
Term Recognition
No results available.
Preview
powered by LetsMT! Machine Translation 2010.08.23. 14:41:19
English-German,Google Translate - Translation Results
Comments
Messages (0)
Term Recognition
Termbase Search
SecondSample.doc.sdlxliff [Translation]*
SecondSample.doc
Termbase Viewer
INS 66.33% 30.61% 3.06%

Road Safety Education in our National Schools

| | | | |
|---|---|---|---|
| 1 | Road Safety Education in our ~~Secondary~~National Schools | 91% | Verkehrserziehung in der Sekundarstufe |
| 2 | Road Safety Education in our ~~Primary~~National Schools | 91% | Verkehrserziehung in der Grundschule |
| 3 | Road Safety Education in our National Schools | 89% | Verkehrserziehung an Schulen |
| 4 | Road Safety Education in our National Schools | AT | Road Safety Education in unserem nationalen Schulen |

| | | | | |
|---|---|---|---|---|
| 1 | Association for Road Safety Conference | 100% | Tagung der Gesellschaft für Verkehrssicherheit | H |
| 2 | Road Safety Education in our National Schools | 91% | Verkehrserziehung in der Sekundarstufe | H |
| 3 | Park Hotel, Dublin 6, Ireland - Friday, May 23, 2008 | AT | | P |
| 4 | The National Association for Road Safety (NARS) invites all primary school teachers to participate in a day long conference on Road Safety education for four to eleven year old children. | AT | | P |
| 5 | This conference presents the new *education programme* unveiled by the Minister for Education last year. | 100% | | |

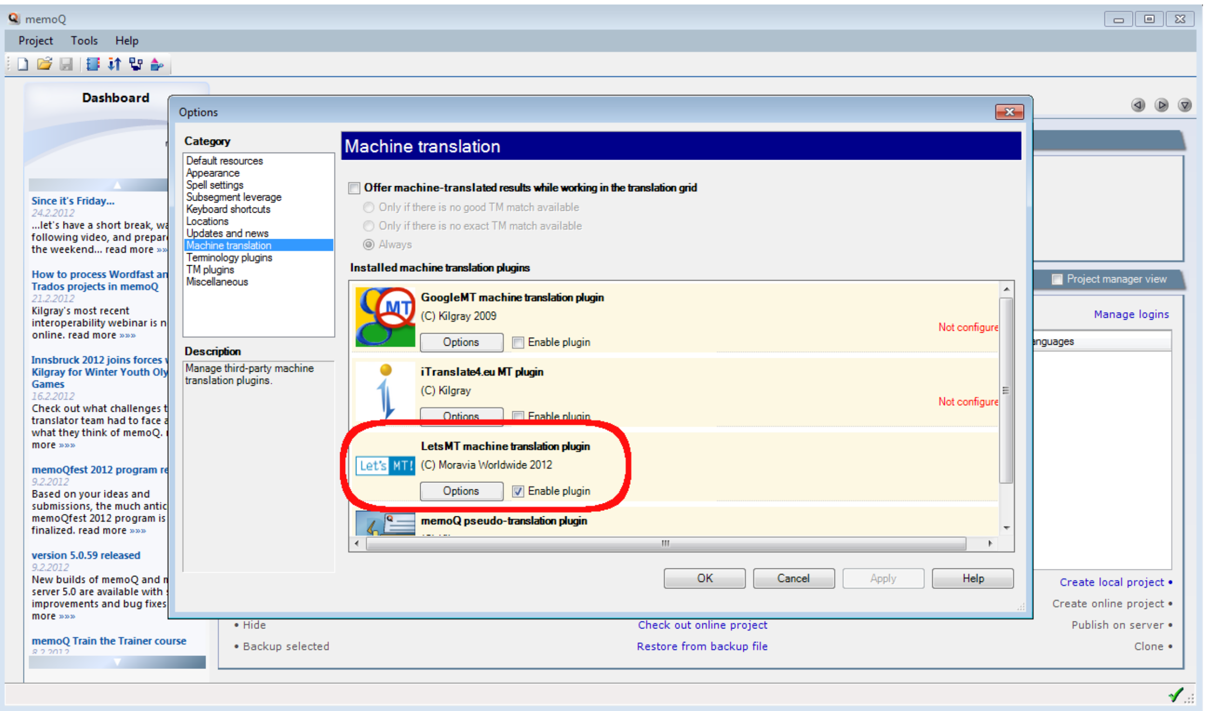
**Figure 7. Translation suggestion from LetsMT! system.**

### 2.3.1.3 Translating with LetsMT! SMT systems in memoQ

Second LetsMT! plug-in will utilize Kilgray's memoQ as another industry CAT tool. The development work is done by Moravia.

When using memoQ, using the menu Tools->Options, pick up Machine translation label, list of MT possibilities is offered (Figure 8. List of MT services). LetsMT! plug-in, as well as any other MT provider, can be enabled or disabled. Settings can be edited by clicking on Options (Figure 9. LetsMT! options). Plug-in's options are similar to the Trados plug-in. There is a list of language pairs where each of them can contain multiple engines. Show only running systems switches on/off the view of systems (engines) that are not currently running. Engine's confidence level threshold can be set up in order to make easier to distinguish between different memoQ's sources (e.g. MT, TM, etc.).

After clicking on User label, a login dialog will appear (Figure 10. Login dialog). Without successful logging procedure, only public engines will be possible to use.



Options

**Category**

- Default resources
- Appearance
- Spell settings
- Subsegment leverage
- Keyboard shortcuts
- Locations
- Updates and news
- Machine translation
- Terminology plugins
- TM plugins
- Miscellaneous

**Description**

Manage third-party machine translation plugins.

**Machine translation**

□ Offer machine-translated results while working in the translation grid

○ Only if there is no good TM match available
○ Only if there is no exact TM match available
● Always

**Installed machine translation plugins**

GoogleMT machine translation plugin
(C) Kilgray 2009
Options □ Enable plugin
Not configure

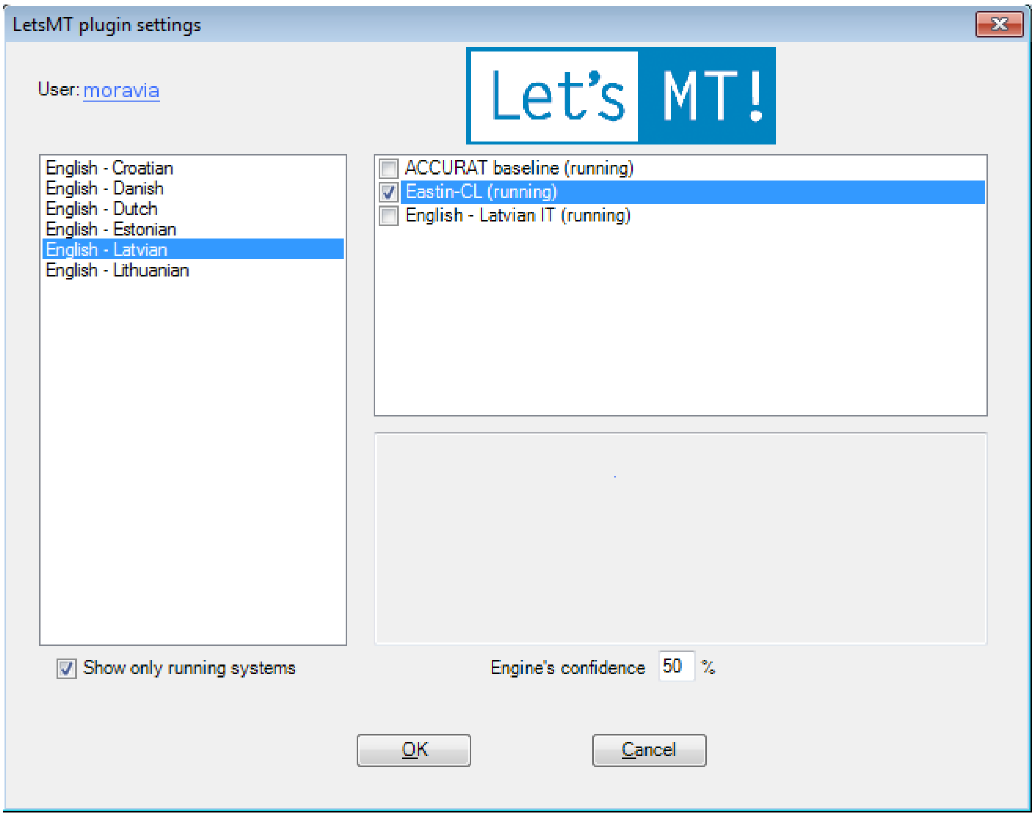
iTranslate4.eu MT plugin
(C) Kilgray
Options □ Enable plugin
Not configure

LetsMT machine translation plugin
(C) Moravia Worldwide 2012
Options ☑ Enable plugin

memoQ pseudo-translation plugin

OK Cancel Apply Help

**Figure 8. List of MT services**


LetsMT plugin settings

User: moravia

- English - Croatian
- English - Danish
- English - Dutch
- English - Estonian
- English - Latvian
- English - Lithuanian

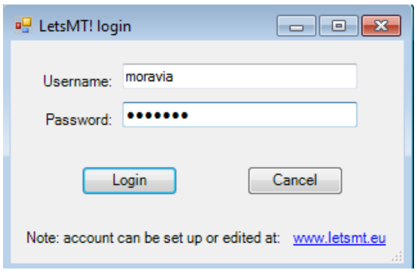
□ ACCURAT baseline (running)
☑ Eastin-CL (running)
□ English - Latvian IT (running)

☑ Show only running systems

Engine's confidence 50 %

OK Cancel

**Figure 9. LetsMT! options**



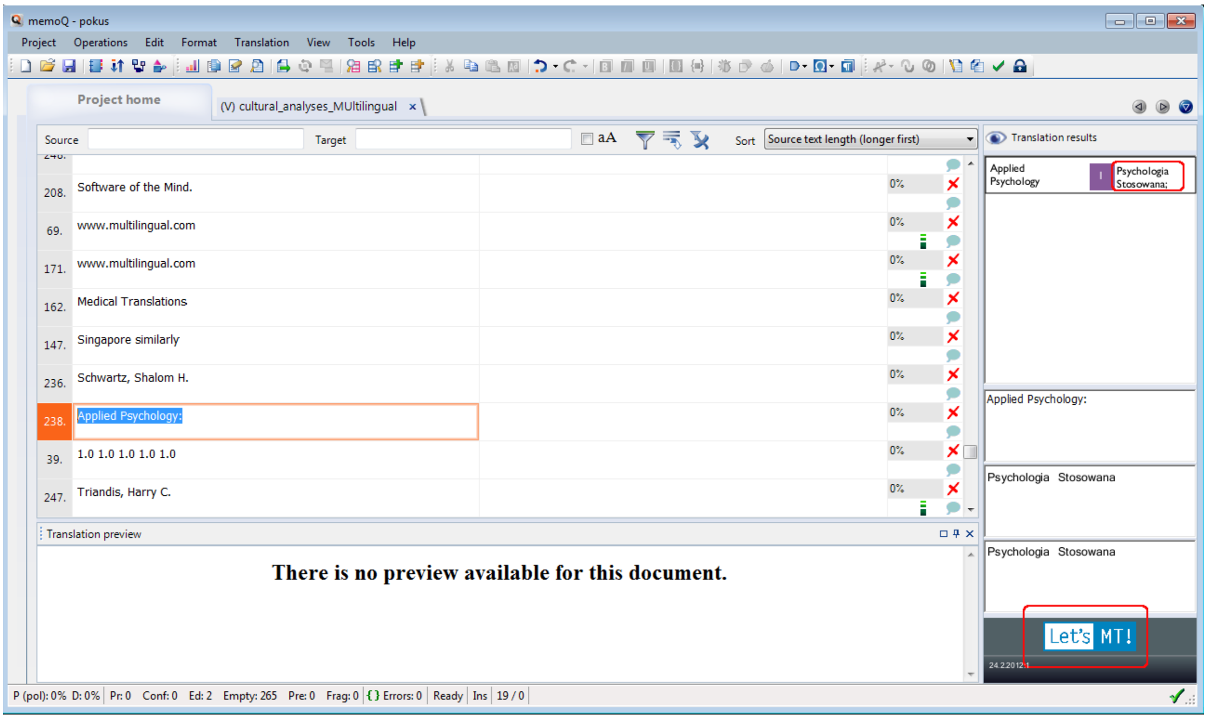
Username: moravia

Password: •••••••

Note: account can be set up or edited at: www.letsmt.eu

**Figure 10. Login dialog**

After LetsMT plug-in is configured, it can be used for MT suggestion (Figure 11. memoQ translation editor with MT suggestion from LetsMT!) or memoQ's Pre-Translation.



memoQ - pokus
Project Operations Edit Format Translation View Tools Help
Project home
(V) cultural_analyses_MUltilingual
Source
Target
aA
Sort
Translation results
Translation preview
Let's MT!

Sort: Source text length (longer first)

| | Source | | |
|---|---|---|---|
| 208. | Software of the Mind. | | 0% |
| 69. | www.multilingual.com | | 0% |
| 171. | www.multilingual.com | | 0% |
| 162. | Medical Translations | | 0% |
| 147. | Singapore similarly | | 0% |
| 236. | Schwartz, Shalom H. | | 0% |
| 238. | Applied Psychology: | | 0% |
| 39. | 1.0 1.0 1.0 1.0 1.0 | | 0% |
| 247. | Triandis, Harry C. | | 0% |

Translation results: Applied Psychology – Psychologia Stosowana;

Applied Psychology:

Psychologia Stosowana

Psychologia Stosowana

There is no preview available for this document.

P (pol): 0% D: 0% | Pr: 0 Conf: 0 Ed: 2 Empty: 265 Pre: 0 Frag: 0 | { } Errors: 0 | Ready | Ins | 19 / 0

**Figure 11. memoQ translation editor with MT suggestion from LetsMT!**

### 2.3.1.4 Translating with LetsMT! Widget

One of the features for the LetsMT! platform is the translation widget – a module that can be integrated into client websites in order to provide multilingual web features. The translation widget is a LetsMT! front-end which implements a client-side functionality powered by LetsMT! web-service (API) in a back-end. See the SemLab Sentiment analysis service (2.3.1.5 on Page 23) for live example of LetsMT! widget implementation.

The translation widget is a Javascript module. The module itself accesses the data on the website, feeds it to the LetsMT! translation web-service and updates web page content with the translated text. There are two basic user profiles or user classes for the translation widget: the web developer and the end user. Following are the usage scenarios for these user profiles.

Web developer:

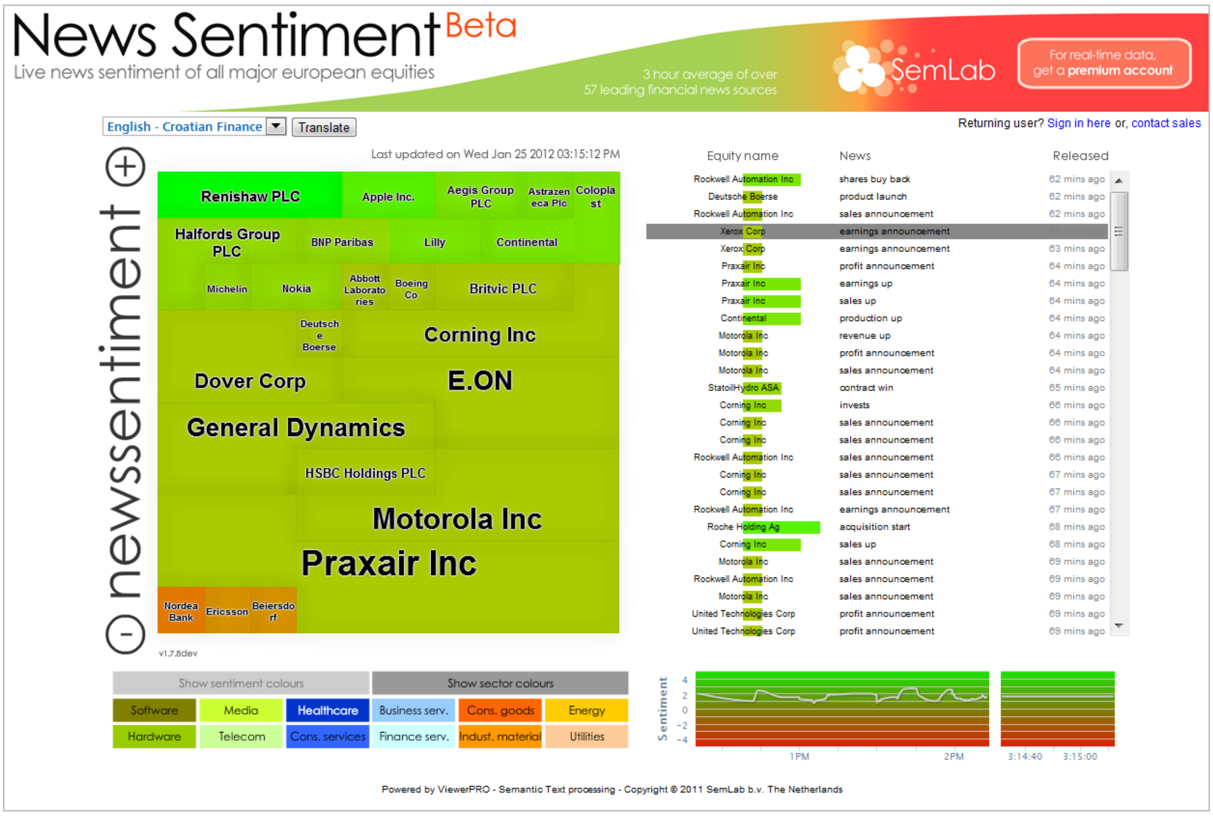
- Web developer goes to the LetsMT! public website and downloads the translation widget module. Downloading the translation widget is coupled by accepting the terms and conditions for its integration and usage within client websites.
- Web developer integrates the module within his/her website in order to provide the translation service to its end users. The module is integrated within the webpage in a visible location, preferably the header menu or similar element of the user interface.

End user:

- The user accesses the client website in which the translation widget was previously integrated.
- Using the simple menu within the widget, the user selects the target language for the translation from the *Target Language* drop-down menu and presses the *Translate* button. A translation is provided to the user inline, within the client website elements previously containing source language text.
- The web page translation is done progressively element by element.

### 2.3.1.5 Translating Financial News

SemLab has built a web-service available at http://letsmt.newssentiment.eu/ which aggregates financial news from different trusted sources and analyzes each message for its "sentiment" or important keywords thus identifying trends in company or brand ratings. This is further visualized in a comprehensive way for easy trend monitoring. The LetsMT! website translation widget is integrated in this News Sentiment webpage (see Figure 12).



| Equity name | News | Released |
|---|---|---|
| Rockwell Automation Inc | shares buy back | 62 mins ago |
| Deutsche Boerse | product launch | 62 mins ago |
| Rockwell Automation Inc | sales announcement | 62 mins ago |
| Xerox Corp | earnings announcement | 63 mins ago |
| Xerox Corp | earnings announcement | 63 mins ago |
| Praxair Inc | profit announcement | 64 mins ago |
| Praxair Inc | earnings up | 64 mins ago |
| Praxair Inc | sales up | 64 mins ago |
| Continental | production up | 64 mins ago |
| Motorola Inc | revenue up | 64 mins ago |
| Motorola Inc | profit announcement | 64 mins ago |
| Motorola Inc | sales announcement | 64 mins ago |
| StatoilHydro ASA | contract win | 65 mins ago |
| Corning Inc | invests | 66 mins ago |
| Corning Inc | sales announcement | 66 mins ago |
| Corning Inc | sales announcement | 66 mins ago |
| Rockwell Automation Inc | sales announcement | 66 mins ago |
| Corning Inc | sales announcement | 67 mins ago |
| Corning Inc | sales announcement | 67 mins ago |
| Rockwell Automation Inc | earnings announcement | 67 mins ago |
| Roche Holding Ag | acquisition start | 68 mins ago |
| Corning Inc | sales up | 68 mins ago |
| Motorola Inc | sales announcement | 69 mins ago |
| Rockwell Automation Inc | sales announcement | 69 mins ago |
| Motorola Inc | sales announcement | 69 mins ago |
| United Technologies Corp | profit announcement | 69 mins ago |
| United Technologies Corp | profit announcement | 69 mins ago |

Powered by ViewerPRO - Semantic Text processing - Copyright © 2011 SemLab b.v. The Netherlands

**Figure 12. SemLab financial news translation service.**

General idea behind the website is illustrated below in Figure 13.

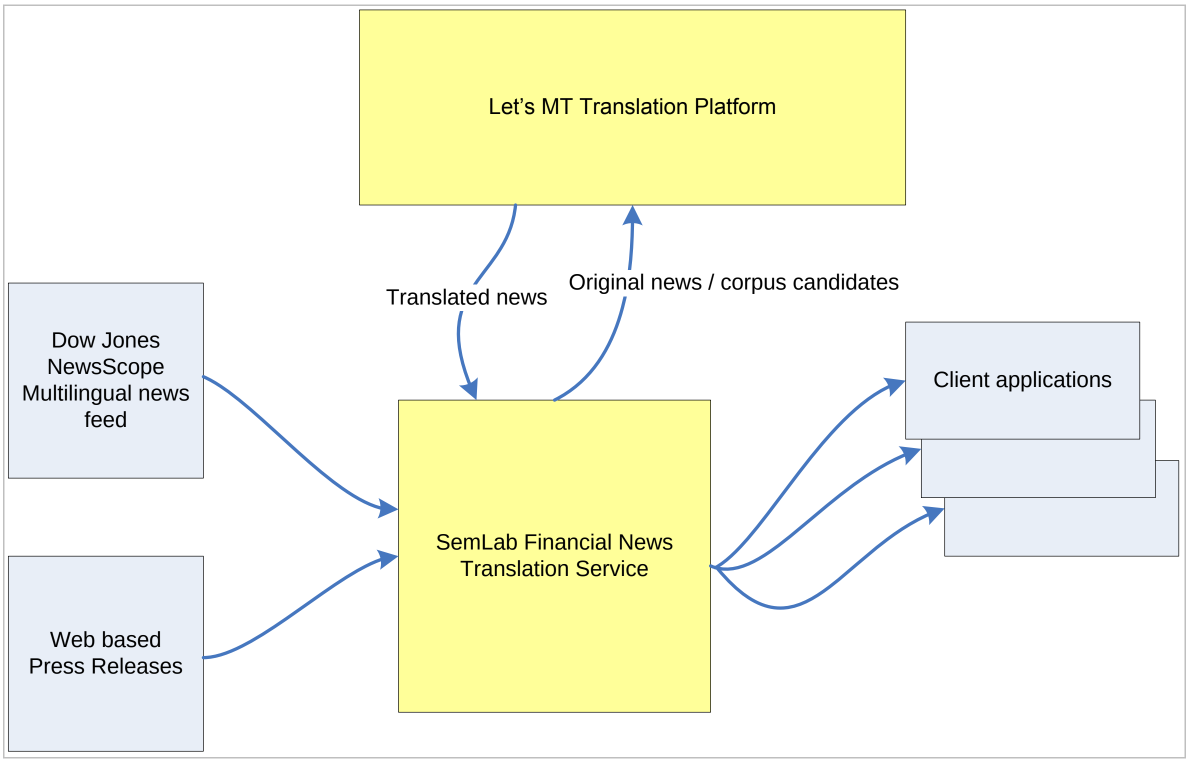

Let's MT Translation Platform
Translated news
Original news / corpus candidates
Dow Jones NewsScope Multilingual news feed
Web based Press Releases
SemLab Financial News Translation Service
Client applications

**Figure 13. SemLab financial news translation service integration logical diagram.**

LetsMT! service is integrated in two ways (i) as a translation widget (see Figure 14) and (ii) using LetsMT! API (Figure 15). User can translate the whole web page or separate news item. LetsMT! webpage translation widget comes in action when whole website needs to be translated.

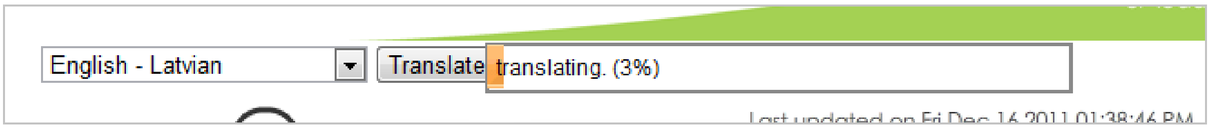
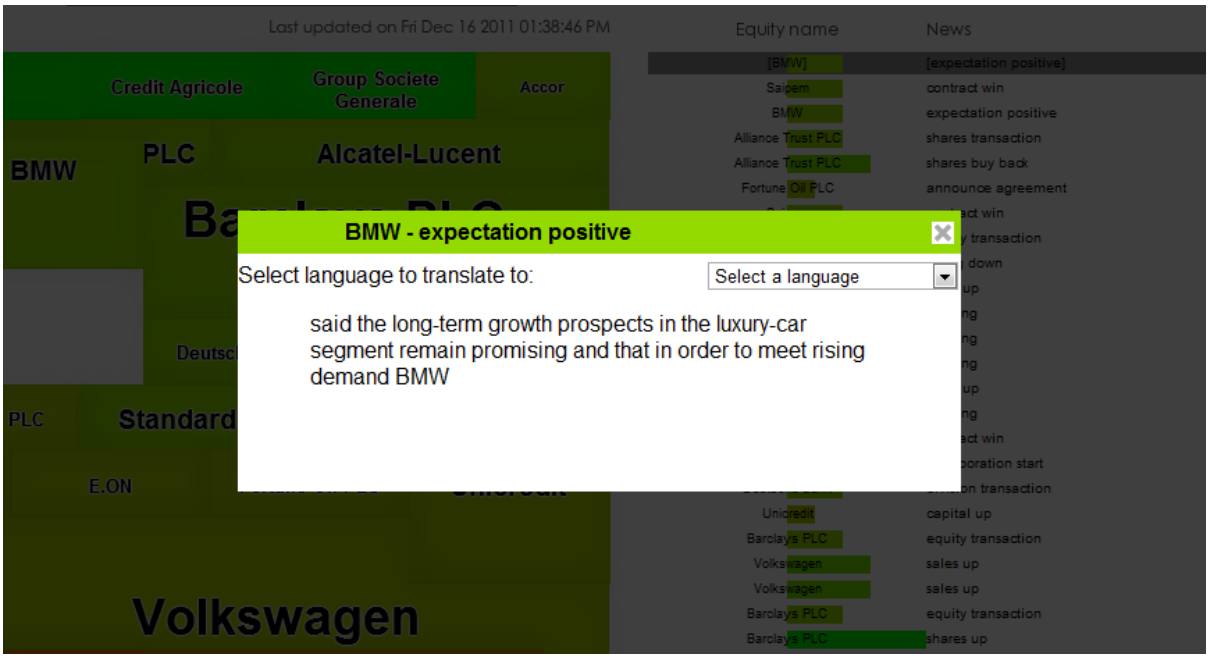


**Figure 14. Web page widget implementation in http://letsmt.newssentiment.eu/.**



BMW - expectation positive

Select language to translate to: Select a language

said the long-term growth prospects in the luxury-car segment remain promising and that in order to meet rising demand BMW

| Equity name | News |
|---|---|
| [BMW] | [expectation positive] |
| Saipem | contract win |
| BMW | expectation positive |
| Alliance Trust PLC | shares transaction |
| Alliance Trust PLC | shares buy back |
| Fortune Oil PLC | announce agreement |
| | act win |
| | y transaction |
| | down |
| | up |
| | ng |
| | ng |
| | ng |
| | up |
| | ng |
| | act win |
| | poration start |
| | on transaction |
| Unicredit | capital up |
| Barclays PLC | equity transaction |
| Volkswagen | sales up |
| Volkswagen | sales up |
| Barclays PLC | equity transaction |
| Barclays PLC | shares up |

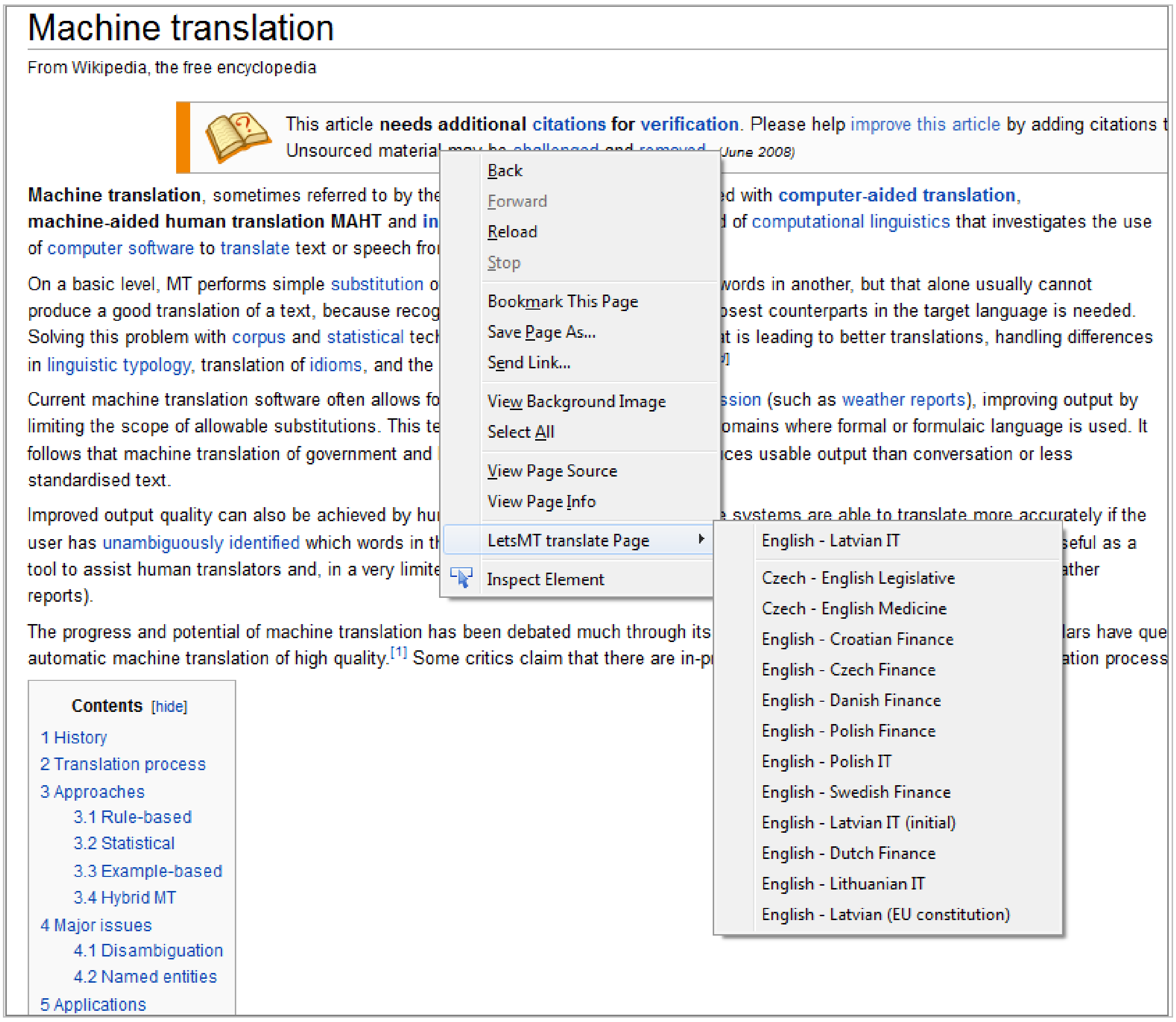
**Figure 15. LetsMT! API implementation in http://letsmt.newssentiment.eu/.**

### 2.3.1.6 Translate Using Web Browser Plug-in

One of the features for the LetsMT! platform is the browser plug-in – an extension for a web browser that enables real-time machine translation of websites by accessing the LetsMT! web-service. Plugin currently supports Mozilla Firefox and Microsoft Internet Explorer. The browser plug-in functionality is similar to the LetsMT! translation widget, as it provides real-time translation on the web. However, the translation widget must be integrated in the website, while the browser plug-in can be installed in the client web browser to provide translations for any website.

The usage scenario is as follows:

- The user goes to the LetsMT! public website and downloads the browser plug-in for his browser.
- The user installs the plug-in. A restart of the browser may be required.
- After loading some website, the user right-clicks in a web page and selects the *LetsMT! translate Page* menu option provided by the LetsMT! plug-in. The context-menu dropdown lists all available SMT systems (see Figure 16).



Machine translation

From Wikipedia, the free encyclopedia

This article **needs additional citations for verification**. Please help improve this article by adding citations t

Unsourced material may be challenged and removed. (June 2008)

**Machine translation**, sometimes referred to by the … d with **computer-aided translation**,

**machine-aided human translation MAHT** and **in** … l of computational linguistics that investigates the use

of computer software to translate text or speech fro

On a basic level, MT performs simple substitution o … words in another, but that alone usually cannot produce a good translation of a text, because recog … osest counterparts in the target language is needed. Solving this problem with corpus and statistical tech … t is leading to better translations, handling differences in linguistic typology, translation of idioms, and the

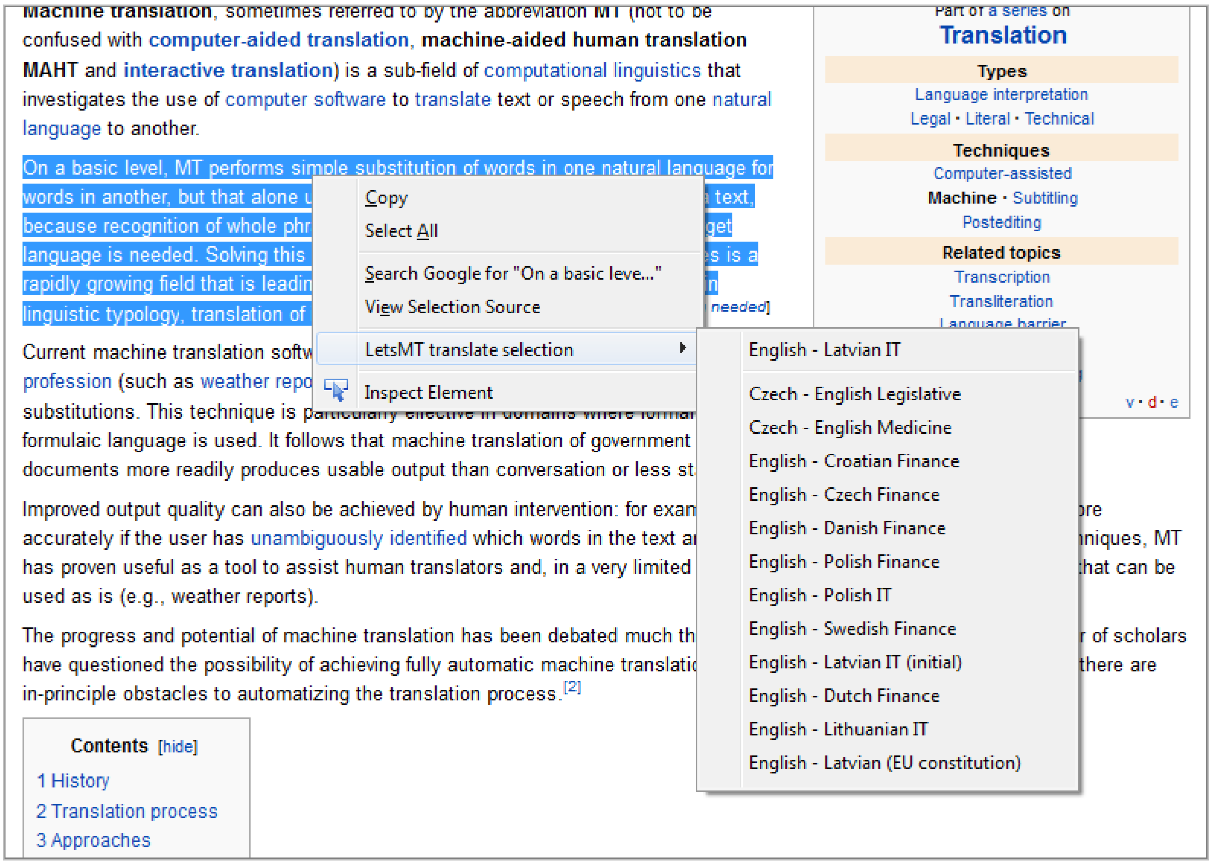
Current machine translation software often allows fo … sion (such as weather reports), improving output by limiting the scope of allowable substitutions. This te … omains where formal or formulaic language is used. It follows that machine translation of government and … ces usable output than conversation or less standardised text.

Improved output quality can also be achieved by hu … e systems are able to translate more accurately if the user has unambiguously identified which words in th … seful as a tool to assist human translators and, in a very limite … ather reports).

The progress and potential of machine translation has been debated much through its … lars have que automatic machine translation of high quality.[1] Some critics claim that there are in-p … ation process

**Figure 16. Translation of the whole web page using LetsMT! web browser plug-in.**

Rather than translating the entire web page, the user can select a certain portion of text from the web page and use the plug-in as described above (see Figure 17).



Machine translation, sometimes referred to by the abbreviation MT (not to be confused with computer-aided translation, machine-aided human translation MAHT and interactive translation) is a sub-field of computational linguistics that investigates the use of computer software to translate text or speech from one natural language to another.

On a basic level, MT performs simple substitution of words in one natural language for words in another, but that alone u … text, because recognition of whole phr … get language is needed. Solving this … s is a rapidly growing field that is leadin … n linguistic typology, translation of … needed]

Current machine translation softw … profession (such as weather repo … substitutions. This technique is particularly effective in domains where formal … formulaic language is used. It follows that machine translation of government documents more readily produces usable output than conversation or less st … ore

Improved output quality can also be achieved by human intervention: for exam … accurately if the user has unambiguously identified which words in the text ar … niques, MT has proven useful as a tool to assist human translators and, in a very limited … hat can be used as is (e.g., weather reports).

The progress and potential of machine translation has been debated much th … r of scholars have questioned the possibility of achieving fully automatic machine translatio … there are in-principle obstacles to automatizing the translation process.[2]

**Figure 17. Translation of selected text using LetsMT! web browser plug-in.**

The text is sent to the LetsMT! web-service by the browser plug-in and the translation is displayed inline within the website.

### 2.3.1.7 Translation web service API

LetsMT! provides an API for integration into other systems. LetsMT! API is used by other LetsMT! components – online public translation service, SDL Trados plug-in, translation widget and browser plug-ins. Developers can build customized solutions using any software development technology that can access web services.

Read more about *Public API* on page 50.

## 2.3.2 Building and Training the SMT system

SMT systems accessible to the user are show in one list in “Systems” page. Due to the high granularity of SMT systems (e.g. for every customer/domain) the number of such systems may be very high, thus LetsMT! provides a simple navigation and flexible filtering mechanism (see Figure 18).

System may have one of the following statuses:

- Green – system is trained and running (system is available for translation)
- Black– system successfully trained, but not running (it must be launched to use it for translation)
- Grey – system training has not been started yet

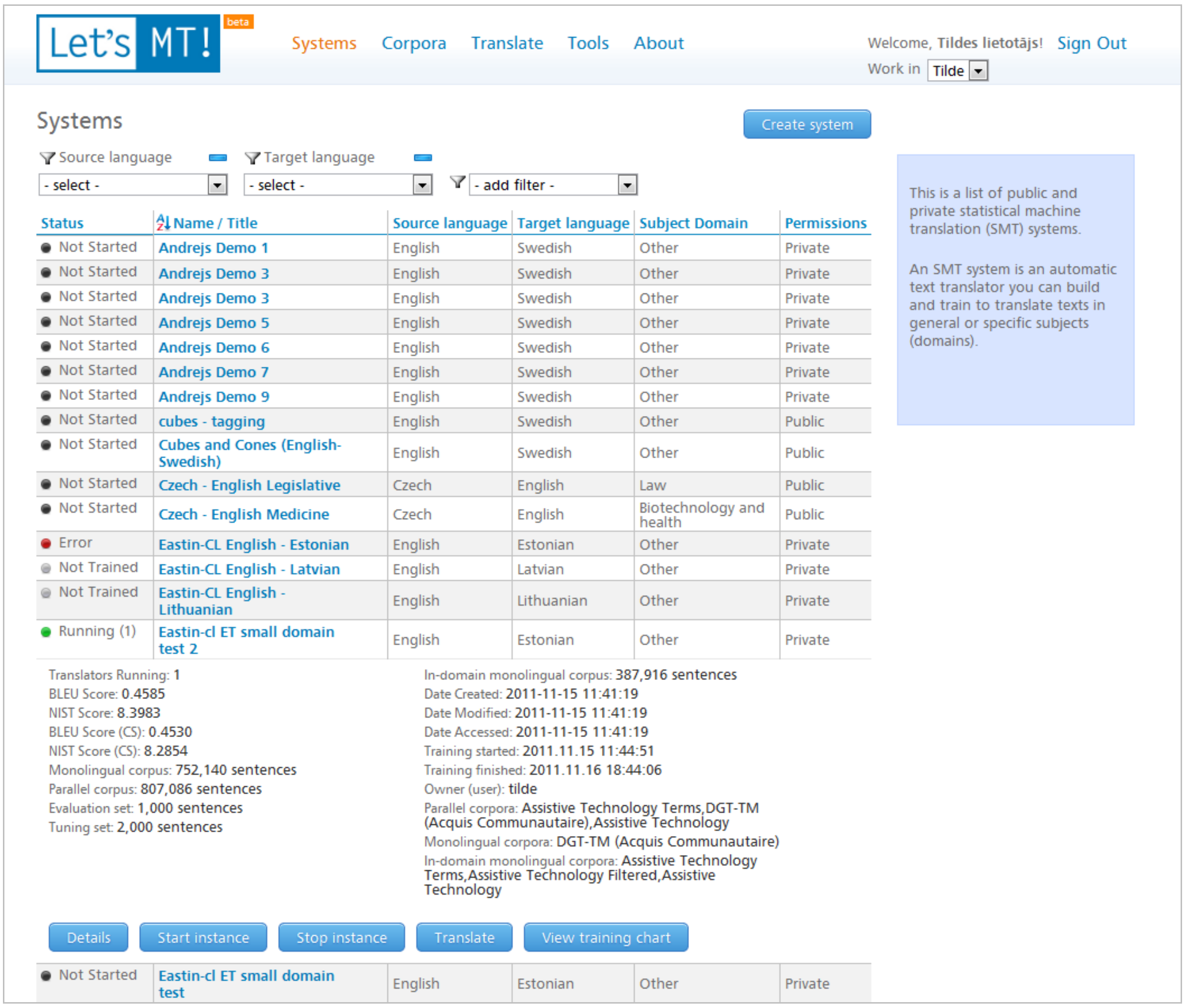
- Yellow – the system training is in progress
- Red – an error has occurred in training process

Clicking on a system expands its details and shows the following information:

- Number of running instances
- Provider and owner
- Corpora used for system training
- Creation and training dates
- Main quality metrics (BLEU, NIST, TER, METEOR)

Under the expanded system's info block depending on user access permissions the following functions may appear:

- Details
- Start/stop instance
- Translate
- View training chart



Systems

| Status | Name / Title | Source language | Target language | Subject Domain | Permissions |
|---|---|---|---|---|---|
| ● Not Started | Andrejs Demo 1 | English | Swedish | Other | Private |
| ● Not Started | Andrejs Demo 3 | English | Swedish | Other | Private |
| ● Not Started | Andrejs Demo 3 | English | Swedish | Other | Private |
| ● Not Started | Andrejs Demo 5 | English | Swedish | Other | Private |
| ● Not Started | Andrejs Demo 6 | English | Swedish | Other | Private |
| ● Not Started | Andrejs Demo 7 | English | Swedish | Other | Private |
| ● Not Started | Andrejs Demo 9 | English | Swedish | Other | Private |
| ● Not Started | cubes - tagging | English | Swedish | Other | Public |
| ● Not Started | Cubes and Cones (English-Swedish) | English | Swedish | Other | Public |
| ● Not Started | Czech - English Legislative | Czech | English | Law | Public |
| ● Not Started | Czech - English Medicine | Czech | English | Biotechnology and health | Public |
| ● Error | Eastin-CL English - Estonian | English | Estonian | Other | Private |
| ● Not Trained | Eastin-CL English - Latvian | English | Latvian | Other | Private |
| ● Not Trained | Eastin-CL English - Lithuanian | English | Lithuanian | Other | Private |
| ● Running (1) | Eastin-cl ET small domain test 2 | English | Estonian | Other | Private |

Translators Running: 1
BLEU Score: 0.4585
NIST Score: 8.3983
BLEU Score (CS): 0.4530
NIST Score (CS): 8.2854
Monolingual corpus: 752,140 sentences
Parallel corpus: 807,086 sentences
Evaluation set: 1,000 sentences
Tuning set: 2,000 sentences

In-domain monolingual corpus: 387,916 sentences
Date Created: 2011-11-15 11:41:19
Date Modified: 2011-11-15 11:41:19
Date Accessed: 2011-11-15 11:41:19
Training started: 2011.11.15 11:44:51
Training finished: 2011.11.16 18:44:06
Owner (user): tilde
Parallel corpora: Assistive Technology Terms,DGT-TM (Acquis Communautaire),Assistive Technology
Monolingual corpora: DGT-TM (Acquis Communautaire)
In-domain monolingual corpora: Assistive Technology Terms,Assistive Technology Filtered,Assistive Technology

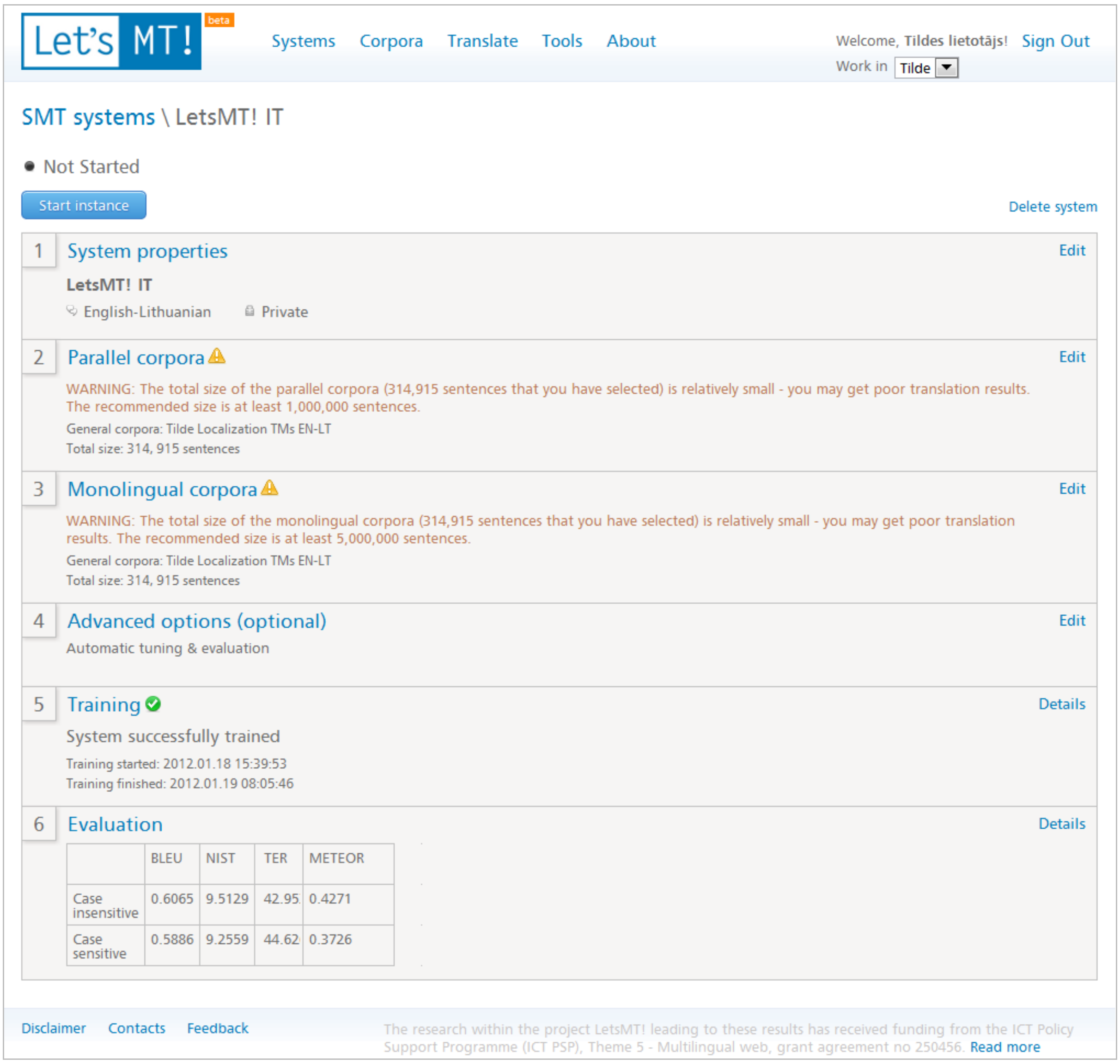
| Status | Name / Title | Source language | Target language | Subject Domain | Permissions |
|---|---|---|---|---|---|
| ● Not Started | Eastin-cl ET small domain test | English | Estonian | Other | Private |

This is a list of public and private statistical machine translation (SMT) systems.

An SMT system is an automatic text translator you can build and train to translate texts in general or specific subjects (domains).

**Figure 18. SMT System List with details of one system expanded.**

SMT system details view is built as an overview of main SMT system parts (see Figure 19). Any warning and error messages are shown here in brief and detailed descriptions are available by expanding a particular section.



SMT systems \ LetsMT! IT

● Not Started

1 System properties

LetsMT! IT

English-Lithuanian    Private

2 Parallel corpora ⚠

WARNING: The total size of the parallel corpora (314,915 sentences that you have selected) is relatively small - you may get poor translation results. The recommended size is at least 1,000,000 sentences.

General corpora: Tilde Localization TMs EN-LT
Total size: 314, 915 sentences

3 Monolingual corpora ⚠

WARNING: The total size of the monolingual corpora (314,915 sentences that you have selected) is relatively small - you may get poor translation results. The recommended size is at least 5,000,000 sentences.

General corpora: Tilde Localization TMs EN-LT
Total size: 314, 915 sentences

4 Advanced options (optional)

Automatic tuning & evaluation

5 Training ✓

System successfully trained

Training started: 2012.01.18 15:39:53
Training finished: 2012.01.19 08:05:46

6 Evaluation

| | BLEU | NIST | TER | METEOR |
|---|---|---|---|---|
| Case insensitive | 0.6065 | 9.5129 | 42.95 | 0.4271 |
| Case sensitive | 0.5886 | 9.2559 | 44.62 | 0.3726 |

The research within the project LetsMT! leading to these results has received funding from the ICT Policy Support Programme (ICT PSP), Theme 5 - Multilingual web, grant agreement no 250456. Read more

**Figure 19. SMT system Details (overview) screen.**

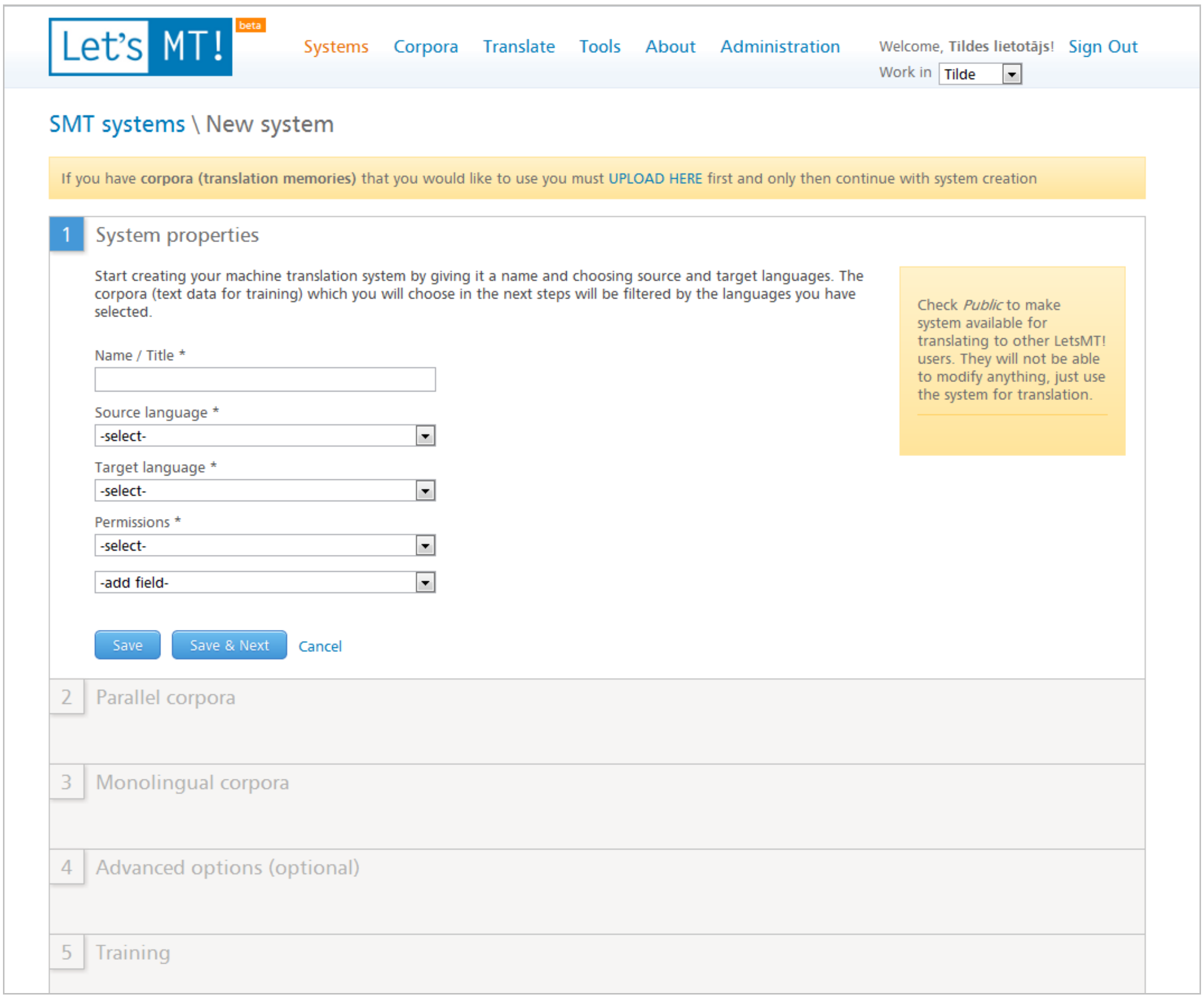
## 2.3.2.1 SMT System Building

The LetsMT! SMT system building and management has been re-thought and simplified to few easy steps. The following steps are a minimum for sufficient SMT system creation:

1. User selects source and target languages and specifies the name of the SMT system. User sets access level of the SMT system – public or private. Optionally the user may select a domain (see Figure 20).
2. User selects parallel corpora to use for system training. A list of corpora matching selected source and target languages is shown and user selects corpora to include in training (see Figure 21).

3. User selects monolingual corpora to use for system training. A list of corpora matching the target languages is shown and user selects corpora to include in training.
4. User saves the changes.
5. SMT system definition is created. The next step is to train it (see chapter 2.3.2.2 *SMT System*).

Any warning and error messages as well as statistics (e.g. total size of selected corpora) are shown after saving a particular step thus informing user to correct issues if any.



SMT systems \ New system

If you have **corpora (translation memories)** that you would like to use you must UPLOAD HERE first and only then continue with system creation

1 System properties

Start creating your machine translation system by giving it a name and choosing source and target languages. The corpora (text data for training) which you will choose in the next steps will be filtered by the languages you have selected.

Check *Public* to make system available for translating to other LetsMT! users. They will not be able to modify anything, just use the system for translation.

Name / Title *

Source language * -select-

Target language * -select-

Permissions * -select-

-add field-

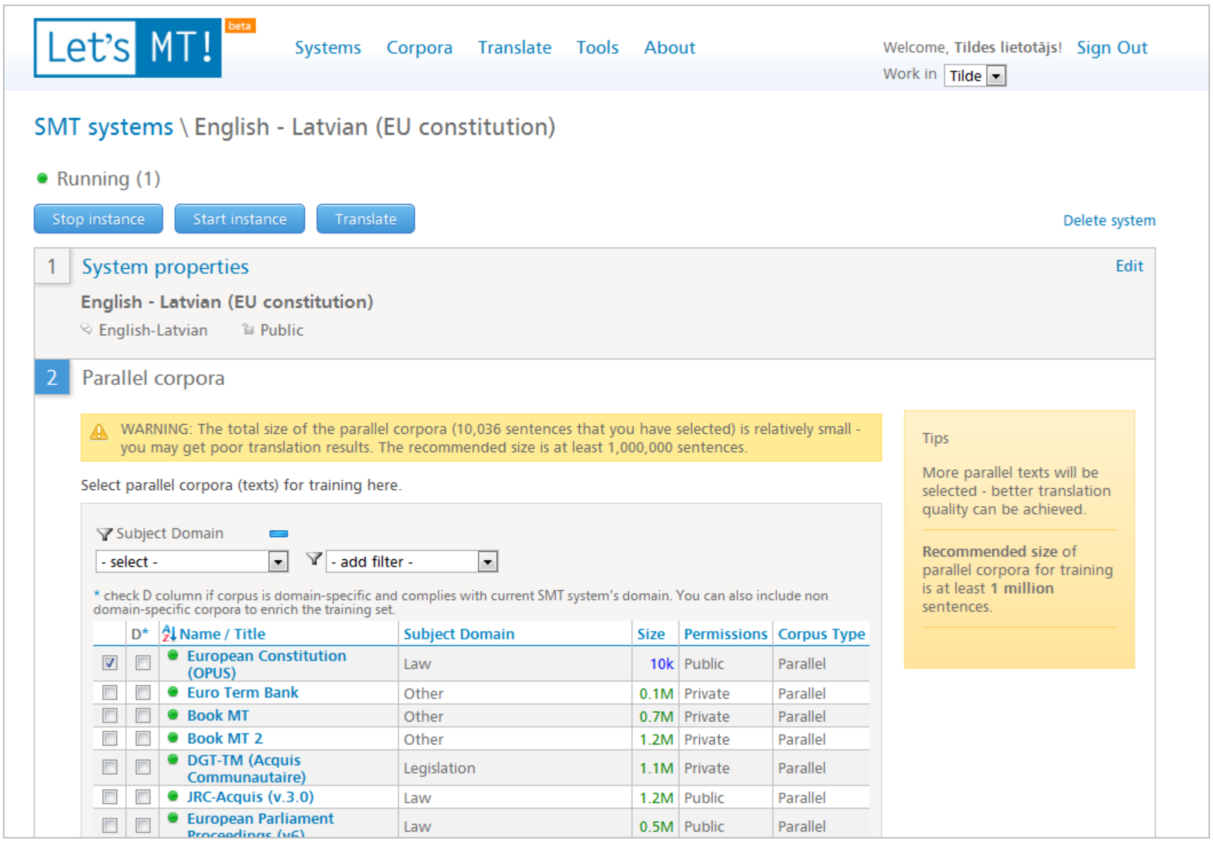
2 Parallel corpora

3 Monolingual corpora

4 Advanced options (optional)

5 Training

**Figure 20. New SMT system definition screen, Step 1.**



SMT systems \ English - Latvian (EU constitution)

● Running (1)

1 System properties

English - Latvian (EU constitution)

English-Latvian Public

2 Parallel corpora

WARNING: The total size of the parallel corpora (10,036 sentences that you have selected) is relatively small - you may get poor translation results. The recommended size is at least 1,000,000 sentences.

Tips

More parallel texts will be selected - better translation quality can be achieved.

Recommended size of parallel corpora for training is at least 1 million sentences.

Select parallel corpora (texts) for training here.

Subject Domain: - select -

- add filter -

* check D column if corpus is domain-specific and complies with current SMT system's domain. You can also include non domain-specific corpora to enrich the training set.

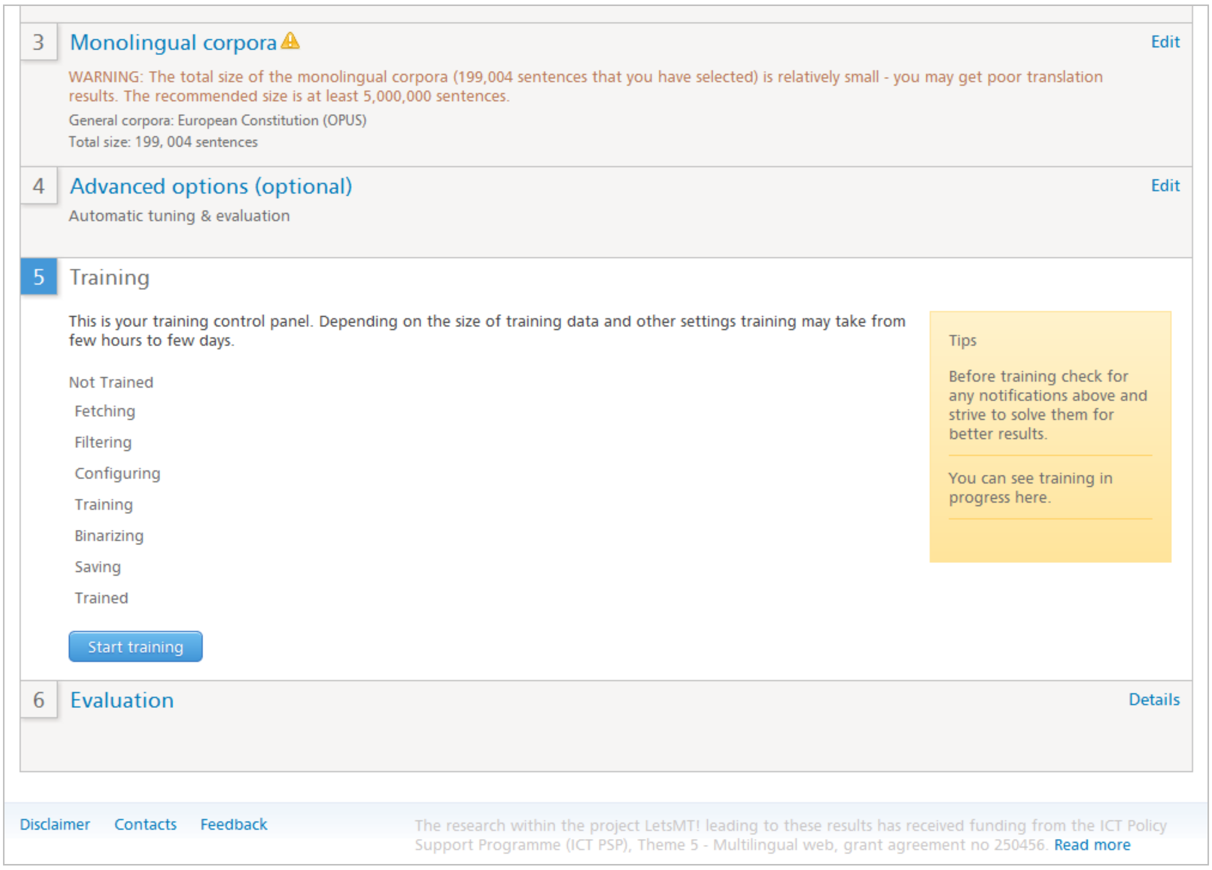
|  | D* | Name / Title | Subject Domain | Size | Permissions | Corpus Type |
|---|---|---|---|---|---|---|
| ☑ | □ | ● European Constitution (OPUS) | Law | 10k | Public | Parallel |
| □ | □ | ● Euro Term Bank | Other | 0.1M | Private | Parallel |
| □ | □ | ● Book MT | Other | 0.7M | Private | Parallel |
| □ | □ | ● Book MT 2 | Other | 1.2M | Private | Parallel |
| □ | □ | ● DGT-TM (Acquis Communautaire) | Legislation | 1.1M | Private | Parallel |
| □ | □ | ● JRC-Acquis (v.3.0) | Law | 1.2M | Public | Parallel |
| □ | □ | ● European Parliament Proceedings (v6) | Law | 0.5M | Public | Parallel |

**Figure 21. SMT system Details screen, Step 2 – parallel corpora selection.**

### 2.3.2.2 SMT System Training

Once the user has specified the SMT system, it goes to *Training* section (Step 5) in SMT system's detail view where he/she hits "Start training" button (see Figure 22) and sees an updating list of training steps reflecting current state training progress (see Figure 23).

The system can be trained only if there is at least one parallel corpus and at least one monolingual corpus selected.



3 Monolingual corpora ⚠ Edit

WARNING: The total size of the monolingual corpora (199,004 sentences that you have selected) is relatively small - you may get poor translation results. The recommended size is at least 5,000,000 sentences.

General corpora: European Constitution (OPUS)

Total size: 199, 004 sentences

4 Advanced options (optional) Edit

Automatic tuning & evaluation

5 Training

This is your training control panel. Depending on the size of training data and other settings training may take from few hours to few days.

Not Trained

Fetching

Filtering

Configuring

Training

Binarizing

Saving

Trained

Start training

Tips

Before training check for any notifications above and strive to solve them for better results.

You can see training in progress here.

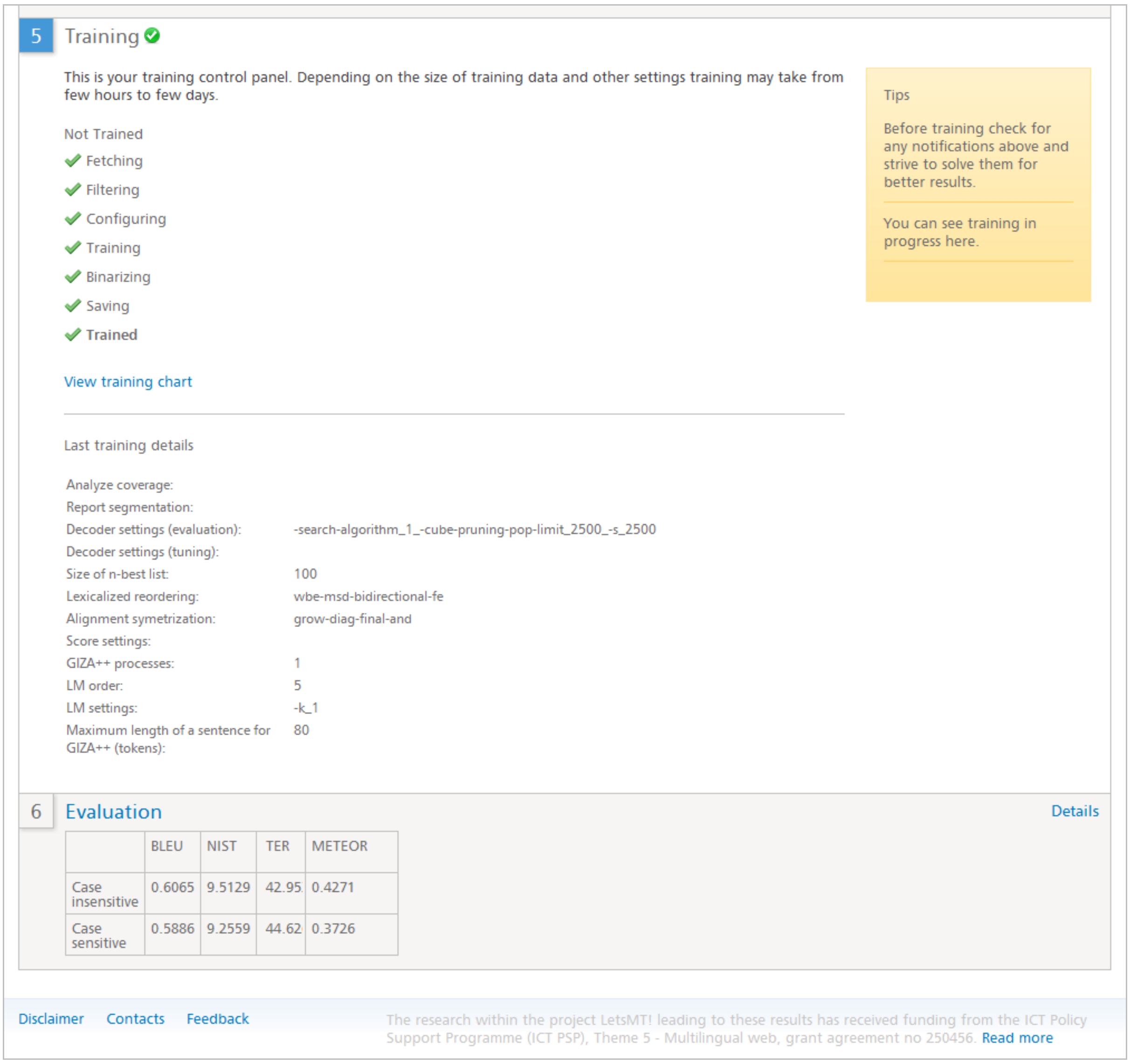
6 Evaluation Details

Disclaimer Contacts Feedback

The research within the project LetsMT! leading to these results has received funding from the ICT Policy Support Programme (ICT PSP), Theme 5 - Multilingual web, grant agreement no 250456. Read more

**Figure 22. SMT system Details view, Step 5 – Start training**

The trained SMT system is automatically evaluated. BLEU, NIST, TER and METEOR evaluation scores are calculated and shown in Step 6 after training is successfully completed (see Figure 23).



5 Training ✓

This is your training control panel. Depending on the size of training data and other settings training may take from few hours to few days.

Not Trained

- ✓ Fetching
- ✓ Filtering
- ✓ Configuring
- ✓ Training
- ✓ Binarizing
- ✓ Saving
- ✓ Trained

View training chart

Tips

Before training check for any notifications above and strive to solve them for better results.

You can see training in progress here.

Last training details

| | |
|---|---|
| Analyze coverage: | |
| Report segmentation: | |
| Decoder settings (evaluation): | -search-algorithm_1_-cube-pruning-pop-limit_2500_-s_2500 |
| Decoder settings (tuning): | |
| Size of n-best list: | 100 |
| Lexicalized reordering: | wbe-msd-bidirectional-fe |
| Alignment symetrization: | grow-diag-final-and |
| Score settings: | |
| GIZA++ processes: | 1 |
| LM order: | 5 |
| LM settings: | -k_1 |
| Maximum length of a sentence for GIZA++ (tokens): | 80 |

6 Evaluation

Details

| | BLEU | NIST | TER | METEOR |
|---|---|---|---|---|
| Case insensitive | 0.6065 | 9.5129 | 42.95 | 0.4271 |
| Case sensitive | 0.5886 | 9.2559 | 44.62 | 0.3726 |

The research within the project LetsMT! leading to these results has received funding from the ICT Policy Support Programme (ICT PSP), Theme 5 - Multilingual web, grant agreement no 250456. Read more

**Figure 23. SMT system Details view, Step 5 – Training.**

The LetsMT! provides another important feature - the **Training Chart** (Figure 24). It provides a detailed visualization of the training process. Several important training details are shown:

- steps queued for executing for a particular training task,
- current execution status of training steps,
- steps where errors occurred (if any).

Training chart becomes available when the training starts and remains available after the training.

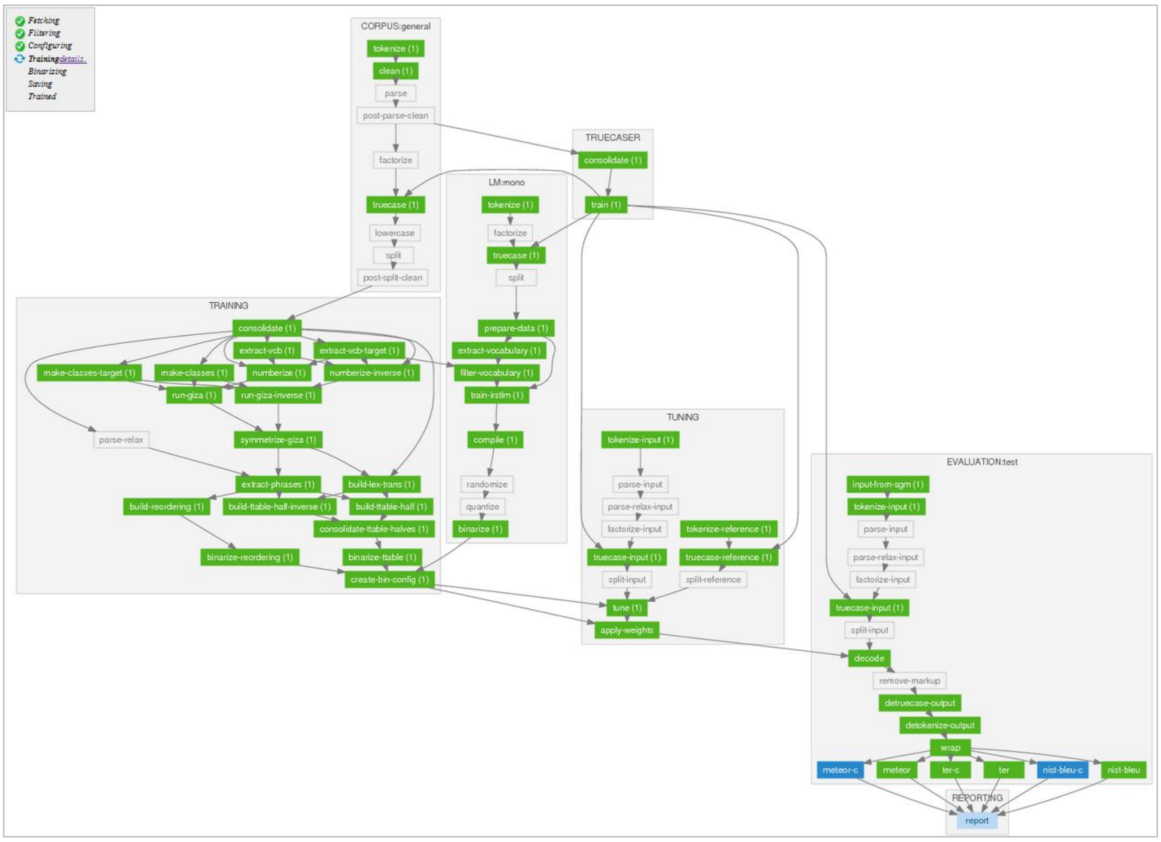


**Figure 24. SMT System Training Chart.**

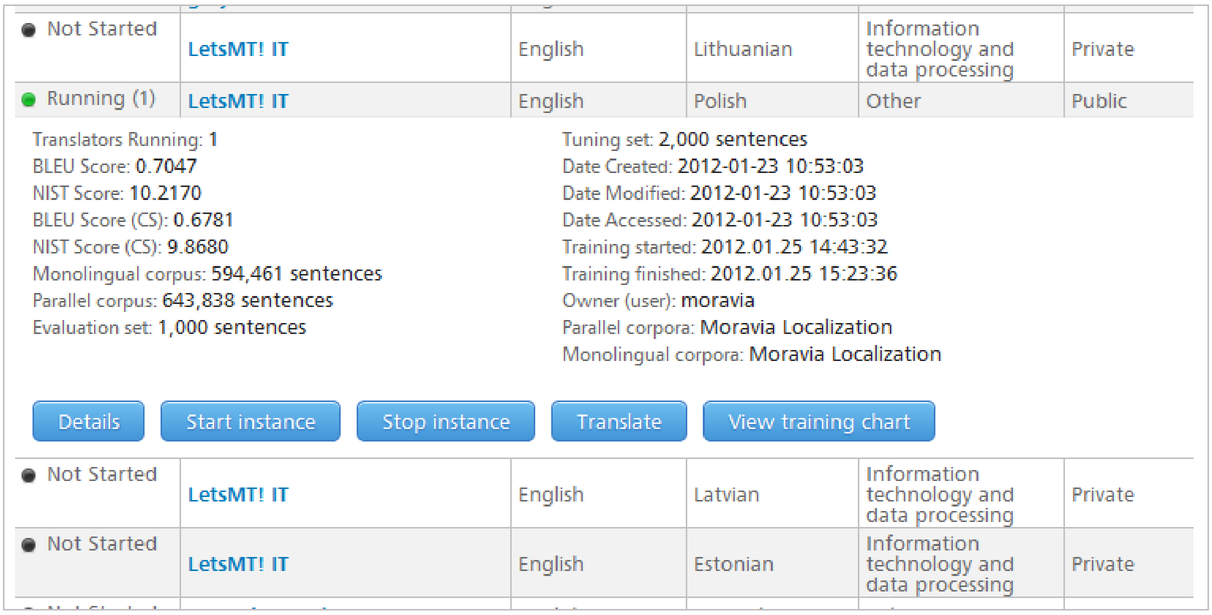
### 2.3.2.3 SMT System Running

When the SMT system is successfully trained, it needs to be started. The LetsMT! allows to start several system instances to speed up translating and balance the SMT system load.

It is possible to start an instance in two ways:

- In SMT system list, expanding the system and clicking “Start instance” (Figure 25).
- In SMT system’s detail view, clicking “Start instance” (Figure 26).

It is possible also to stop one or all running instances in a similar way.



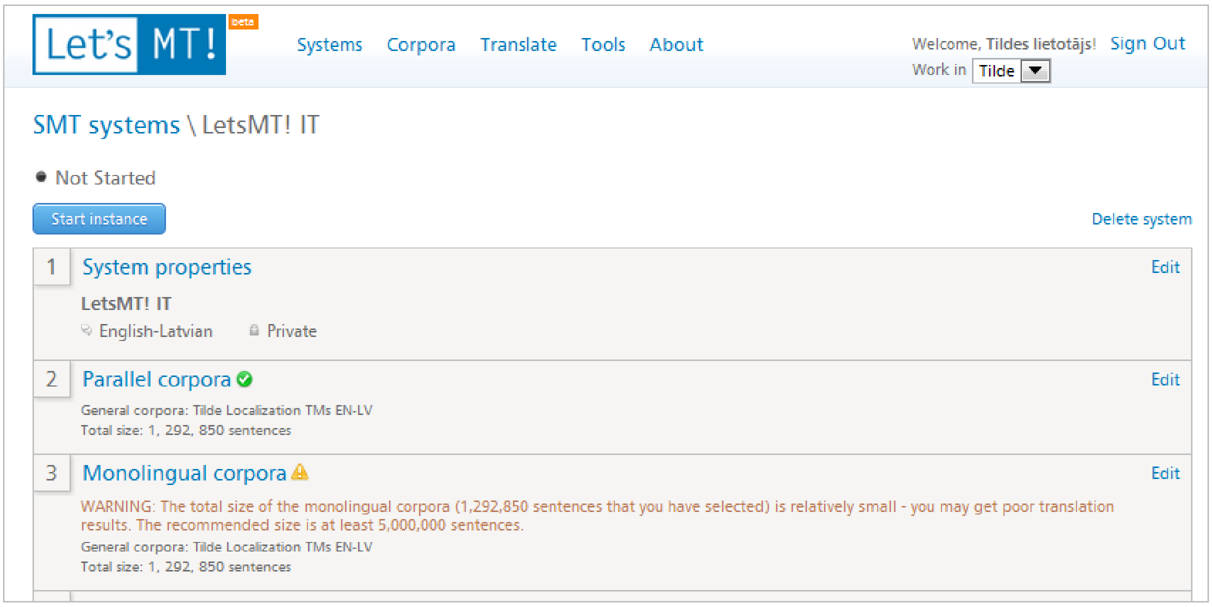
| | | | | | |
|---|---|---|---|---|---|
| ● Not Started | LetsMT! IT | English | Lithuanian | Information technology and data processing | Private |
| ● Running (1) | LetsMT! IT | English | Polish | Other | Public |

Translators Running: 1
BLEU Score: 0.7047
NIST Score: 10.2170
BLEU Score (CS): 0.6781
NIST Score (CS): 9.8680
Monolingual corpus: 594,461 sentences
Parallel corpus: 643,838 sentences
Evaluation set: 1,000 sentences

Tuning set: 2,000 sentences
Date Created: 2012-01-23 10:53:03
Date Modified: 2012-01-23 10:53:03
Date Accessed: 2012-01-23 10:53:03
Training started: 2012.01.25 14:43:32
Training finished: 2012.01.25 15:23:36
Owner (user): moravia
Parallel corpora: Moravia Localization
Monolingual corpora: Moravia Localization

| | | | | | |
|---|---|---|---|---|---|
| ● Not Started | LetsMT! IT | English | Latvian | Information technology and data processing | Private |
| ● Not Started | LetsMT! IT | English | Estonian | Information technology and data processing | Private |

**Figure 25. Starting an instance from system list expanded view.**



SMT systems \ LetsMT! IT

● Not Started

1 System properties

LetsMT! IT
English-Latvian   Private

2 Parallel corpora ✓

General corpora: Tilde Localization TMs EN-LV
Total size: 1, 292, 850 sentences

3 Monolingual corpora ⚠

WARNING: The total size of the monolingual corpora (1,292,850 sentences that you have selected) is relatively small - you may get poor translation results. The recommended size is at least 5,000,000 sentences.
General corpora: Tilde Localization TMs EN-LV
Total size: 1, 292, 850 sentences

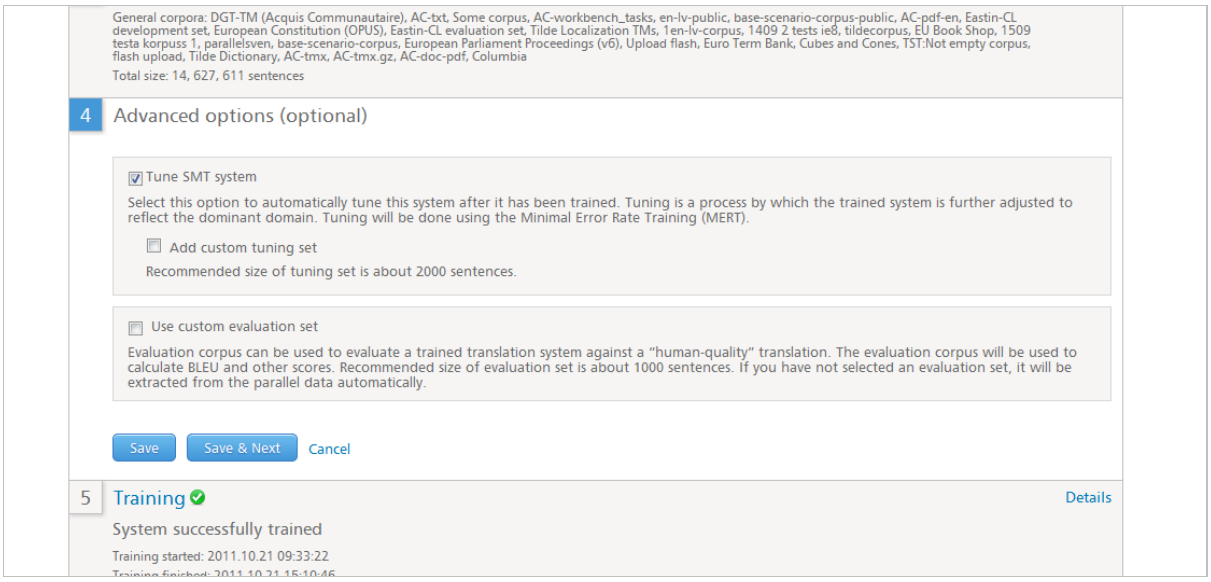
**Figure 26. Starting an instance from system's details view.**

### 2.3.2.4 Advanced Training Options

In *Advanced options* (optional) step user can select whether to use custom tuning and evaluation corpora in the training (see Figure 27).

If the user selects to tune the system and has not explicitly specified any tuning corpus then a random sample of 2,000 sentences will be automatically extracted from the training data and used for tuning (MERT), these sentences will not be used to build translation and language models.

If the user has not specified any evaluation corpus then a random sample of 1,000 sentences will be automatically extracted from the training data and used for evaluation (BLEU and other scores), these sentences will not be used for training.



General corpora: DGT-TM (Acquis Communautaire), AC-txt, Some corpus, AC-workbench_tasks, en-lv-public, base-scenario-corpus-public, AC-pdf-en, Eastin-CL development set, European Constitution (OPUS), Eastin-CL evaluation set, Tilde Localization TMs, 1en-lv-corpus, 1409 2 tests ie8, tildecorpus, EU Book Shop, 1509 testa korpuss 1, parallelsven, base-scenario-corpus, European Parliament Proceedings (v6), Upload flash, Euro Term Bank, Cubes and Cones, TST:Not empty corpus, flash upload, Tilde Dictionary, AC-tmx, AC-tmx.gz, AC-doc-pdf, Columbia

Total size: 14, 627, 611 sentences

4 Advanced options (optional)

☑ Tune SMT system

Select this option to automatically tune this system after it has been trained. Tuning is a process by which the trained system is further adjusted to reflect the dominant domain. Tuning will be done using the Minimal Error Rate Training (MERT).

□ Add custom tuning set

Recommended size of tuning set is about 2000 sentences.

□ Use custom evaluation set

Evaluation corpus can be used to evaluate a trained translation system against a "human-quality" translation. The evaluation corpus will be used to calculate BLEU and other scores. Recommended size of evaluation set is about 1000 sentences. If you have not selected an evaluation set, it will be extracted from the parallel data automatically.

5 Training ✓

System successfully trained

Training started: 2011.10.21 09:33:22

Training finished: 2011.10.21 15:10:46

**Figure 27. Advances SMT training options.**

### 2.3.3 Storing and Sharing Corpora

Corpora are the building blocks of SMT systems. LetsMT! provides simple but powerful corpora upload and management environment.

**All data imported into LetsMT! repository is kept there in repository format and can't be directly downloaded by any means. The uploaded data can be used only for SMT training.**

Corpora in LetsMT! are **logical containers** which has some meta-data and holds monolingual or parallel texts.

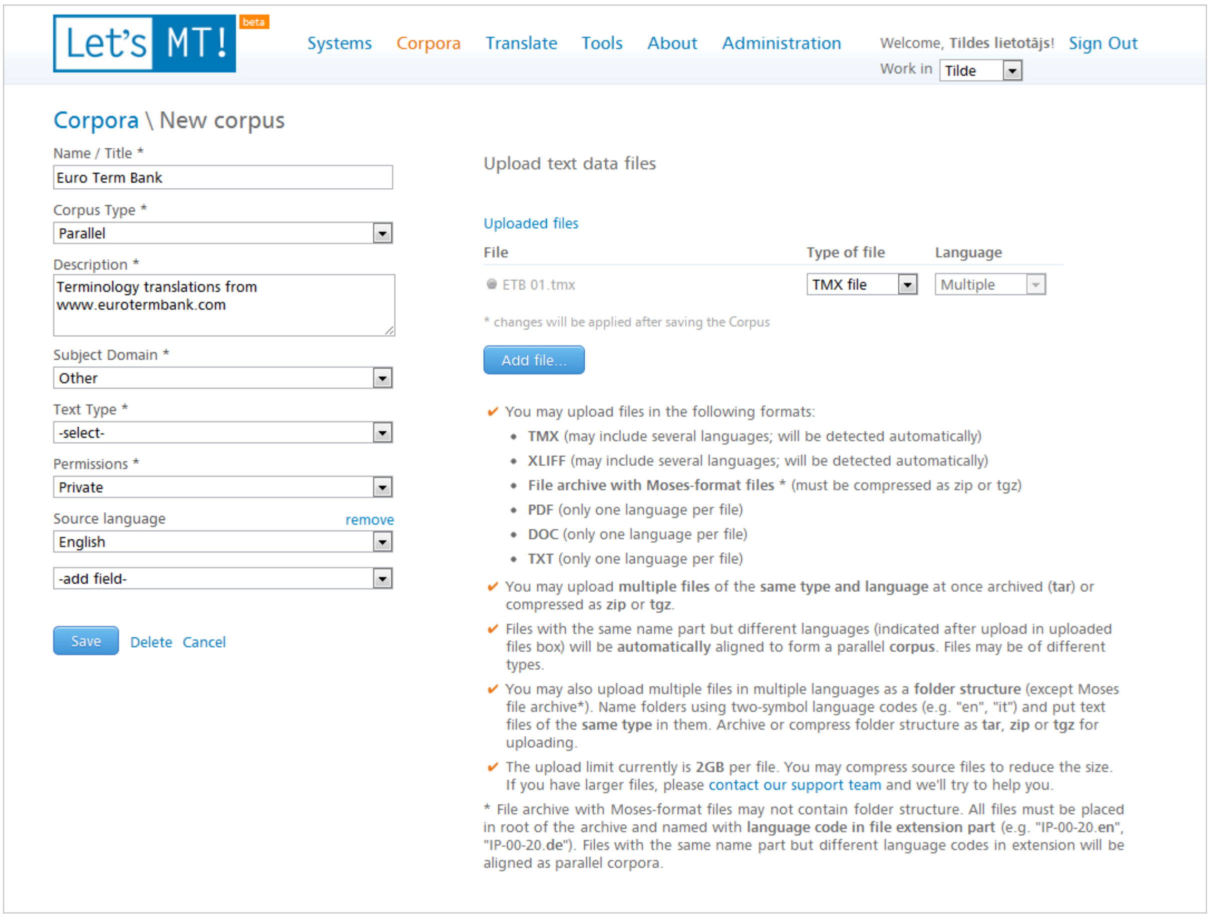
General corpus creation process consists of several steps:

1. Providing meta-data (title, corpus type, description, text type, permissions, etc.);
2. Uploading files and specifying language (if needed)
3. Saving corpus

When corpus is uploaded and saved, system:

- extracts files from the archive if they are archived,
- checks file validity (e.g. validates XML, checks encoding, etc.) and creates error reports in case of invalid files,
- converts files to system internal representation (This process is called importing and for big files it might take quite long time). This process is different for different upload formats and it may also include document aligning.
- updates status information,
- counts number of sentences and other statistics and saves this information in metadata.

New corpus creation is done in a single screen (see Figure 28).



Let's MT! beta
Systems Corpora Translate Tools About Administration
Welcome, Tildes lietotājs! Sign Out
Work in Tilde

Corpora \ New corpus

Name / Title *: Euro Term Bank

Corpus Type *: Parallel

Description *: Terminology translations from www.eurotermbank.com

Subject Domain *: Other

Text Type *: -select-

Permissions *: Private

Source language: English

remove
-add field-
Save Delete Cancel

Upload text data files

Uploaded files

| File | Type of file | Language |
|---|---|---|
| ETB 01.tmx | TMX file | Multiple |

* changes will be applied after saving the Corpus

✔ You may upload files in the following formats:

- **TMX** (may include several languages; will be detected automatically)
- **XLIFF** (may include several languages; will be detected automatically)
- **File archive with Moses-format files** * (must be compressed as zip or tgz)
- **PDF** (only one language per file)
- **DOC** (only one language per file)
- **TXT** (only one language per file)

✔ You may upload **multiple files** of the **same type and language** at once archived (**tar**) or compressed as **zip** or **tgz**.

✔ Files with the same name part but different languages (indicated after upload in uploaded files box) will be **automatically** aligned to form a parallel **corpus**. Files may be of different types.

✔ You may also upload multiple files in multiple languages as a **folder structure** (except Moses file archive*). Name folders using two-symbol language codes (e.g. "en", "it") and put text files of the **same type** in them. Archive or compress folder structure as **tar**, **zip** or **tgz** for uploading.

✔ The upload limit currently is **2GB** per file. You may compress source files to reduce the size. If you have larger files, please contact our support team and we'll try to help you.

* File archive with Moses-format files may not contain folder structure. All files must be placed in root of the archive and named with **language code in file extension part** (e.g. "IP-00-20.**en**", "IP-00-20.**de**"). Files with the same name part but different language codes in extension will be aligned as parallel corpora.

**Figure 28. Corpus creation and file upload.**

The system supports corpora upload up to 2GB per file in TMX, XLIFF, Moses plain text, PDF, DOC, TXT formats. Files can be archived as ZIP, TAR, TGZ or GZ files before uploading.

There are three different types of corpora which can be uploaded in LetsMT! – (i) localization file formats, e.g., translation memories, (2) Moses format and (3) document file formats. All these formats have their specific requirements for data preparation, upload, and processing. The detailed description about each of them is given in next chapters.

### 2.3.3.1 Uploading corpora in localization file formats (TMX/XLIFF)

The LetsMT! platform allows localization companies and other users to prepare and upload corpora in standard data exchange formats widely used in localization – TMX, XLIFF.

TMX (Translation Memory eXchange) is an open XML standard for the exchange of translation memory data. It may contain both monolingual text (in one or more languages) and parallel text (in two or more languages).

XLIFF (XML Localization Interchange File Format) is a vocabulary format to store localizable data and carry it from one step of the localization process to the other, while allowing interoperability between tools. It may contain monolingual text in one language or parallel text in one language pair.

TMX and XLIFF formats hold some meta-information about to data. The LetsMT! corpora upload mechanism employs advantage of the meta-information by reading language information for each translation unit simplifying upload process for the user.

Each corpus in LetsMT! may contain one or more TMX or XLIFF files. User can upload them all at one or one by one. User can add more data the corpus at any time and user can replace previously uploaded files with the new files.

To speed up the data upload process the user can put TMX and XLIFF files in ZIP, TAR, GZ or TGZ archives. One or more TMX or XLIFF files are allowed in one archive file.

#### 2.3.3.2 Uploading corpora in Moses file format

The LetsMT! platform allows academic and other users to prepare and upload parallel corpora in a format used in Moses SMT framework[1]. Training data has to be provided sentence aligned (one sentence per line), in two files, one for the source language sentences, one for the target language sentences. Files must be encoded in UTF-8 encoding without BOM. Both files must have the same file name and use language code (ISO 639-1, e.g. "en", "fr") as the file extension (e.g., file.en, file.fr). Both files in Moses format must be archived in one ZIP, TAR or TGZ archive to be uploaded to the LetsMT! User can also put more than two files in Moses format in a single archive (e.g., file1.en, file1.fr, file2.en, file2.fr) and user can create a corpus in LetsMT! platform which contains more than one archive with files in Moses format.

#### 2.3.3.3 Uploading and auto-aligning document file formats

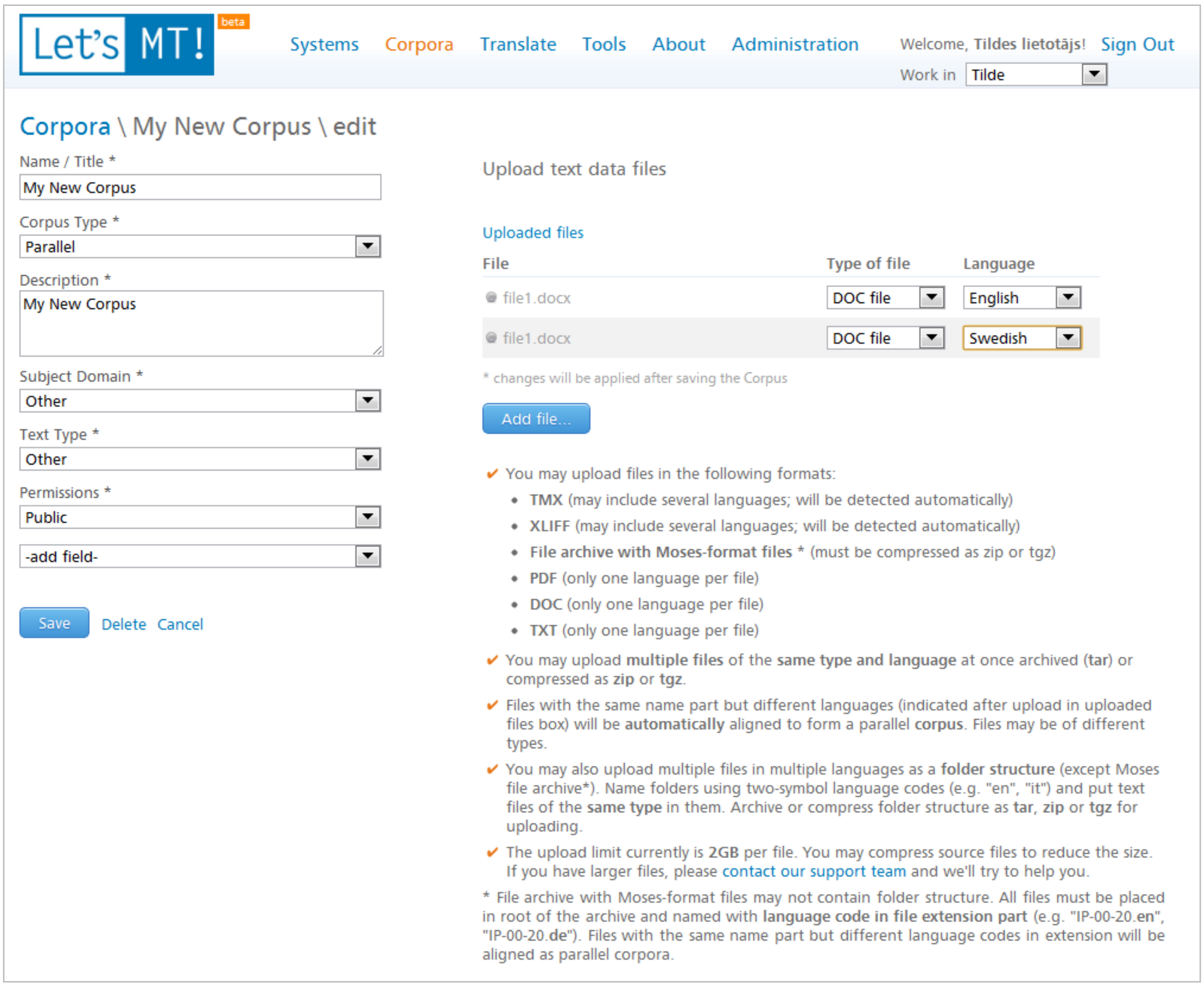
The LetsMT! platform allows users to make a corpus from a collection of translated documents. If user has a set of documents in DOC, DOCX, PDF or TXT format and he has these documents translated in other language, then he can create the parallel corpus using these documents.

There are 3 ways how user can upload documents and get them automatically aligned.

1. User can upload documents one by one specifying the language for each document (See Figure 29). Documents with the same filename (but in different languages) will be automatically aligned.
2. User can archive all one language documents in one ZIP, TAR or TGZ archive and other language document in other archive. Then he can upload both archives specifying the document type and the language for each of them. All documents with the same filename but in different archive files will be automatically aligned. (See Figure 30)
3. User can put all one language documents in one directory named using language code ((ISO 639-1, e.g. "en", "fr") and other language document in other directory. Then he can archive both directories in a single ZIP, TAR or TGZ archive and upload it. In this case user must specify language “Multiple” (See Figure 31). All documents with the same filename but in different directories will be automatically aligned.

Using all 3 methods described above user can create also parallel corpora in more than two languages. And user can automatically align files with different file formats as well, for example, file MYFILE.DOC can be aligned with MYFILE.PDF.

[1] http://www.statmt.org/moses/?n=FactoredTraining.PrepareTraining



Work in Tilde

Corpora \ My New Corpus \ edit

Name / Title *: My New Corpus

Corpus Type *: Parallel

Description *: My New Corpus

Subject Domain *: Other

Text Type *: Other

Permissions *: Public

-add field-

Upload text data files

Uploaded files

| File | Type of file | Language |
|---|---|---|
| file1.docx | DOC file | English |
| file1.docx | DOC file | Swedish |

* changes will be applied after saving the Corpus

✔ You may upload files in the following formats:

- **TMX** (may include several languages; will be detected automatically)
- **XLIFF** (may include several languages; will be detected automatically)
- **File archive with Moses-format files** * (must be compressed as zip or tgz)
- **PDF** (only one language per file)
- **DOC** (only one language per file)
- **TXT** (only one language per file)

✔ You may upload **multiple files** of the **same type and language** at once archived (**tar**) or compressed as **zip** or **tgz**.

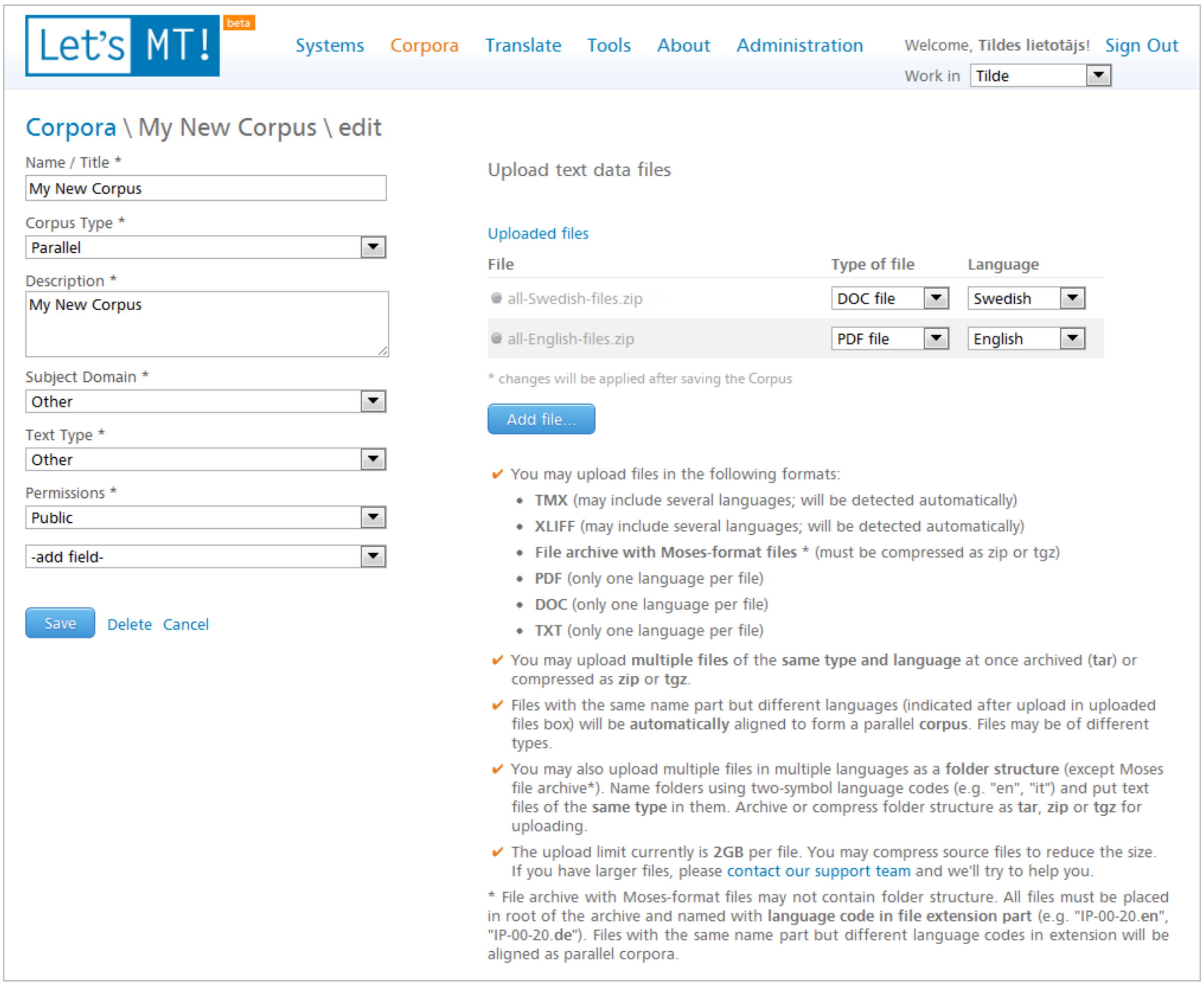
✔ Files with the same name part but different languages (indicated after upload in uploaded files box) will be **automatically** aligned to form a parallel **corpus**. Files may be of different types.

✔ You may also upload multiple files in multiple languages as a **folder structure** (except Moses file archive*). Name folders using two-symbol language codes (e.g. "en", "it") and put text files of the **same type** in them. Archive or compress folder structure as **tar**, **zip** or **tgz** for uploading.

✔ The upload limit currently is **2GB** per file. You may compress source files to reduce the size. If you have larger files, please contact our support team and we'll try to help you.

* File archive with Moses-format files may not contain folder structure. All files must be placed in root of the archive and named with **language code in file extension part** (e.g. "IP-00-20.**en**", "IP-00-20.**de**"). Files with the same name part but different language codes in extension will be aligned as parallel corpora.

**Figure 29. Uploading corpus files one by one.**



Corpora \ My New Corpus \ edit

Name / Title *: My New Corpus

Corpus Type *: Parallel

Description *: My New Corpus

Subject Domain *: Other

Text Type *: Other

Permissions *: Public

-add field-

Upload text data files

Uploaded files

| File | Type of file | Language |
|---|---|---|
| all-Swedish-files.zip | DOC file | Swedish |
| all-English-files.zip | PDF file | English |

* changes will be applied after saving the Corpus

✔ You may upload files in the following formats:

- **TMX** (may include several languages; will be detected automatically)
- **XLIFF** (may include several languages; will be detected automatically)
- **File archive with Moses-format files** * (must be compressed as zip or tgz)
- **PDF** (only one language per file)
- **DOC** (only one language per file)
- **TXT** (only one language per file)

✔ You may upload **multiple files** of the **same type and language** at once archived (**tar**) or compressed as **zip** or **tgz**.

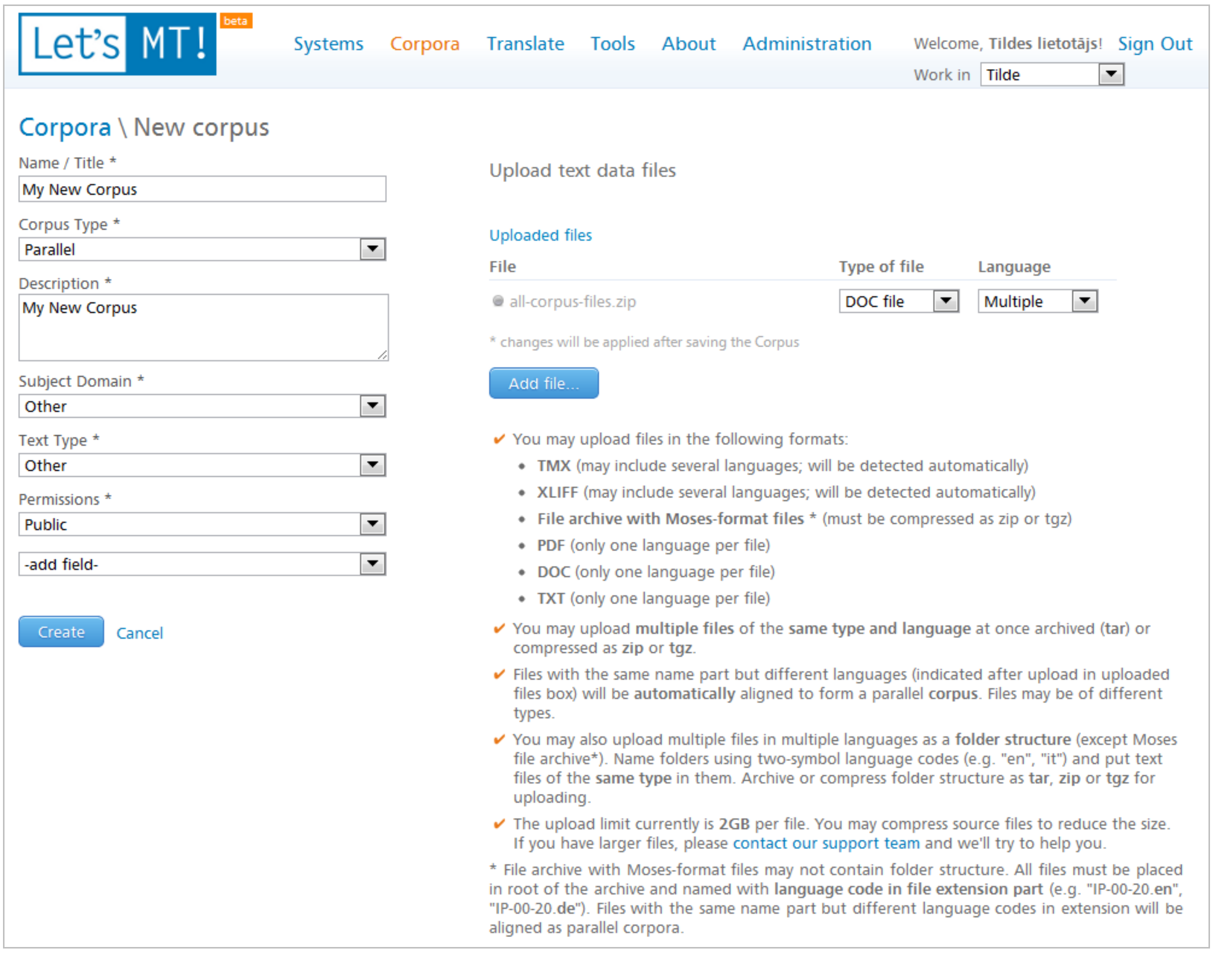
✔ Files with the same name part but different languages (indicated after upload in uploaded files box) will be **automatically** aligned to form a parallel **corpus**. Files may be of different types.

✔ You may also upload multiple files in multiple languages as a **folder structure** (except Moses file archive*). Name folders using two-symbol language codes (e.g. "en", "it") and put text files of the **same type** in them. Archive or compress folder structure as **tar**, **zip** or **tgz** for uploading.

✔ The upload limit currently is **2GB** per file. You may compress source files to reduce the size. If you have larger files, please contact our support team and we'll try to help you.

* File archive with Moses-format files may not contain folder structure. All files must be placed in root of the archive and named with **language code in file extension part** (e.g. "IP-00-20.**en**", "IP-00-20.**de**"). Files with the same name part but different language codes in extension will be aligned as parallel corpora.

**Figure 30. Uploading corpus as archived document files (each language files in a separate archive)**



Welcome, **Tildes lietotājs!**

Work in: Tilde

Corpora \ New corpus

Name / Title *: My New Corpus

Corpus Type *: Parallel

Description *: My New Corpus

Subject Domain *: Other

Text Type *: Other

Permissions *: Public

-add field-

Upload text data files

Uploaded files

| File | Type of file | Language |
|---|---|---|
| all-corpus-files.zip | DOC file | Multiple |

* changes will be applied after saving the Corpus

✓ You may upload files in the following formats:

- **TMX** (may include several languages; will be detected automatically)
- **XLIFF** (may include several languages; will be detected automatically)
- **File archive with Moses-format files** * (must be compressed as zip or tgz)
- **PDF** (only one language per file)
- **DOC** (only one language per file)
- **TXT** (only one language per file)

✓ You may upload **multiple files** of the **same type and language** at once archived (**tar**) or compressed as **zip** or **tgz**.

✓ Files with the same name part but different languages (indicated after upload in uploaded files box) will be **automatically** aligned to form a parallel **corpus**. Files may be of different types.

✓ You may also upload multiple files in multiple languages as a **folder structure** (except Moses file archive*). Name folders using two-symbol language codes (e.g. "en", "it") and put text files of the **same type** in them. Archive or compress folder structure as **tar**, **zip** or **tgz** for uploading.

✓ The upload limit currently is **2GB** per file. You may compress source files to reduce the size. If you have larger files, please contact our support team and we'll try to help you.

* File archive with Moses-format files may not contain folder structure. All files must be placed in root of the archive and named with **language code in file extension part** (e.g. "IP-00-20.**en**", "IP-00-20.**de**"). Files with the same name part but different language codes in extension will be aligned as parallel corpora.

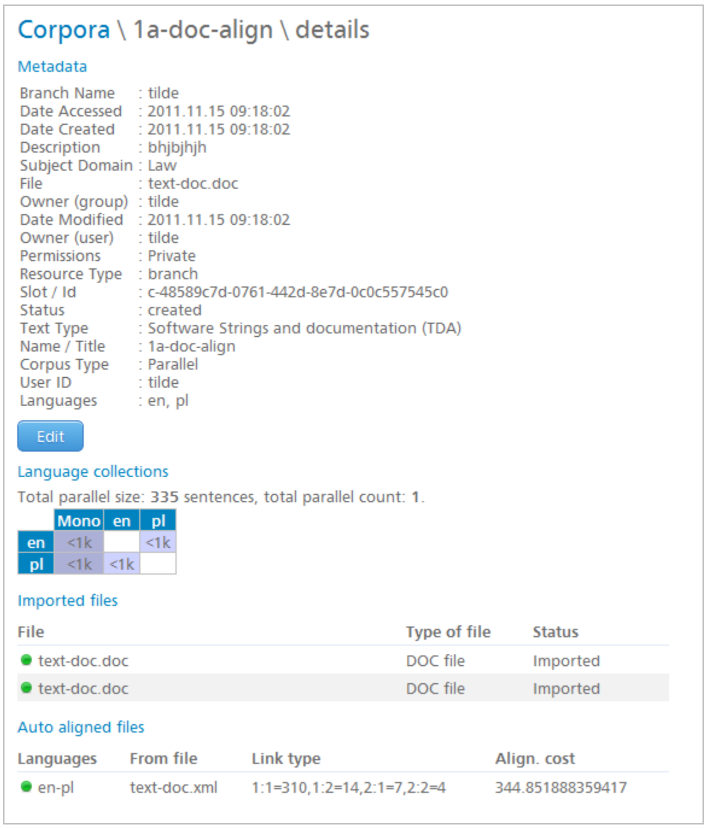
**Figure 31. Uploading the whole corpus as archived directories with files**

All archives may only contain files of the same type and language.

When user uploads files in document file format for the automatic alignment, the system (i) extracts text from the document, (ii) breaks the text into sentences, (iii) finds two matching documents and (iv) aligns them at the sentence level.

The user can also upload files in document file formats to create monolingual corpus, in this case system will just extract text from documents and break it into sentences.

Import and alignment results are shown in corpus details view (see Figure 32), where the user can see status of each uploaded file and detailed information about automatic alignment of documents.



Corpora \ 1a-doc-align \ details

Metadata

| | |
|---|---|
| Branch Name | : tilde |
| Date Accessed | : 2011.11.15 09:18:02 |
| Date Created | : 2011.11.15 09:18:02 |
| Description | : bhjbjhjh |
| Subject Domain | : Law |
| File | : text-doc.doc |
| Owner (group) | : tilde |
| Date Modified | : 2011.11.15 09:18:02 |
| Owner (user) | : tilde |
| Permissions | : Private |
| Resource Type | : branch |
| Slot / Id | : c-48589c7d-0761-442d-8e7d-0c0c557545c0 |
| Status | : created |
| Text Type | : Software Strings and documentation (TDA) |
| Name / Title | : 1a-doc-align |
| Corpus Type | : Parallel |
| User ID | : tilde |
| Languages | : en, pl |

Language collections

Total parallel size: 335 sentences, total parallel count: 1.

| | Mono | en | pl |
|---|---|---|---|
| en | <1k | | <1k |
| pl | <1k | <1k | |

Imported files

| File | Type of file | Status |
|---|---|---|
| ● text-doc.doc | DOC file | Imported |
| ● text-doc.doc | DOC file | Imported |

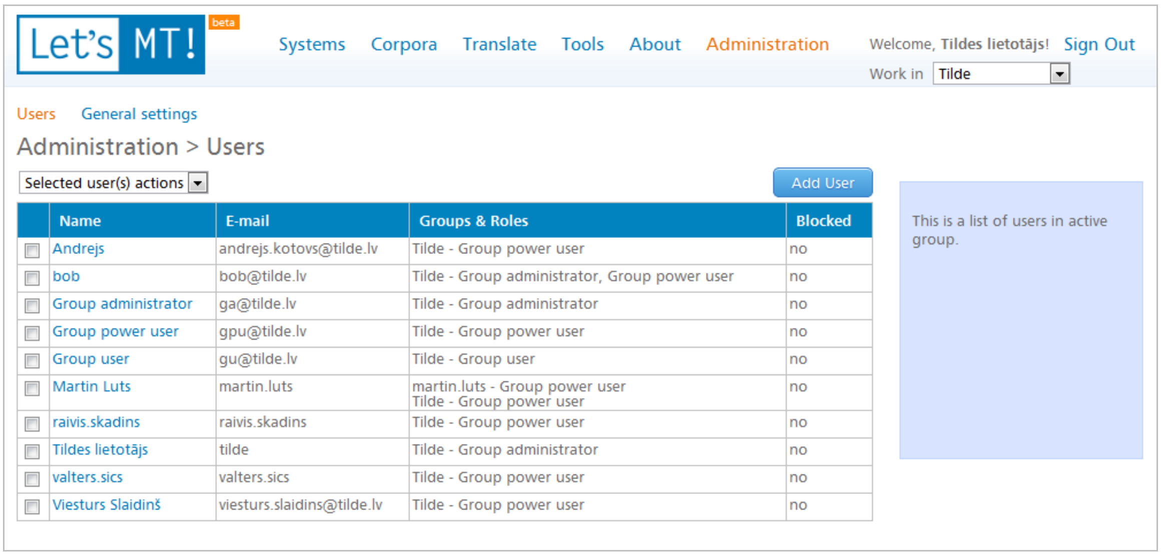
Auto aligned files

| Languages | From file | Link type | Align. cost |
|---|---|---|---|
| ● en-pl | text-doc.xml | 1:1=310,1:2=14,2:1=7,2:2=4 | 344.851888359417 |

**Figure 32. The status of uploaded files and automatic aligning.**

## 2.3.4 Managing users

LetsMT! provides Group and User management interface (see Figure 33).



Welcome, Tildes lietotājs!

Work in Tilde

Administration > Users

| | Name | E-mail | Groups & Roles | Blocked |
|---|---|---|---|---|
| □ | Andrejs | andrejs.kotovs@tilde.lv | Tilde - Group power user | no |
| □ | bob | bob@tilde.lv | Tilde - Group administrator, Group power user | no |
| □ | Group administrator | ga@tilde.lv | Tilde - Group administrator | no |
| □ | Group power user | gpu@tilde.lv | Tilde - Group power user | no |
| □ | Group user | gu@tilde.lv | Tilde - Group user | no |
| □ | Martin Luts | martin.luts | martin.luts - Group power user<br>Tilde - Group power user | no |
| □ | raivis.skadins | raivis.skadins | Tilde - Group power user | no |
| □ | Tildes lietotājs | tilde | Tilde - Group administrator | no |
| □ | valters.sics | valters.sics | Tilde - Group power user | no |
| □ | Viesturs Slaidinš | viesturs.slaidins@tilde.lv | Tilde - Group power user | no |

This is a list of users in active group.

**Figure 33. User management screen.**

User may be assigned one or more roles in one or more groups (see 2.4.1 *User Accounts*). A user has a unique identifier in LetsMT! system – e-mail address.

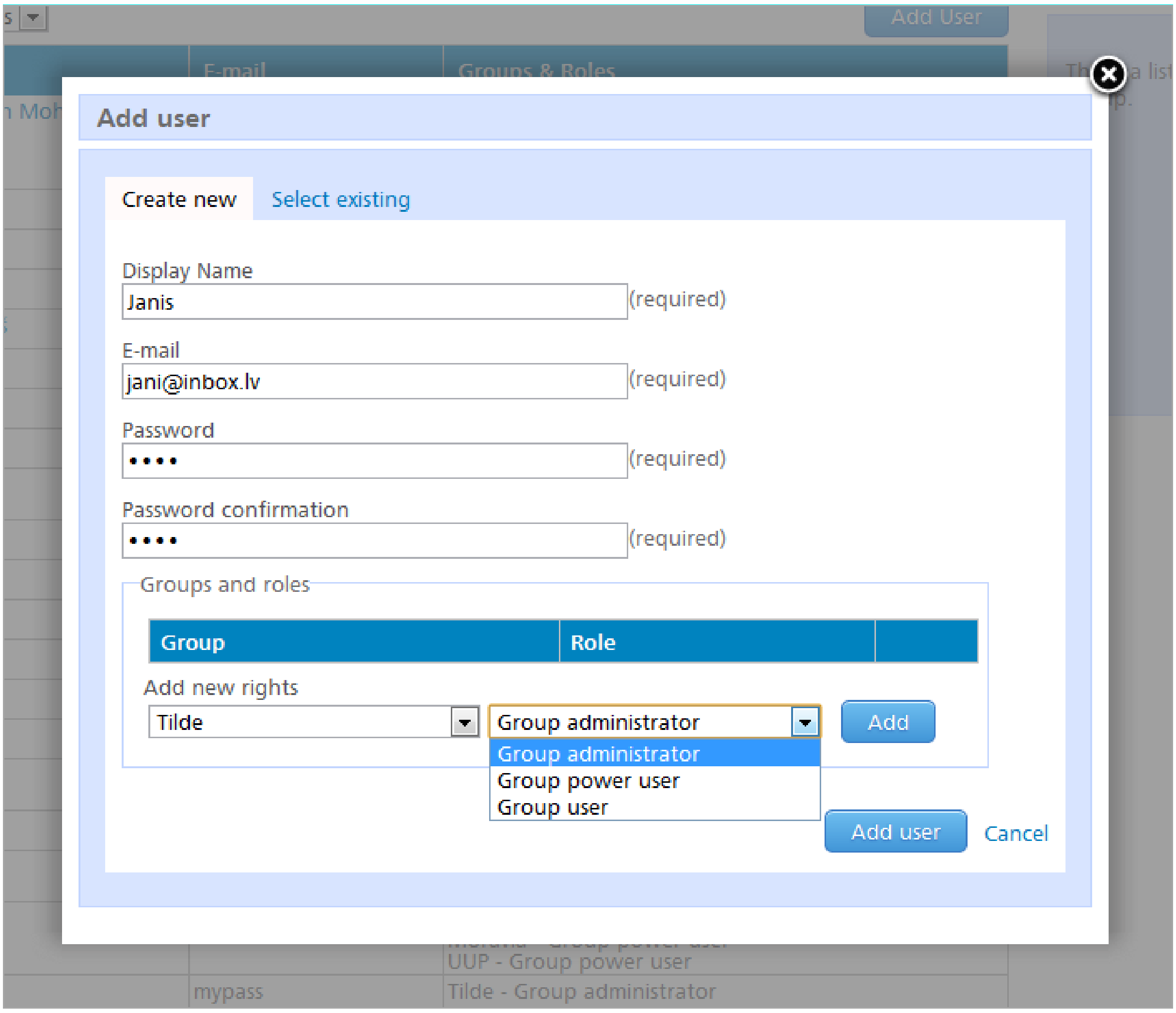
Group Administrator (GA) and System Administrator (SA) roles can manage users in groups.

LetsMT! provides two ways of adding a new user to current working group:

1. Creating a new LetsMT! account for new user (see Figure 34) and
2. Adding existing LetsMT! user.

Creating a new account is done in the following steps:

- click "Add User" and select "Create new" tab (selected by default),
- enter user's name, e-mail address and password (twice),
- add at least one permission to some group by choosing user's role,
- click "Add user".



Add user

Create new | Select existing

Display Name: Janis (required)

E-mail: jani@inbox.lv (required)

Password: •••• (required)

Password confirmation: •••• (required)

Groups and roles

| Group | Role |
|---|---|

Add new rights: Tilde | Group administrator

**Figure 34. New user creation form.**

Adding existing LetsMT! user is done in the following steps:

- Enter valid e-mail address of existing LetsMT! user,
- click "Add user".

Editing user data is available only to group administrators (GA) of user's primary group. User's primary group is assigned when user is created and is set to current working group.

## 2.4 *System security overview*

LetsMT! has user authentication and authorization mechanisms based on user names (e-mail addresses) and passwords. User authorization is used to control:

- Login to website and rights to access functions available through web page interface,
- Access rights to training data (corpora), trained models and SMT systems,
- Initiation and management of training tasks,
- Access through external APIs (data upload, translation, CAT tools)

The system employs a security infrastructure in order to meet multi-tenancy, IPR and reliability requirements.

Users have to identify themselves when they access the system. The system keeps the record what kind of user or external agent is accessing it and to which service subscriber it belongs. The external systems have to send not only translation queries and other commands but also information necessary for authentication.

Authenticated users access the LetsMT! system via a secure internet connection (HTTPS[2]) to ensure a secure transfer of authentication information, data, translation requests and translations.

### 2.4.1 User Accounts, Roles and Groups

By default, users access the system through a web interface and get access permissions of the 'Anonymous User' role. Anonymous users can enter user name and password and get authenticated. Authenticated users belong to one or more user groups. Each group is associated with LetsMT! service subscriber. The LetsMT! system has several user roles defining which functions are allowed to the user.

The following system user roles are defined for the LetsMT! system:

- Anonymous User (AN) – Any user from the Internet. No authentication is done. Can access public information and use public SMT systems for translation.
- User (U) – User who can use trained SMT systems. He can access both public SMT systems and SMT systems allowed to his group. He can use SMT systems both using the LetsMT! web interface and an API.
- Power User (PU) – User who manages training data, training tasks and trained SMT systems of his group.
- Group Administrator (GA) – User who has full control over his group. He manages corpora, SMT systems and users in the group. He can grant rights also to *Power User*.
- System Administrator (SA) – User who has the maximum allowed rights. Can manage groups (LetsMT! service subscribers); manage all users, resources, training tasks, trained SMT systems etc.

New group with one user (group administrator) gets created each time when new service subscriber gets registered. Group administrator can create new accounts for other users

[2] HTTPS: http://en.wikipedia.org/wiki/HTTP_Secure

(*Users* and *Power Users*) in his group. All users belonging to the group can access only public resources and resources belonging to their group.

Changes in users and roles assignments are stored in audit log.

### 2.4.2 Permission matrix

The following table lists entities and functions that can be accessed by pre-defined roles.

**Table 4. Function-to-role authorization matrix**

| **Permission\Role** | **AN** | **U** | **PU** | **GA** | **SA** |
|---|---|---|---|---|---|
| List public corpora | √ | √ | √ | √ | √ |
| List public systems | √ | √ | √ | √ | √ |
| Translate using public SMT systems | √ | √ | √ | √ | √ |
| Translate using private* SMT systems |  | √ | √ | √ | √ |
| List private corpora |  | √ | √ | √ | √ |
| List private SMT systems |  | √ | √ | √ | √ |
| Edit private corpora |  |  | √ | √ | √ |
| Delete private corpora |  |  | √ | √ | √ |
| Upload private or public corpora |  |  | √ | √ | √ |
| Start/stop private SMT systems |  |  | √ | √ | √ |
| View details of private SMT systems |  |  | √ | √ | √ |
| Create new SMT system |  |  | √ | √ | √ |
| Delete private SMT systems |  |  | √ | √ | √ |
| Edit private SMT systems |  |  | √ | √ | √ |
| Train private or public SMT system |  |  | √ | √ | √ |
| List users of group |  |  |  | √ | √ |
| Add users in group |  |  |  | √ | √ |
| Delete users from group |  |  |  | √ | √ |
| Edit users from group |  |  |  | √ | √ |
| List users of system |  |  |  |  | √ |
| Add users in system |  |  |  |  | √ |
| Delete users from system |  |  |  |  | √ |
| Edit users of system |  |  |  |  | √ |
| List groups of the system |  |  |  |  | √ |
| Add groups in system |  |  |  |  | √ |
| Delete groups of system |  |  |  |  | √ |
| Edit groups of system |  |  |  |  | √ |

* here private means – corpora and systems which are allowed to the group user belongs.

# 3 Logical Design

## 3.1 *Overview of System Architecture*

The LetsMT! system has a multi-tier architecture. It has (i) an interface layer for user interface and APIs with external systems; (ii) an application logic layer for the system logic, and (iii) a data storage layer consisting of file and database storage. The LetsMT! system performs various time and resource consuming tasks; these tasks are defined by the application logic and the data storage and are sent to the High Performance Computing (HPC) Cluster for execution.

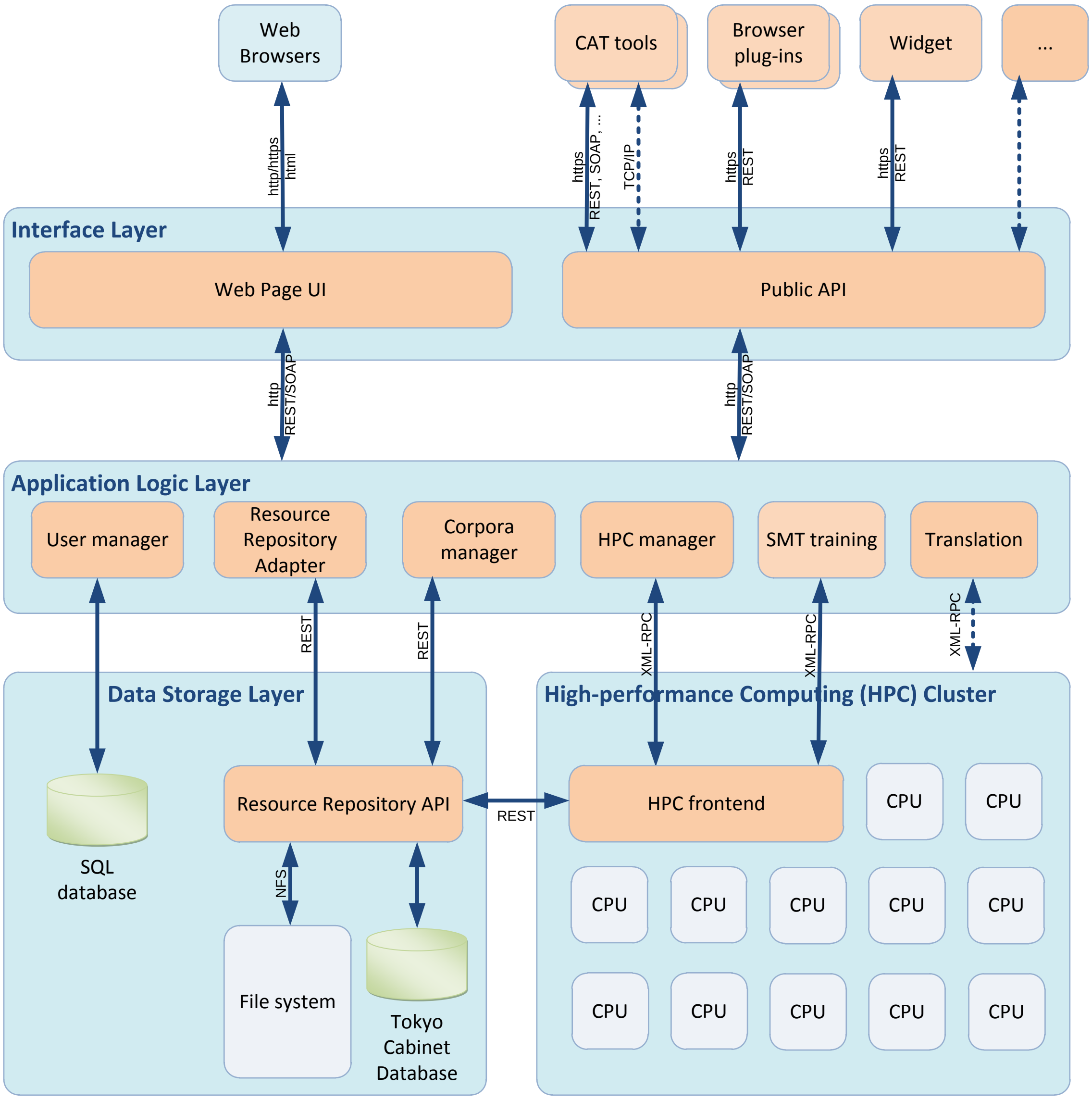


**Figure 35. Software architecture of LetsMT! platform**

The interface layer provides interfaces between the LetsMT! system and external users. The system has both human and machine users. Human users will access the system through web browsers by using the LetsMT! web page interface. External systems such as CAT tools and browser plug-ins will access the LetsMT! system through a public API. The public API is available through both REST[3] (with XML or JSON[4] serialization) and SOAP[5] protocol web services. Some CAT tools or other external systems may require different interfaces; they might be introduced if necessary. A secure HTTPS protocol is used to ensure secure user authentication and secure data transfer.

An application logic layer contains a set of modules responsible for the main functionality or logic of the systems. It receives queries and commands from the interface layer and prepares answers or performs tasks using data storage and the HPC cluster. This layer contains several modules such as the User Manager, the Resource Repository Adapter, the Corpora Manager, the HPC Manager, the SMT Training Manager and the Translation Manage. The interface layer accesses the application logic layer services through both REST and SOAP protocol.

The LetsMT! system as a data sharing and MT platform stores a huge amount of SMT training data (parallel and monolingual corpora) as well as trained models of SMT systems. The data is stored in one central Resource Repository (RR). The RR is also used to store SMT systems. A key-value store in RR is used to keep metadata and statistics about training data, about trained SMT systems, and about internal RR users and data access management. Modules from the application logic layer and HPC cluster access RR through a REST-based web service interface.

A HPC cluster is used to execute many different data processing tasks, corpora import, SMT system training, and running trained SMT systems. Modules from the application logic layer and RR create jobs and send them to the HPC cluster to execute. HPC cluster is responsible for accepting, scheduling, dispatching, and managing the remote and distributed execution of large numbers of jobs. The LetsMT! HPC cluster is based on Oracle Grid Engine[6] (SGE). The HPC cluster accesses data stored in the data storage layer using the Resource Repository API.

## *3.2 Interface Layer*

The Interface Layer consists of two main modules (see Figure 36. Interface Layer and communication with client tools of the LetsMT! platform on page 48): the Web Page User Interface (UI) and the Application Programming Interface (API) for translation automation and integration with external applications. Both are available directly from Internet.

[3] REST: http://en.wikipedia.org/wiki/Representational_State_Transfer

[4] JSON: http://www.json.org/, http://en.wikipedia.org/wiki/JSON

[5] SOAP: http://www.w3.org/TR/soap/, http://en.wikipedia.org/wiki/SOAP

[6] Oracle Grid Engine, previously known as Sun Grid Engine (SGE): http://gridengine.org/, http://en.wikipedia.org/wiki/Sun_Grid_Engine

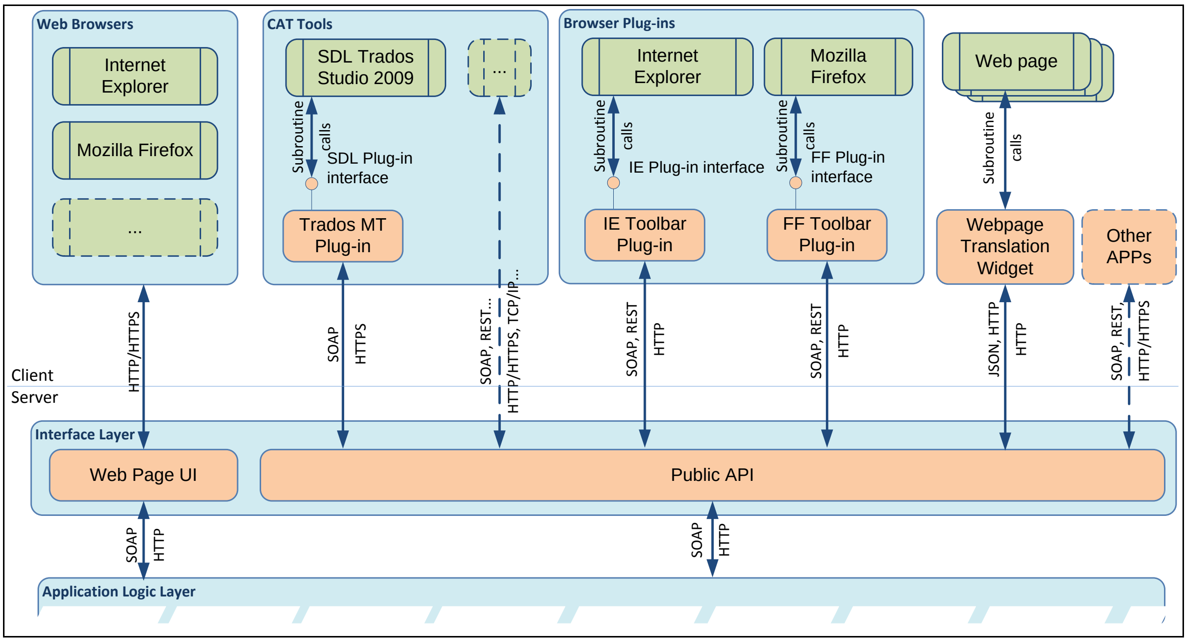


**Figure 36. Interface Layer and communication with client tools of the LetsMT! platform**

### 3.2.1 Web Page UI

All LetsMT! services are accessible through the Web page User Interface (UI). Business and professional users of the localization and translation industry may use the non-public part of the LetsMT! platform for uploading their parallel and monolingual corpora, building custom SMT systems from the specified collections of training data and accessing these SMT systems in their productivity environments (typically, various CAT tools). All services are available using **TLS/SSL**[7] cryptographic methods to protect users' content that is sent over the network.

Web page UI provides following main functionality:

- User authentication
- Corpus file upload
- Training data management
- SMT system training
- SMT system management
- Translating
- Data sharing including corpora and SMT systems
- User management

[7] http://en.wikipedia.org/wiki/Transport_Layer_Security

### 3.2.1.1 Uploading large files

The Web Page UI supports uploading large data volumes over the Internet.

Large file upload is related to bilingual and monolingual training and evaluation data uploads. Many different file formats are relevant in connection with upload data. Supported upload file formats are TMX, XLIFF, TXT or plain text files, PDF, DOC, DOCX, and Moses archive format files.

Together with uploaded corpora files metadata should be provided. See deliverable “D4.1 Specification of administration of training data and IPR draft agreement” for meta-data type requirements.

### 3.2.1.2 SMT system training

Training job management of the Application Logic layer is accessible through the Web Page UI. A user is able (depending on security restrictions) to define SMT system configuration by selecting:

- Translation direction (source/target language),
- Monolingual and bilingual corpora,
- Training and evaluation corpora,
- Access level of the system (weather it will be public or restricted for use within particular LetsMT! service subscriber),
- System’s name, description and other tags,
- And additional metadata and training options could be specified.

For all defined system configurations appropriate SMT system training job is queued using the Training job management.

### 3.2.1.3 SMT system browsing

When a user wants to translate a text or perform other actions on a SMT system he/she first has to locate translation system to use. A user is able to see all available SMT system, depending on its access rights.

Various filters by source/target language, domain and other metadata are implemented to relieve system selection.

User permissions and LetsMT! service subscriber membership restrictions is taken into account when providing SMT system list to user.

### 3.2.1.4 Translating

Web page UI contains website for text translation. The users of the website are able to enter a text for translation and translate it with selected SMT system. The text for translation is passed to Translation Manager of the Application Logic Layer which distribute the text over the HPC Cluster for translation with the SMT system.

### 3.2.2 Public API

The Public API is a public interfacing component that provides LetsMT! functionality to external applications like CAT tools, webpage translation widgets, web browsers and other translation applications that might be integrated with LetsMT! services. Web developers are able to use this API to integrate LetsMT! services in their products. The Public API is able to communicate with external application using SOAP, REST (using XML or JSON serialization) over HTTPS protocol.

In general, The Public API mainly works as a proxy of the Application Logic Layer components providing subset of its functionality. Additionally it can serve specific custom functions based on functionality of the Application Logic Layer or do data transformations that are necessary for integration with external applications, e.g., CAT tools.

The Public API serves functionality that is associated with a translation process:

- User authentication;
- Translation system querying;
- Translation;
- Sentence breaking.

The Public API is implemented as a public web service that will be accessible directly from external applications. Detailed description of LetsMT! public API is available in deliverable “D3.4 SMT Web Service Ready for Integration”

### 3.2.3 Authentication

When Web Page UI and Public API users access the system they have to identify themselves. The users have to send their username and password before accessing non-public functionality and resources.

Web pages UI uses HTML Form based authentication[8] method, which is supported by all web browsers. Public API uses basic access authentication[9] that is part of the HTTP 1.1 protocol. To make both mentioned authentication methods secure those are combined with cryptographic methods of application layer like TLS/SSL.

## *3.3 Client components*

The solution delivers the following client components to integrate LetsMT! services with external tools:

- translation widget provided for inclusion into websites to translate their content;
- browser plug-ins that provides the quickest access to translation;
- integration in CAT tools.

[8] http://en.wikipedia.org/wiki/HTTP%2BHTML_Form_based_authentication

[9] http://en.wikipedia.org/wiki/Basic_access_authentication

### 3.3.1 CAT tools

Localisation and translation industry business and translation professionals are able to use their custom SMT systems from the specified collections of training data, public systems of the LetsMT! platform and accessing these solutions in their productivity environments.

It has been decided to integrate the LetsMT! platform with SDL Trados Studio 2009 and Kilgray memoQ. Integration is made directly into the translators' daily workflow. Users are able to use SMT systems as easily as translation memories (TM) are used. Particular translation result from the SMT system is displayed together with the suggestion from the TM.

#### 3.3.1.1 Functionality

Two main functions are implemented for LetsMT! plug-ins for CAT tools:

- Ability to choose which SMT system to use in the translation process. LetsMT! platform might provide more SMT systems for the same language direction and even domain, so each user should be able to choose SMT system that targets the most specific needs.
- Receive translation results from LetsMT! platform of segments to translate.

Detailed description of LetsMT! CAT tools – functionality, implementation, etc., is available in deliverable "Deliverable D6.1 Integration in CAT tools".

#### 3.3.1.2 Deployment

LetsMT! CAT plug-ins are client side component that has to be installed on client computer. User can obtain CAT plug-ins from the LetsMT! web page..

#### 3.3.1.3 Security

User has to be registered in the LetsMT! platform to use the LetsMT! platform via CAT tools. Authentication is necessary to identify which SMT systems are available for the user. TLS/SSL data encryption is used to protect the users' content that is sent over the network.

### 3.3.2 Browser Plug-ins

The LetsMT! browser plug-ins and the web page translation widget serve the purpose of demonstrating the basic capabilities of the LetsMT! platform to a wide spectrum of potential users. The browser plug-in itself is envisioned as a tool for quick and easy access to the basic translation services of the LetsMT! platform by using the user interfaces of client web browsers.

Currently, only selected version of Microsoft Internet Explorer and Mozilla Firefox web browsers are supported. After a quick and easy installation to the client browser, the user translates websites by using the provided plug-in interface within the browser.

The plug-ins are implemented as software interfaces to the LetsMT! web service (SOAP API). They mediate between the Public API of LetsMT! platform and the client browser interface and web presentation layer.

### 3.3.2.1 Functionality

LetsMT! browser plug-ins provide the user with machine translation on the web by using the LetsMT! platform facilities and the interface of selected versions of Internet Explorer and Mozilla Firefox web browsers. The basic functionality of this tool includes:

- The ability to translate custom client websites using the browser interface. Basically, the user accesses a website and uses the browser interface to forward the website to the LetsMT! web service (API) and receive the results within the interface, maintaining the structure of the website.
- A configuration tool within the browser plug-in configuration back-end. The back-end tool enables configuration and fine-tuning of the browser plug-in according to specific user requirements such as possibly the choice of translation system(s), etc.

Detailed description of LetsMT! browser plugins – functionality, implementation, etc., is available in deliverable “Deliverable D5.1 Widget and browser plug-ins for web page translation”.

### 3.3.2.2 Deployment

Users can obtain the browser plug-ins from the LetsMT! public website and install them into their browsers by using the plug-in installation interface provided by the browsers. User can obtain plug-ins from the LetsMT! web page.

### 3.3.2.3 Security

Users shall authenticate implicitly, by using the LetsMT! Public API authentication mechanism before using the SMT systems of LetsMT! platform. Authentication is necessary to identify which custom SMT systems are available for the user. If the user does not provide his user credentials within the plug-in, then only publicly available SMT systems will be available. TLS/SSL data encryption is used to protect the user content sent over the network.

## 3.3.3 Web Page Translation Widget

The LetsMT! browser plug-ins and the web page translation widget serve the purpose of demonstrating the basic capabilities of the LetsMT! platform to a wide spectrum of potential users. The web page translation widget is envisioned for source code integration within the client website in order to enable multilingual web features to these websites by seamless integration of the widget and the LetsMT! platform web service (API).

The widget is to be developed in Javascript and available as open source software for integration within client websites.

The widget is implemented as an interface to the Public API of the LetsMT! platform. It mediates between the platform API and the client web presentation layer.

### 3.3.3.1 Functionality

The web page translation widget provides prospective users, i.e. web developers and end users of their websites, with the possibility of enabling multilingual web features in their websites through the usage of LetsMT! platform services. There are two basic functionalities offered by the translation widget:

- Custom integration of the widget source code within the client website. This task is carried out by website developers wishing to enable multilingual features within their websites.
- Translation of websites by using the LetsMT! platform. Basically, the user accessing the website containing the widget has the option to translate the source language into a target language of his/her choice. The widget forwards the content of the current web page to the LetsMT! platform for translation and displays the translated results to the user, preserving the structure of the web page.

Detailed description of LetsMT! browser plugins – functionality, implementation, etc., is available in deliverable "Deliverable D5.1 Widget and browser plug-ins for web page translation".

### 3.3.3.2 Deployment

The widget is available to the general public from the LetsMT! public website. Website developers wishing to use the widget will integrate them within their websites according to their own design goals.

### 3.3.3.3 Security

Authentication is necessary to identify the website and which custom SMT systems are available for the website. If the owner of the website does not provide his credentials then only publicly available SMT systems for the widget will be used. TLS/SSL data encryption is used to protect user content sent over the network.

## *3.4 Application Logic Layer*

The Application Logic Layer is the interfacing component between high-level frontend services (LetsMT! website and the Public API) and the low-level backend services (Resource Repository and HPC Cluster). The Application Logic Layer manages the application logic of LetsMT! Platform, which includes the following functional responsibilities:

- User authentication and permission control
- Resource Repository management
- Training job management
- Translation service management
- User settings management
- HPC Cluster management
- Report management

The Application Logic Layer is implemented as a web service that provides an internal API of the LetsMT! platform. This API will not be accessible from outside, and will be interfaced by the services of the Interface Layer.

### 3.4.1 User authentication and permission control

The internal service of the Application Logic Layer allows only authenticated requests. User authentication is necessary to access non-public part of the LetsMT! website and to access Public API. User credentials from the services of the Interface Layer are passed to the service of the Application Logic Layer to authorize access to certain data and functions. User authentication is used to control access to resources in the Resource Repository and system functions.

### 3.4.2 Resource Repository management

This subcomponent will provide an interface for browsing and searching the training corpora and MT systems, editing metadata and data upload.

### 3.4.3 Training job management

This subcomponent is for preparing and submitting training jobs to the HPC Cluster, as well as monitoring their progress.

The training process is a complex task involving many steps that depend on the training configuration. The training is managed by EMS (Experiment Management System) that is included in Moses toolkit. The EMS supports training job execution on HPC Cluster.

The subcomponents role is to prepare metadata to generate the configuration for EMS and submit the training job to the HPC Cluster. The configuration is template based, requiring the user to specify some required fields (the training corpora) and several optional fields.

### 3.4.4 Translation service management

This subcomponent is for running the translation engines on HPC Cluster. It is monitoring the state of running SMT systems, launch additional instances of a SMT system and terminate them on demand. To mitigate high demand and response times, several instances of a SMT system can be launched. The subcomponent distributes the translation load between the instances.

### 3.4.5 Non-functional requirements

The Application Logic Layer will perform various control and management tasks and pass the information between its interfaced components. These tasks are not heavily CPU or I/O dependent.

As the Application Logic Layer is a potential point of failure for the whole platform, the non-functional requirements include fault tolerance.

In case the web service providing the Application Logic Layer is restarted, the implementation will reinitialize its state by querying the HPC Cluster and resource Repository.

## *3.5 Resource Repository*

The Resource Repository provides data processing facilities and general storage capacities for the LetsMT! platform. All training resources and trained SMT systems are stored in the Resource Repository. Users can browse the repository and search appropriate resources for

SMT training. The repository store metadata information about all resources and controls permission to protect or to share uploaded data.

Detailed description of Resource Repository – architecture, functionality, implementation, etc., is available in deliverable "D2.2 SMT resource repository and data processing facilities ready for integration".

## *3.6 High Performance Computing Cluster*

The High Performance Computing (HPC) Cluster provide computing environment to process variety of tasks. The tasks are data processing (including data extraction from variety of file formats supported by RR, data analysis and alignment calculation), and SMT system training and running. Tasks are initiated by the Application Logic Layer which initiate SMT system training and running tasks, and Resource Repository which initiate data processing of corpus files. HPC cluster is responsible of accepting, scheduling, managing distributed execution, and status providing of large numbers of tasks. The HPC Cluster is based on Oracle Grid Engine[10] (SGE).

Computing instances of the HPC Cluster are identical machines with Linux operation system and the same tool setup for desired task computing.

The HPC Cluster is well scalable. More computing instances can be dynamically allocated to satisfy growing demand of computing resources, and the resources can be released once computing demand reduces. Instances of the HPC Cluster are allocated and released on demand using Amazon Web Services. This help to reduce task processing speeds and optimize maintenance costs on less system load.

The HPC Cluster is built to process three types of tasks:

- SMT system training which consists of tenth of different training tasks,
- SMT system translation server hosting,
- Data upload and processing tasks.

SMT system training is performed with state-of-the-art Moses SMT toolkits and other necessary tools (Giza++, IRST LM, etc.). The SMT system training is launched and managed by EMS (Experiment Management System) which is included in Moses toolkit.

A single SMT system training consists of a number of independent tasks with its own mutual dependencies. All the fractions of the whole SMT system training task and its dependencies are managed by EMS. EMS submits training tasks to the HPC Cluster as batch-processes.

Detailed description of SMT system training facilities is available in deliverable "D3.3 SMT Training Facilities".

To start translation with trained SMT system on the LetsMT! platform, a SMT system translation server is launched for the SMT system on the HPC Cluster. The SMT system translation server use Moses decoder, tag translation components and supporting text processing tools to translate text. Multiple translation servers can be launched for a SMT system to increase translation capacity. The SMT translation services are started, stopped

[10] Oracle Grid Engine, previously known as Sun Grid Engine (SGE): http://gridengine.org/, http://en.wikipedia.org/wiki/Sun_Grid_Engine

and managed by the Application Logic Layer. The Application Logic Layer distributes translation requests over the SMT system translation server farm using XML-RPC [11]protocol.

## 3.7 SQL Database

LetsMT! SQL database is used for storing information about LetsMT! user, user groups, user permissions, user access control information and user activity log. MySQL database engine is currently used as SQL database for LetsMT! platform.

LetsMT! system's resources such as uploaded corpora and trained SMT systems are stored in the Resource Repository. Metadata associated with these resources are also stored in metadata storage of the Resource Repository.

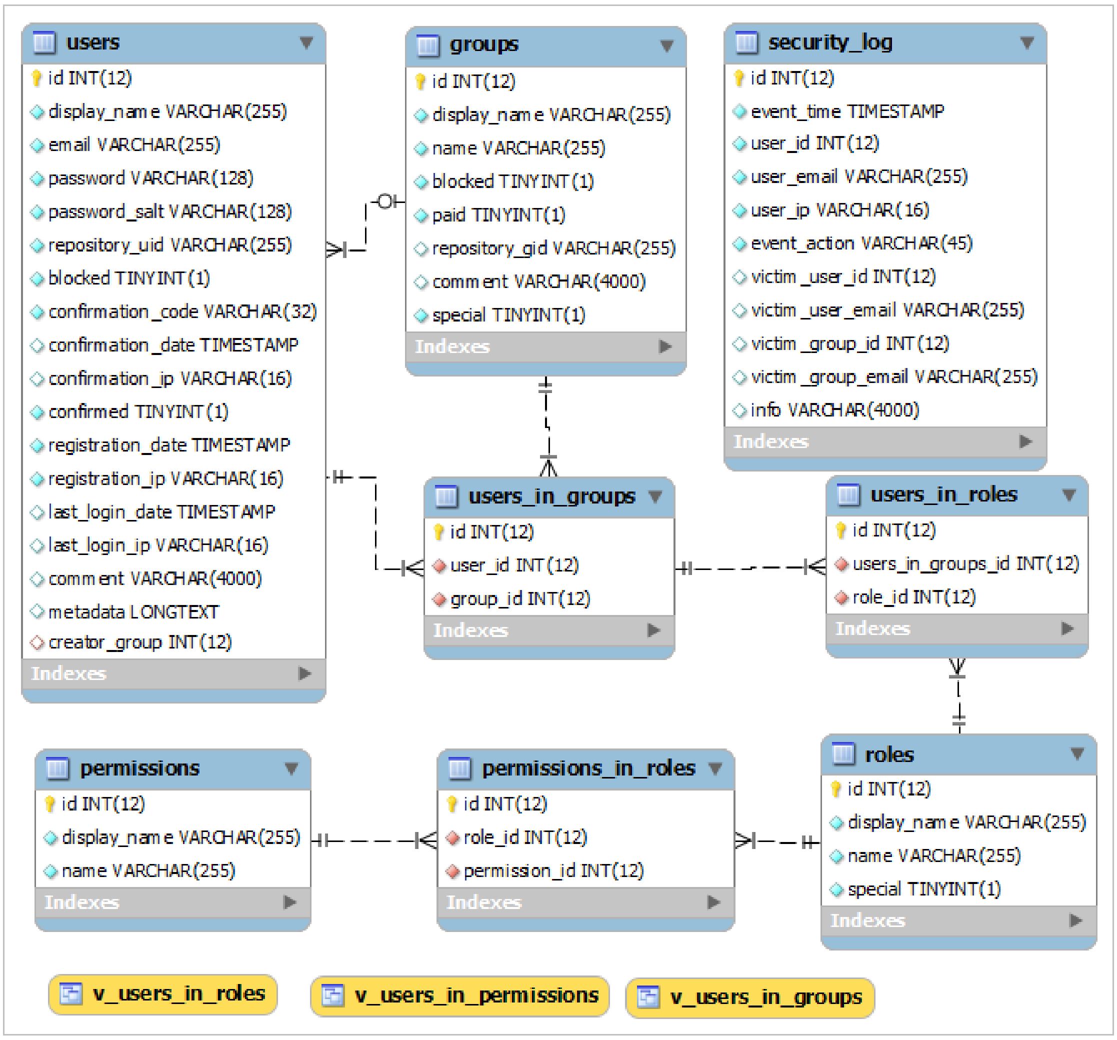


**Figure 37. ER model of LetsMT! SQL database**

For storing LetsMT! user and user management data following database tables are used:

- „users" – stores information about LetsMT! users;
- „groups" – stores information about groups of LetsMT! users;
- „permissions" – stores information about user permissions;

[11] XML Remote Procedure Call (XML-RPC): http://xmlrpc.scripting.com/

- „roles” – stores information about roles;
- „users_in_groups” – joins tables “users” and “groups”;
- „users_in_roles” – joins tables “users” and “roles”;
- „permissions_in_roles” – joins tables “permissions” and “roles”;
- “security_log” – information about user activities.

**Table „users”**

| Field name | Data type | Mandatory | Description |
|---|---|---|---|
| id | INT(12) | Yes | User ID |
| display_name | VARCHAR(256) | Yes | Display name |
| email | VARCHAR(256) | Yes | E-mail |
| password | VARCHAR(128) | Yes | Password + salt hash using SHA-2. |
| password_salt | VARCHAR(128) | Yes | Password salt |
| respository_uid | VARCHAR(256) | Yes | User UID in Resource Repository |
| blocked | TINYINT(1) | Yes | User blocked ? |
| confirmation_code | VARCHAR(32) | Yes | E-mail confirmation code |
| confirmation_date | TIMESTAMP | | E-mail conformation date |
| confirmation_ip | VARCHAR(16) | | Confirmation IP address |
| confirmed | TINYINT(1) | Yes | Is user e-mail confirmed? |
| registration_date | TIMESTAMP | Yes | Data of registration |
| registration_ip | VARCHAR(16) | Yes | Registration IP |
| last_login_date | TIMESTAMP | | Last login date |
| last_login_ip | VARCHAR(16) | | Last login IP |
| comment | VARCHAR(4000) | | Comments |
| metadata | LONGTEXT | | User metadata in XML |
| creator_group | INT(12) | | Initial user group |

**Table „groups”**

| Field name | Data type | Mandatory | Description |
|---|---|---|---|
| id | INT(12) | Yes | Group  ID |
| display_name | VARCHAR(256) | Yes | Group display name |
| name | VARCHAR(256) | Yes | Group name |
| blocked | TINYINT(1) | Yes | Is group blocked? |
| repository_gid | VARCHAR(256) | | Resource Repository group ID |

| Field name | Data type | Mandatory | Description |
|---|---|---|---|
| owner_id | INT(12) | Yes | Group ID in Resource Repository |
| comment | VARCHAR(4000) | | Comments |

**Table „users_in_groups"**

| Field name | Data type | Mandatory | Description |
|---|---|---|---|
| id | INT(12) | Yes | ID |
| user_id | VARCHAR(256) | Yes | User ID |
| group_id | INT(12) | Yes | Group ID |

**Table „permissions"**

| Fiel name | Data type | Mandatory | Description |
|---|---|---|---|
| id | INT(12) | Yes | ID |
| display_name | VARCHAR(256) | Yes | Permission display name |
| name | VARCHAR(256) | Yes | Permission name |

**Table „roles ”**

| Field name | Data type | Mandatory | Description |
|---|---|---|---|
| id | INT(12) | Yes | Role ID |
| display_name | VARCHAR(256) | Yes | Role display name |
| name | VARCHAR(256) | Yes | Role name |

**Table „permissions_in_roles"**

| Field name | Data type | Mandatory | Description |
|---|---|---|---|
| id | INT(12) | Yes | ID |
| role_id | INT(12) | Yes | Role ID |
| permission_id | INT(12) | Yes | Permission ID |

**Table „users_in_roles"**

| Field name | Data type | Mandatory | Description |
|---|---|---|---|
| id | INT(12) | Yes | ID |
| users_in_groups_id | INT(12) | Yes | User group ID |

| Field name | Data type | Mandatory | Description |
|---|---|---|---|
| role_id | INT(12) | Yes | Role ID |

**Tabula „sercurity_log"**

| Field name | Data type | Mandatory | Description |
|---|---|---|---|
| id | INT(12) | Yes | Event ID |
| event_time | TIMESTAMP | Yes | Timestamp |
| user_id | INT(12) | Yes | User's ID |
| user_email | VARCHAR(256) | Yes | User's e-mail |
| user_ip | VARCHAR(16) | Yes | User's IP |
| event_action | VARCHAR(45) | Yes | Event description |
| victim_user_id | INT(12) | | ID of involved user |
| victim_user_email | VARCHAR(256) | | E-mail of involved user |
| victim_group_id | INT(12) | | ID of involved user group |
| victim_group_email | VARCHAR(256) | | E-mail of involved user group |
| info | VARCHAR(4000) | | Additional information |

### 3.7.1 Non-functional requirements

The amount of metadata stored in the SQL database will not be big; therefore, scalability is not an issue.

## *3.8 Software Components*

LetsMT! system is built on top of many existing software packages providing the  necessary functionality. This section outlines what software is used, what adaptations to the existing software are planned and implemented and what software has been developed specially for LetsMT! needs.

### 3.8.1 Statistical Machine Translation

The core functionality of LetsMT! system is SMT training and running of SMT systems. In recent years SMT has provided a major breakthrough in development providing a cost effective and fast way to build SMT systems. This development was particularly facilitated by the open-source corpus alignment tool GIZA++[12], the MT training and decoding tool Moses[13] and other tools. The LetsMT! system is mainly based on these existing tools. The Moses toolkit is the core of the LetsMT! platform and it has been enriched with new features.

[12] GIZA++: http://fjoch.com/GIZA++.html

[13] Moses: http://www.statmt.org/moses/

For detailed information about Moses toolkit adaptation for LetsMT! platform needs see deliverable “D3.1 Adapted Moses toolkit”.

#### 3.8.1.1 Alignment

The first step in building SMT translation models from parallel corpora is automatic word alignment. This part of the process is especially complicated and requires a great deal of computational power especially for large-scale corpora. Standard word alignment for SMT are the IBM models and the HMM alignment model implemented in the freely available tool GIZA++. It can be used as a black-box tool in connection with the Moses toolkit which supports all the necessary steps to build a phrase-based SMT system from a given sentence aligned parallel corpus. The word alignment is carried out in an unsupervised way using EM re-estimation procedures and a cascaded combination of alignment models. Various settings can be adjusted in the alignment procedure and phrase table extraction. The word alignment is time consuming and requires large amounts of internal memory for extensive data sets. Fortunately, there are extensions and alternative tools available with improved efficiency. MGIZA++ is a multi-threaded version of GIZA++[14] and it can run several word alignment processes in parallel on a multi-core machines. Furthermore, the same author provides a cluster-based version of GIZA++ that can be used to distribute word alignment over various machines. An alternative approach that can also run a parallel alignment procedure is implemented in the MTTK toolkit[15].

#### 3.8.1.2 Training

A significant breakthrough in SMT was achieved by the EuroMatrix project[16]. Among the project objectives were translation systems for all pairs of EU languages and the provision of an open source MT technology including research tools, software and data. The project resulted in the improved open source SMT toolkit Moses developed by the University of Edinburgh. The Moses SMT toolkit is a complete translation system distributed under LGPL (Lesser General Public License). Moses includes components needed to pre-process data, train language models and translation models. Moses is widely used in the research community and has also reached the commercial sector. While the use of the software is not closely monitored (there is no need to sign any license agreement), Moses is known to be in commercial use by companies such as Systran, Asia Online, Autodesk, Matrixware, Translated.net. The EuroMatrix project has demonstrated how open source tools and publicly available data can be used to generate SMT systems for all language pairs of EU official languages.

The LetsMT! project extends the use of these existing state-of-the-art SMT tools enabling users to build custom tailored SMT systems through simple web based interface. LetsMT! uses Moses as a language independent SMT solution and integrate it as a cloud-based service into the LetsMT! online platform.

Important advancement of the LetsMT! is an adaptation of the Moses toolkit to fit into the rapid training, updating and interactive access environment of the LetsMT! platform. The SMT training pipeline implemented in Moses currently involves a number of steps that each

[14] MGIZA++: http://www.cs.cmu.edu/~qing/ and http://code.google.com/p/giza-pp/

[15] http://mi.eng.cam.ac.uk/~wjb31/distrib/mttkv1/

[16] http://www.euromatrix.net/

require a separate program to run. In the framework of LetsMT! this process is streamlined and made automatically configurable given a set of user-specified variables (training corpora, language model data, dictionaries, tuning sets). The SMT training process is based on improved Moses Experiment Management system[17].

### 3.8.1.3 SMT Decoder

LetsMT! system is based on Moses SMT decoder. We have improved Moses decoder to better suite LetsMT! requirements. First the Moses decoder is adjusted to work with (i) incrementally built and suffix array based translation models (more in chapter 3.8.1.4 Translation Models) and (ii) stream-based randomized language models (more in chapter 3.8.1.5 Language Models).

The prior implementation of the Moses decoder had deeply integrated language and translation model functionality. Models were loaded into the Moses process memory. It means that we have to run the whole decoding system in one process on one machine. The machine must have sufficient amount of RAM and disk space, and the Moses decoder has a long launch time as it must load models in launch time. The distributed language models have been introduced to solve this issue.

### 3.8.1.4 Translation Models

An additional important improvement of Moses that will be implemented as a part of LetsMT! is the incremental training of translation  models. These advancements will be based on an online version of the EM algorithm (Levenberg et al., 2010) for the word aligning and suffix arrays (Chris Callison-Burch et al., 2005) for translation models.

### 3.8.1.5 Language Models

Currently the Moses decoder works with the following language models:

- SRI language modeling toolkit[18]
- IRST language modeling toolkit[19]
- RandLM language modeling toolkit[20]
- KenLM language model interface[21]

IRSTLM toolkit compared to SRILM handles LM formats which permit to reduce both storage and decoding memory requirements, and to save time in LM loading. RandLM allows building of the largest LMs possible (for example, a 5-gram trained on several billions of words, such as the whole of the Gigaword Corpus). KenLM is a library that can train its own structure from estimated language model, i.e., using the models estimated by previously mentioned toolkits. KenLM is fast and memory-efficient, and it can outperform previously mentioned language modeling tools in both of the figures.

Since we are expecting users to add new amounts of additional monolingual training data in frequent intervals, we will need to retrain language models frequently. All the mentioned

[17] Experiment Management System: http://www.statmt.org/moses/?n=FactoredTraining.EMS
[18] SRILM: http://www.speech.sri.com/projects/srilm/
[19] IRSTLM: http://sourceforge.net/projects/irstlm/
[20] RandLM: http://sourceforge.net/projects/randlm/
[21] KenLM: http://kheafield.com/code/kenlm/

LM toolkits currently supported in Moses are batch-based and retraining is computationally demanding. Stream-based Randomized Language Models (Levenberg et al., 2009) will be integrated in Moses toolkit to streamline building of frequently increasing language models.

In LetsMT! we will use language model software depending on size and nature of data used to train the model.

### 3.8.2 Processing of training data

Detailed description of processing of training data – functionality, implementation, etc., is available in deliverables "D2.1 Specification of data formats allowed" and "D2.2 SMT resource repository and data processing facilities ready for integration".

### 3.8.3 Operating systems, frameworks and servers

The main LetsMT! functions are sharing and processing of training data, MT training and MT running. All these functions are implemented by integrating and adapting existing tools which are designed for Linux platform. It means that all modules in data storage layer and HPC cluster will be running in Linux environment.

HPC Cluster is used to execute many different data processing and SMT training and running tasks. Modules from the application logic layer are creating jobs and send them to HPC cluster to execute. LetsMT! HPC cluster is based on Sun Grid Engine[22] (SGE) running on Linux platform.

LetsMT! database is developed using SQL server. It is important to make database portable and scalable and easy deployable in different hardware infrastructures. For example we have local servers for test and development environments and cloud based hardware solution for public use. For current LetsMT! platform deployment MySQL is used as database server.

LetsMT! web page, public API and many parts of application logic is developed using Microsoft .NET 4.0 framework as it provides powerful and easy scalable tools for web application and web service development and Tilde has long experience and skills working with this technology.

### 3.8.4 Software developed in scope of LetsMT! project

There are areas in LetsMT! system where we cannot use existing software and we have implemented new features in scope of the project.

First of all we have implemented all functionality of interface layer and application logic layer as this is very project specific functionality. The LetsMT! application logic integrates together all tools which we are used for data processing, SMT training and SMT running.

Many data pre-processing tasks are based on existing tools such as aligners, file format converters, but we need to develop tools to process some file formats (e.g. TMX and XLIFF) to validate results and to integrate all pre-processing tools together in one pre-processing chain or workflow.

[22] SGE: http://gridengine.sunsource.net, http://en.wikipedia.org/wiki/Sun_Grid_Engine

# 4 Infrastructure Design

The hardware infrastructure of LetsMT! platform is heterogeneous:

1) The majority of services are running on Linux platform (Giza++, Moses, data processing tasks);

2) Other services runs on Windows platform.

System hardware architecture is designed to provide high scalability and availability.

The Consortium, instead of buying servers, intends to lease capacity. This is economically efficient and provides flexibility in adding new resources exactly when necessary.

It is planned to deploy the LetsMT! platform completely within Amazon Web Services (AWS)[23] as this is the tested 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! system. The LetsMT! platform also can be deployed on local server infrastructure with obvious deviation from AWS deployment.

Detailed description of LetsMT! hardware infrastructure and deployment is available in deliverable “D1.3 Hardware infrastructure”.

[23] Amazon Web Services (AWS): http://aws.amazon.com/

# 5 References

Abby Levenberg, Chris Callison-Burch and Miles Osborne. 2010. Stream-based Translation Models for Statistical Machine Translation. NAACL, Los Angeles, USA.

Abby Levenberg and Miles Osborne. 2009. Stream-based Randomised Language Models for SMT. EMNLP, Singapore.

Chris Callison-Burch, Colin Bannard, and Josh Schroeder. 2005. Scaling phrase-based statistical machine translation to larger corpora and longer phrases. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), pages 255–262, Ann Arbor, Michigan, June. Association for Computational Linguistics