LetsMT!

**Platform for Online Sharing of**

**Training Data and Building User Tailored MT**

Grant Agreement number 250456

letsmt.eu

project.letsmt.eu

# Final Publishable Summary

**31.10.2012**

# Table of Content

1 EXECUTIVE SUMMARY ..........4
2 PROJECT CONTEXT AND OBJECTIVES ..........6
2.1 Project context ..........6
2.2 Project objectives ..........7
3 WORK CARRIED OUT AND MAIN RESULTS ..........8
3.1 LetsMT! key features ..........8
3.2 SMT Resource Repository and data processing facilities ..........9
3.3 SMT training facilities and SMT web service ..........11
3.4 Hardware infrastructure and support services ..........13
3.5 Collecting the training data ..........13
3.6 LetsMT! widgets and browser plugins ..........16
3.7 LetsMT! Mobile translation application ..........18
3.8 Machine translation services for business and financial news ..........19
3.9 Facilities and evaluation of MT usage in localisation ..........22
4 MAIN DISSEMINATION ACTIVITIES AND EXPLOITATION OF RESULTS ..25
5 THE PROJECT’S PARTICIPANTS PUBLIC WEBSITE AND CONTACT DETAILS ..........29

# 1 Executive summary

The proliferation of information enabled by the meteoric advance of the Internet, and the growing tide of multilingual data has resulted in the need for a paradigm change in the way we process information.

In the European Union there are 23 official languages and more than 80 languages in daily use. Language diversity, one of the cornerstones of EU unity and a source of identity and national pride is also a barrier to communication and commerce in the single market. Recent advances in the shift from rules based to statistical methods have produced cardinal changes in the way technology can serve communication especially for large languages, however, it has not reduced the digital divide for smaller languages. Our work on the cloud-based SMT platform LetsMT! started with the aim of creating a revolutionary solution that simplifies MT for smaller languages and specific domains. This concept started our work on the cloud-based platform letsmt.com.

The initial project specifications defined the following requirements of a good service platform for generating SMT engines:

- easy to use – no complicated code or a multitude of manual steps and processes; process and store your data in various widely used formats;
- align the parallel texts that will be used to train systems;
- significant amounts of language and domain specific resources available to add to a user's own data
- training a SMT engine should take no more than few mouse clicks
- access and use the results in a familiar tool whenever and wherever translation is required

The LetsMT! platform delivers and has exceeded expectations on the original requirements. The latest SMT research and technological solutions have been transformed into a commercially viable service, putting custom MT in reach of otherwise underserved SME's. The cloud-based LetsMT! platform web interface (letsmt.com) is an intuitive, self-explanatory environment that beta testers have evaluated as very user-friendly. Just a few clicks and through a step-by-step process and even an MT neophyte can upload data and train and use custom Machine Translation.

The core technology of the platform fully automates the underlying open-source Moses tookit and all the necessary supporting functions for generating SMT engines.

Data is the key for quality SMT results. Smaller languages, narrow domains suffer from a lack of sufficient data. The project has endeavored to remedy this inequality by seeding LetsMT! with large quantities of sharable data. At the time of writing, the platform repository contains resources in excess of project requirements:

- 104 languages;
- 1.5 billion sentence pairs;
- over 2 billion monolingual sentences;
- 39 public MT systems.

With these substantial initial resources, there is a solid basis for training baseline systems for a large number of language pairs. There are a great many resources available in various private collections and we anticipate that sharing resources will become a common activity as the benefits of sharing and reusing data can be illustrated. This positive example is necessary to motivate users to upload more data to increase the coverage of the system in terms of language pairs and domains.

More than 100 SMT systems in different subject domains and different language pairs have already been trained using the LetsMT! platform.

Since commercial implementation is the key outcome of the project, a number of industry use scenarios were elaborated in the original requirements have been evaluated during the project. SMT is vital for increasing productivity in the localization/translation industry. Another current need is the speedy and economic translation of fast-turnaround but short shelf-life content such as financial news.

For each scenario specific features were developed to meet their specific demands and criteria. For the localisation industry usage scenario tools were developed to integrate the MT results directly into the translation workflow. The widely used SDL Trados Studio and Kilgray MemoQ CAT tools were selected to develop plug-ins for. During the project 6 MT systems were trained in the localization domain. According to the automatic evaluation measures all the LetsMT! systems achieved significantly better scores than the translation output by *Google Translate*.

In addition, several evaluations were made measuring impact of SMT in the localisation workflow using assistance from LetsMT! machine translation. The translators were given translation suggestions from the SMT engine in addition to the usual translation memory suggestion. The evaluations of using MT results in translation indicated a strong increase in translators' productivity by 25-33% while keeping the error rate at the acceptable quality level.

The major goals of the project related to the news translation scenario have also been achieved. For this use scenario 6 LetsMT! systems – all trained for text within specific domains – outperformed *Google Translate* for the chosen evaluation sets. Widget and browser plugin software components have been developed and can be used to translate websites with trained SMT systems running on the LetsMT! platform.

The project use scenarios are backed-up by enthusiastic reviews from entities interested in integrating SMT into various commercial scenarios. Currently 80 potential LetsMT! platform users have expressed interest to use LetsMT! services and have applied to try the LetsMT! platform. User feedback has been universally positive, and instrumental in suggesting in-project adjustments and improvements to the platform.

Additional features such as tag translator and use of LetsMT! on mobile devices that were not foreseen in the original specification, have also been developed.

Significant efforts from consortium partners were dedicated to the dissemination of the LetsMT! project results. Project partners presented LetsMT! at more than 30 localisation, industry and research events contributing to the high profile and interest in the LetsMT! project. Scientific results of the project are published in 14 peer-reviewed research papers. LetsMT! success is also reported in the NewScientist magazine and localization industry magazine MultiLingual.

The initial requirements of the project have been more than fully implemented and fulfilled. The LetsMT! platform has made quality custom SMT accessible for under-resourced languages and narrow domains. The technology contributes to European business competitiveness, increases the pool of open data resources for language and is another constructive step toward closing the digital information divide.

# 2 Project context and objectives

## 2.1 Project context

The web and explosion of its multilingual content is driving the need for solutions to access information in a multitude of languages. We have been witnesses to the evolution from human to machine assisted, to human assisted translation. There is a number of factors that have made this a propitious time for rapid advances in language technologies. Key among them are the availability of scalable computing power which can support the complex algorithms that drive statistical machine translation (SMT), the explosion of texts and other linguistic resources necessary to feed the statistical systems, and powerful open source tools like Moses[1]. In recent years statistical machine translation has become a major developmental breakthrough by providing a cost efficient and fast way to build and use machine translation systems.

However, even recent advances fall short of fulfilling expectations regarding MT systems. The quality of a statistical MT system largely depends on the size of training data. Obviously the majority of available parallel data has been generated by major languages. MT works by statistically comparing the parallel corpora of two languages and calculating the probabilities that are used to generate the most likely translation. As a result SMT systems for the most widely used languages are of much better quality compared to systems for under-resourced languages.

This quality gap is further broadened by the complex linguistic structure of many smaller languages. Languages like Latvian, Lithuanian and Estonian, to name just a few, have a complex morphological structure and free word order. To learn this additional complexity from corpus data by statistical methods, proportionally much larger volumes of training data are needed than for languages with a simpler linguistic structure.

Another drawback preventing wider implementation of MT is its general nature. Although free web translation systems 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. For example, Google Translator currently provides MT for more than 50 languages. However, for smaller languages such as Latvian or Estonian, translation quality is quite poor, particularly for domain specific texts, making it an unsuitable productivity tool for use in the translation industry.

While large languages have the benefit of large markets that successfully amortise investments in proprietary systems, smaller languages also suffer from smaller consumer markets and lower overall translation volumes. Many producers of goods and services supply content mostly in larger languages because the cost of human translation into smaller languages is prohibitively high and the quality of existing MT solutions is insufficient.

In the localisation and translation industry, huge pools of parallel texts in a variety of industry formats have been accumulated, but the application of this data has not yet been fully utilised in modern MT. At the same time, this industry is experiencing unrelenting pressure for efficiency and performance, clients expect more to be translated in nearly real time at lower prices.

[1] http://www.statmt.org/moses/

Presently, integration of MT in translation and localisation services is in its early stages, and is mostly in the realm of large agencies working with the large languages. The cost of developing specialised MT solutions is prohibitive for most players in the localisation and translation industry, while the quality and confidentiality afforded by the free generic MT offerings are not sufficient to reap substantial efficiency gains in the professional localisation industry setting.[2]

These considerations have led us to evaluate possibilities provided by the recent advances in machine translation and seek for innovation to find cost-efficient and user friendly methods to make custom machine translation affordable.

## 2.2 Project objectives

The objective of the LetsMT! project was to design and to build a platform that makes custom MT affordable for everyone and encourages collaboration and reuse of resources contributed by users for generating MT engines. The goals of the project were to increase the availability of parallel language resources and to make MT services of good and acceptable quality for under resourced languages accessible to those whose current MT systems perform poorly due to the limited availability of training data.

To fully exploit the huge potential of existing open SMT technologies LetsMT! aimed to create an innovative online collaborative platform for data sharing and MT building. This platform should support upload of public as well as proprietary MT training data and building of multiple MT systems, public or proprietary, by combining and prioritising this data. The project should extend the use of existing state-of-the-art SMT methods by applying them to data supplied by users to increase quality, scope and language coverage of machine translation.

Two particular application scenarios have been the focus of LetsMT! –free online translation of business and financial news and implementation in the localisation and translation industry. At the same time, LetsMT! should be of interest for a variety of users: web users in general, speakers of less-covered languages, academia, multinational enterprises, government entities, etc.

For the localisation and translation industry, LetsMT! should provide facilities for training of SMT systems on their data and generating custom SMT solutions to be used by localisation service providers as well as enterprises and organisations with multilingual translation needs. Integration of SMT solutions in professional productivity environments should be ensured.

For readers of business and financial news, LetsMT! should provide free and instant MT services with emphasis on less-covered languages. Their quality should be ensured by the use for training of a large pool of domain-specific resources and subsequent evaluation cycles.

[2] Andrejs Vasiļjevs, Indra Sāmīte (2012), *Machine translation for less-resourced languages*, MultiLingual Magazine, January/February 2012

# 3 Work carried out and main results

The project has accomplished its goals by providing elaborated and evaluated functionality now available for usage in several application scenarios. The Platform has a simple-to-use interface which intuitively guides users through the process of creating custom machine translation engines.

The results of the project have exceeded original expectations for the project implementation with a number of additional features such as incremental training which facilitates adding batches of fresh resources without re-training an entire engine. Advanced processing of text formatting, dealing with multiple additional file formats and other specific additional features that were not foreseen in the original specification have also been developed.

LetsMT! delivers the following **core functionality**:

- Website for the upload of parallel corpora and the building -of MT engines;
- Website for the translation of shorter texts;
- Widget to translate website content;
- Browser plug-ins that provide the quickest access to translation;
- Integration in CAT tools to enhance productivity;
- Sophisticated customization features for advanced users;

The LetsMT! Platform has been developed, integrated, tested, and is being implemented in a commercial translation scenario. It is available 24/7 through the Amazon Web Services computing cloud platform. The project partners have gathered and made available to users of the LetsMT! system an extensive collection - 1.5 billion parallel sentences, 102 language pairs and 39 pre-trained systems for public use.

## *3.1 LetsMT! key features*

LetsMT! provides a simple interface to guide the user though the process of creating their custom MT engine. To attack the most difficult problem for small languages and specialized domains – the lack of training resources – the core concept of the Platform is to share resources. The Platform provides a repository of data collected by project partners and by the users of letsmt.eu. The users uploads their data, supplements it with the data already provided on the Platform and generates their custom MT engine.

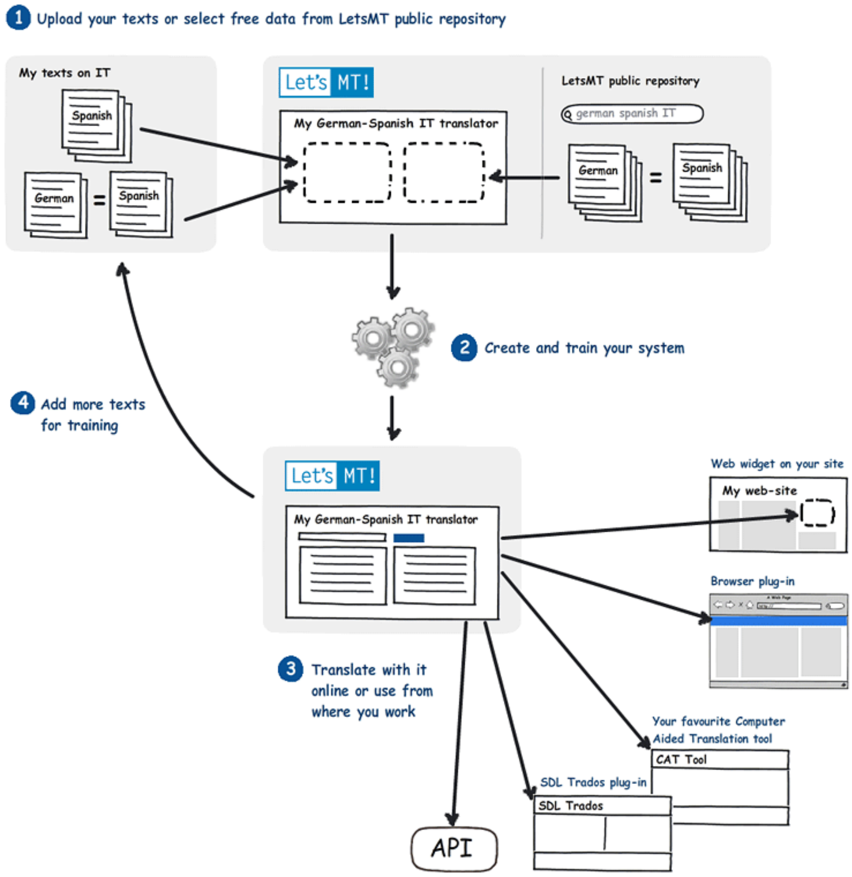

1 Upload your texts or select free data from LetsMT public repository
My texts on IT
Spanish
German
=
Spanish
Let's MT!
My German-Spanish IT translator
LetsMT public repository
german spanish IT
German
=
Spanish
2 Create and train your system
4 Add more texts for training
Let's MT!
My German-Spanish IT translator
Web widget on your site
My web-site
Browser plug-in
3 Translate with it online or use from where you work
Your favourite Computer Aided Translation tool
CAT Tool
SDL Trados plug-in
SDL Trados
API

**Figure 1** Conceptual workflow of the LetsMT!

The LetsMT! Platform has fully succeeded in providing all of the initially planned key **features**:

- An intuitive web interface overlaying the open-source Moses toolkit;
- Supports collaboration by providing easy upload of parallel texts for users that will contribute their content;
- Universally accessible directory of all public corpora gathered by LetsMT!;
- Automated training of SMT systems from specified collections of training data;
- Enables customisation of MT engines from both public and/or proprietary data;
- Evaluation of generated MT with widely used automated scores (BLEU, NIST, etc.);
- Cloud based 24/7 accessibility.

## 3.2 SMT Resource Repository and data processing facilities

One of the essential modules of the LetsMT! Platform is the resource repository (RR). Success and quality of statistical machine translation is strongly connected to the availability of appropriate training data in terms of large parallel, translated documents and even larger monolingual texts in the target language. Training satisfactory quality, that is useful in real-life scenarios machine translation engines requires copious amounts of domain-specific and usage-specific data to achieve high quality results. The RR is not only a place where users can store and process their own data, but it is also a repository for large collections of public corpora. The quality of the output of SMT engines is largely correlated with the volume of training data used. In many cases commercial users do not have sufficient resources of their own to create viable SMT engines. The public corpora are available to users to use together with their own data in order to improve the quality of the resulting engines. In addition the RR includes various tools to automatically perform all the necessary conversion and data processing tasks in order to create appropriate data sets from possibly noisy user uploads.

The RR is the main storage facility for the LetsMT! Platform. Users can upload their resources in various widely used data formats such - SRT (SubRip text format), XLZ, RTF, OpenDocument, basic HTML and standalone XML.

The LetsMT! RR was designed with the following requirements in mind:

- The repository needs to be scalable in order to support a growing multi-user environment.
- The upload and conversion mechanisms need to be flexible yet simple and mostly automatic.
- Uploaded data needs to be stored in a save state securing user and data integrity.
- Data sets need to be cleaned and prepared for immediate training of translation engines.

For readers interested in technical details, an illustration of the general architecture of the repository software is provided in Figure 2.

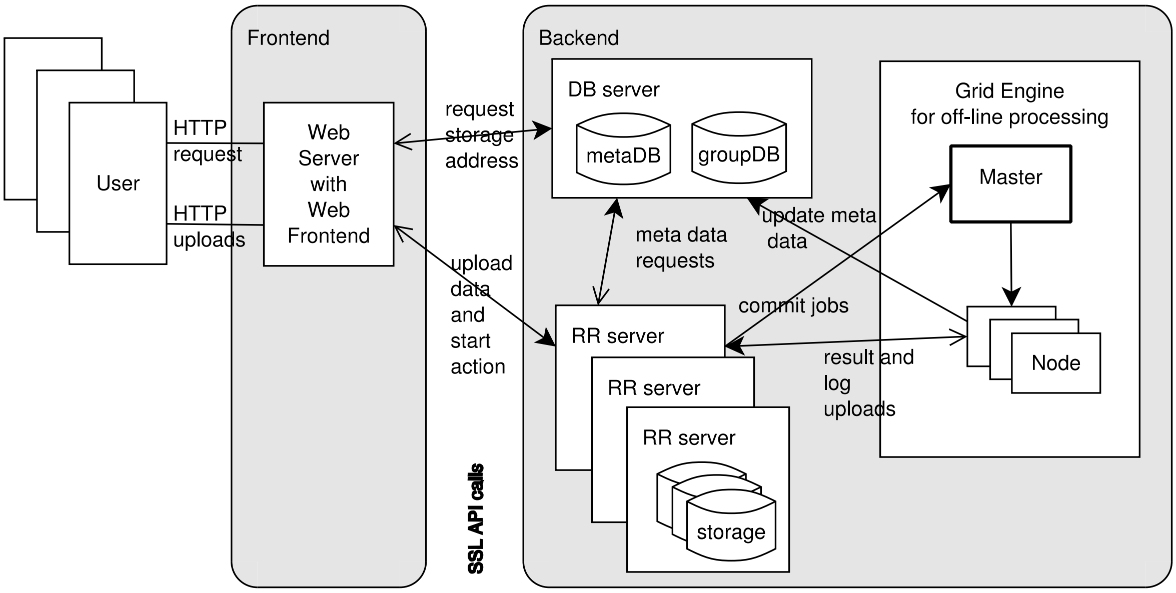


**Figure 2** General architecture of the LetsMT! Resource Repository.

On the client's side (web browser), the data upload mechanism is implemented using Flash technology. The first step is to upload - data from the user's machine to a temporary folder on the webserver. During the next steps, data is a-synchronically uploaded to the Resource Repository

One important aspect of the RR and the entire LetsMT! Platform is the support of user uploaded content that can be used to train user-specific translation engines. It is, therefore, necessary to take care of user provided data in various formats and make it suitable for training statistical MT models. In the translation and data creation industries many different document formats are in use. In order to make the RR user-friendly, it is possible to import documents and files of various data formats and for RR to implements automatic data validation and conversion tools. The uploaded documents are processed automatically in the RR and the result is training data in a format which is appropriate for use in training SMT engines. The main steps are summarized by the following tasks:

- Data validation;
- Data conversion;
- Data alignment.

The beauty of the system is in the simplicity of its modular and extendable architecture.

LetsMT! Platform automatically detects character encodings, identifies and verifies the language used and detects sentence boundaries . Another feature is the automatic alignment of the data. In case translated documents are detected (using name and size rules), it also triggers an automatic sentence alignment process. Statistical methods are used to extract parallel sentences from translated documents, which are then the basis for the translation models used in SMT.

## *3.3 SMT training facilities and SMT web service*

The SMT training facility and web service is built on top of the Moses machine translation toolkit. Significant improvements to the Moses toolkit were implemented within the framework of the LetsMT! project.

A new feature to the Moses toolkit is the distributed language model. Valuable feature of distributes the training of especially large language models across several processors at once – thus reducing the time that it takes to train models. The gains in time and hence resource efficiency are notable. The distributed language model has the fallowing two advantages: A decoder wishing to use it only needs to connect to it, thereby eliminating the slow start-up time associated with all other language models. Second, the distributed language model can be shared across other decoders, thereby further reducing the amount of memory needed. Google has a similar cluster-based language model.

Another new feature is incremental training. This is a particularly complex and involved feature that is a step closer to the holy grail of SMT: continuous improvement using freshly translated content. Incremental training allows batches of fresh data to be added to an existing SMT engine on the Platform. Since the effects of adding additional data can only be felt if the proportion of fresh to existing resources is significant incremental training is an economic way to improve the quality incrementally, without taking the time and computing resources to retrain the whole engine. The benefits are that it more closely resembles the working of traditional translation memories which are updated continuously, and it reduces the resources and time required to add incremental resources from translation work in progress.

**Speed and efficiency improvements.** The Moses decoder was continuously improved in terms of speed and space it needs for translation.

**Bug fixes.** Like all complex pieces of software, Moses contained bugs. Throughout the lifetime of the project we supported Moses, fixed bugs and feeding the updates to the distribution. LetsMT! benefited from these updates.

Building SMT systems is a complex affair, involving numerous steps (for example: pre-processing of data, building word alignments, creating phrase tables, tuning, etc)

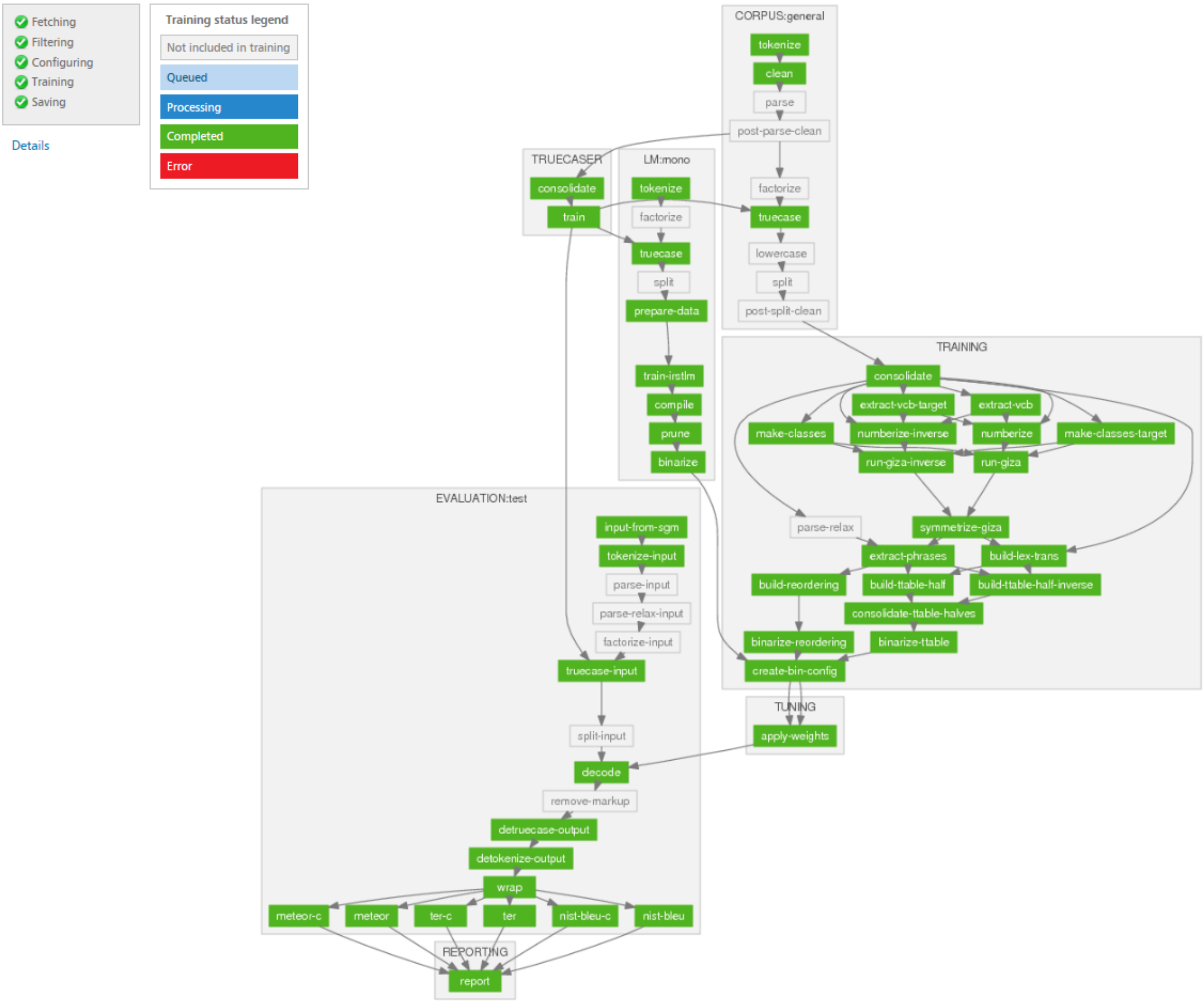


**Figure 3** LetsMT! graphically represents the progress of an MT engine training process.

Figure 3 shows a typical example of an MT system in the process of being built. Boxes represent tasks and lines showing the flow of control. Without the process automation provided by the LetsMT! Platform each one of these steps would require a manual intervention to run.

All the SMT systems available through the LetsMT! Platform is accessible through the SMT web service. The MT web page, the widget for webpage translation, and interfaces for integration in CAT tools will all use this web service to access a particular SMT system. The LetsMT! Platform can run many translation engines simultaneously, and translation engines can be run on several servers to ensure necessary system performance.

The LetsMT! Public Translation API supports the key features like translation, sentence breaking, language detection, client identification and authentication (basic authentication over SSL tunnel).

Training time is an important factor for the LetsMT! Platform. The less time users have to wait to get results for their systems, the faster they can tune and develop them and make better use of the facilities LetsMT! provides. By spreading the load of system training across multiple machines, more efficient use will be made of the hardware resources the project has invested in.

LetsMT! automatically evaluates trained MT systems by combining the traditional BLEU/NIST evaluation with TER evaluation scores as these can be calculated language independently. Scores are based on a basic version of METEOR because it does not require language dependent resources.

## 3.4 Hardware infrastructure and support services

To ensure flexibility and scalability of hardware resources, the LetsMT! Platform is deployed on the cloud infrastructure. It allows easy scaling up and down based on customer usage. Fallowing a cost-benefit analysis Amazon Web Services were chosen as the most suitable solution.

The system hardware architecture is designed to be highly scalable. LetsMT! Platform runs on multiple servers and can accommodate both continuous and on-demand availability.

To ensure efficient and reliable operations of the service administration and system maintenance features have been developed:

- Web based administration console;
- Elaborated monitoring of hardware utilisation;
- Selection of different types of instances to optimize the usage of the cloud infrastructure;
- e-mail notifications informing users about the status of SMT system training;
- Detailed SMT training log;
- User activity log;
- Advanced management of training data files on the cloud.

The functionality testing in LetsMT! has been a continuous task starting very early in the project. After the final test delivery, the platform functionality has continuously been used and tested to ensure that updates have improved the functionality. Aside from testing by project partners, numerous beta testers from the translation industry have tested the platform functionality and provided valuable insights, suggestions and feedback.

## 3.5 Collecting the training data

A great deal of the consortium effort was put into the data collection to address the under-resourced areas of machine translation. These activities have resulted in a large amount of training data that was previously not identified and aligned. Collected data was processed and is provided on the LetsMT! Platform for use in training the MT engines.

The complete data flow from initiating a batch of parallel documents through converting files to appropriate formats, aligning parallel sentences, cleaning up bad sentence alignments, ending up with a sentence aligned corpus is handled by the LetsMT! Platform. Users can now handle the data processing workflow by using LetsMT!, and large amounts of training data are made publicly available for general use and in different subject domains.

In order texts could be used as training data it is important to have information and to be able to administer such information about the origin of the texts, subject domain, text formats, etc. (metadata), same as information about possible use and the IPR of the data.

The needs for metadata were identified from the analysis of user requirements and the functional specification report. Through an examination of various metadata standards, the consortium decided to use the Dublin Core Metadata (DC) for the metadata elements when possible, defining an add-on covering the special metadata needs of the project. Specifications on how the training data should be administrated with relation to upload to a resource repository, responsibility for upload, quality assurance, persistency, and accessibility of data were also determined

The general contents of the agreements with text providers and future users have been established and 52 Data Provider Agreements and/or User Agreements have been registered

As shown in Table 1, the targeted volumes of collected data have been strongly exceeded.

**Table 1** Planned and actual results of the LetsMT! data collection efforts.

| Initial target at the beginning of the project | Results achieved by the end of the project |
|---|---|
| 20 language pairs | **104 languages** in total with **225 language pairs**. |
| 1.5M sentences per language pair | For 225 language pairs the amount of collected parallel data **by far exceeds 1.5M**.<br>For the 10 language pairs in focus more than 1.5M sentences for each language have been collected. |
| 5 under resourced languages | Besides the initially planned five languages, **many more under resourced languages** are represented on the Platform. |
| Data specific to 5 domains | **6 domains** are represented with more than 1M sentences and **two domains** represented by a smaller amount of data. |

By the end of the project, total of 1,480 billion aligned sentences have been collected. The majority of these, 1240 billion aligned sentences, are available for a public use. The project selected ten under-resourced languages which have been in special focus, i.e., languages of the project partner countries Croatian, Czech, Danish, Dutch, Latvian, Swedish plus Slovak, Polish, Lithuanian, and Estonian. The total number of parallel sentences for these under-resourced languages is 361M.

Total amount of collected and uploaded training data for the 10 languages in focus is summarised in Table 2, where data is represented in millions (M) of parallel sentences per language or language pair.

**Table 2** LetsMT! language statistics of the 10 languages in focus

| Languages/size | English (EN) | Croatian (HR) | Czech (CS) | Danish (DA) | Estonian (ET) | Latvian (LV) | Lithuanian (LT) | Polish (PL) | Slovak (SK) | Swedish (SV) | Dutch (NL) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| English (EN) | | 1.6M | 28.6M | 22.9M | 14.9M | 18.1M | 16.1M | 17.2M | 11M | 18.7M | 24.3M |
| Croatian (HR) | 1.6M | | | | 0.1M | 0.1M | 0.1M | 0.1M | 0.1M | 0.1M | 0.1M |
| Czech (CS) | 28.6M | | | 4.6M | 5.1M | 5.0M | 5.1M | 5.3M | 10.3M | 4.5M | 4.5M |
| Danish (DA) | 22.9M | | 4.6M | | 3.7M | 4.7M | 4.8M | 4.9M | 4.7M | 6.1M | 6.5M |
| Estonian (ET) | 14.9M | 0.1M | 5.1M | 3.7M | | 5.6M | 5.6M | 5.9M | 5.6M | 4.9M | 5.1M |
| Latvian (LV) | 18.1M | 0.1M | 5.0M | 4.7M | 5.6M | | 6.6M | 5.8M | 5.6M | 4.7M | 4.9M |
| Lithuanian (LT) | 16.1M | 0.1M | 5.1M | 4.8M | 5.6M | 6.6M | | 5.9M | 5.5M | 4.8M | 5M |
| Polish (PLT) | 17.2M | 0.1M | 5.3M | 4.9M | 5.9M | 5.7M | 5.9M | | 4.5M | 3.7M | 5.2M |
| Slovak (SK) | 11M | 0.1M | 10.3M | 4.7M | 5.5M | 5.5M | 5.5M | 4.5M | | 3.7M | 3.9M |
| Swedish (SV) | 18.7M | 0.1M | 4.5M | 6.1M | 4.9M | 4.7M | 4.8M | 3.7M | 3.7M | | 5.1M |
| Dutch (NL) | 24.3M | 0.1M | 4.5M | 6.5M | 5.1M | 4.9M | 5M | 5.2M | 3.9M | 5.1M | |

Total amount of collected and uploaded training data by subject domains is summarised in Table 3.

**Table 3** Collected data by subject domains

| Domain | Size in parallel sentences |
|---|---:|
| Law | 862M |
| Biotechnology and Health | 243M |
| Finance | 33.3M |
| Information technology and data processing | 27.5M |
| National and international affairs | 6M |
| Education | 1.4M |
| Electronics | 0.4 M |
| Tourism | 2.1K |
| Other | 67M |

To train MT engines for business and financial news translation scenario, parallel corpora was collected for Polish, Swedish, Dutch, Danish, and Czech, each matched with the parallel text in English. To ensure a high quality of parallel data, corporate business and finance files were collected manually. The document type mainly is financial documentation, but there are also news, announcement, reports, and research. In total, approximately 34M words were collected in the above-mentioned 5 languages. All of this data is provided for public use.

The collected amount was sufficient to train efficient translation engines as tested by benchmarking with results of *Google Translate*. For this specific domain, LetsMT! engines achieved better translation results.

**Table 4** Size of collected business and finance news corpora in words

| Language pairs | Subject domain | Volume in words |
|---|---|---|
| **PL-EN** | Business/Finance | >5M |
| **SV-EN** | Business/Finance | > 11,5M |
| **NL-EN** | Business/Finance | > 8M |
| **DA-EN** | Business/Finance | > 6M |
| **CS-EN** | Business/Finance | > 1M |

To make LetsMT! usable for training of MT systems in the IT domain, a large number of private localisation corpora has been uploaded to the LetsMT! Platform. Localisation data are mostly private, as these usually are IPR protected and cannot easily be made available for a broader audience.

LetsMT! allows language service providers to upload their own private data to LetsMT! as private data, and in training to combine their own private data with the public data from the repository.

The localisation corpora have been successfully used to train and evaluate SMT systems.

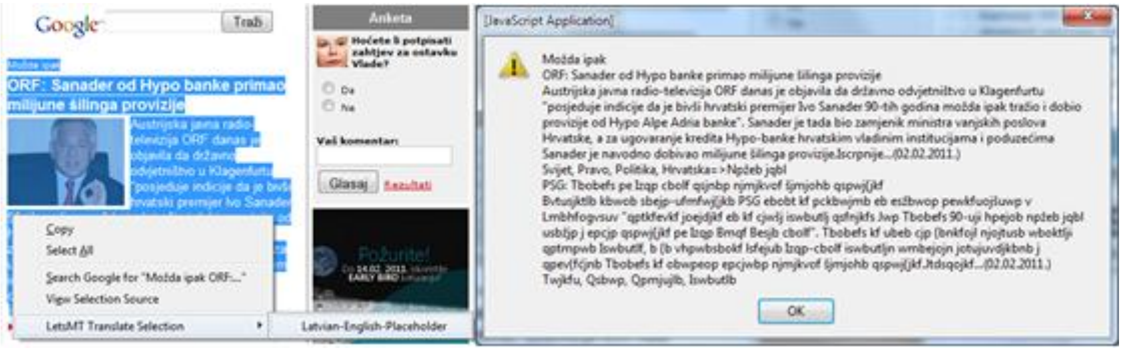
## 3.6 LetsMT! widgets and browser plugins

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 meant to be source code integrated 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. The widget offers two basic functions to prospective users: custom integration within the client website and translation of websites by using the LetsMT! translation facilities.

The widget user interface is very simple. After a user selects a pair of languages, (s)he can click on the "Translate" button. JavaScript running in the user's browser will issue a number of requests to the translate service (to be more precise, currently that is the redirect script), and as soon as the results start arriving from the translate service they will be used to replace the contents of the original page. When the user clicks on the translation button, multiple asynchronous translation requests are issued. The number of translation requests that are issued is approximately equal to the number of different non-empty DOM text elements minus the number of inline elements. As soon as translation results are received, the contents of the translated elements are replaced. Sometimes old inline elements and text nodes are deleted and are replaced with new elements. The structure of the translated page should not be affected by the translation.

In order to translate a page that does not contain the LetsMT! widget or to translate only a part of some webpage, a user needs to install a browser plug-in. The LetsMT! Firefox extension is a single file with .xpi extension. User installs it by double-clicking the file and later clicking on the installation button in Firefox. Most of the functionality is available via the context menu that can be accessed by a right-click (or control-click on a Mac) on the web page. This is illustrated in Figure 4.

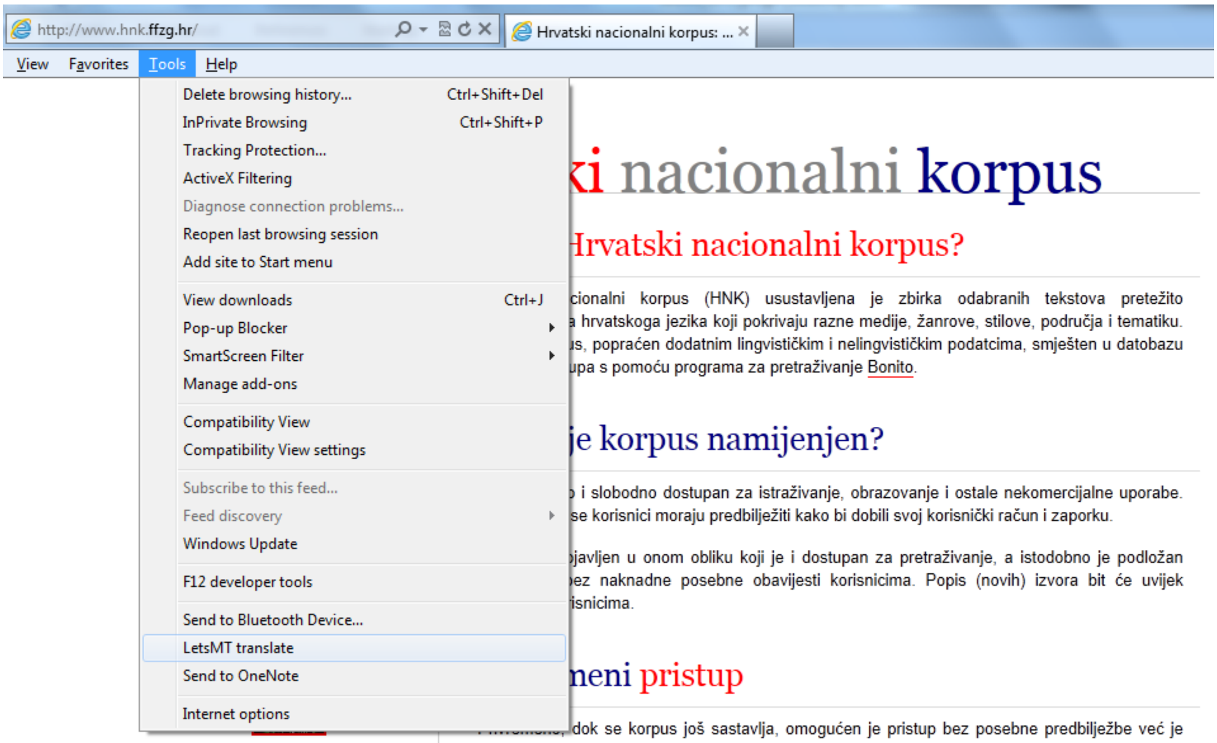

Google
Traži
ORF: Sanader od Hypo banke primao milijune šilinga provizije
Anketa
Hoćete li potpisati zahtjev za ostavku Vlade?
Da
Ne
Vaš komentar:
Glasaj
Rezultati
Požurite!
Copy
Select All
Search Google for "Možda ipak ORF-..."
View Selection Source
LetsMT Translate Selection
Latvian-English-Placeholder
[JavaScript Application]
OK

Možda ipak
ORF: Sanader od Hypo banke primao milijune šilinga provizije
Austrijska javna radio-televizija ORF danas je objavila da državno odvjetništvo u Klagenfurtu "posjeduje indicije da je bivši hrvatski premijer Ivo Sanader 90-tih godina možda ipak tražio i dobio provizije od Hypo Alpe Adria banke". Sanader je tada bio zamjenik ministra vanjskih poslova Hrvatske, a za ugovaranje kredita Hypo-banke hrvatskim vladinim institucijama i poduzećima Sanader je navodno dobivao milijune šilinga provizije.Iscrpnije...(02.02.2011.)
Svijet, Pravo, Politika, Hrvatska=>Npžeb jqbl
PSG: Tbobefs pe Izqp cbolf qsjnbp njmjkvof ŧjmjohb qspwj{jkf
Bvtusjktlb kbwob sbejp-ufmfwj{jkb PSG ebobt kf pckbwjmb eb esžbwop pewkfuoj[uwp v Lmbhfogvsuv "qptkfevkf joejdjkf eb kf cjwdj iswbutlj qsfnjkfs Jwp Tbobefs 90-uji hpejob npžeb jqbl usbžjp j epcjp qspwj{jkf pe Izqp Bmqf Besjb cbolf". Tbobefs kf ubeb cjp {bnkfojl njojtusb wboktlji qptmpwb Iswbutlf, b {b vhpwbsbokf lsfejub Izqp-cbolf iswbutljn wmbejojn jotujuvdjkbnb j qpev{fćjnb Tbobefs kf obwpeop epcjwbp njmjkvof ŧjmjohb qspwj{jkf.Jtdsqojkf...(02.02.2011.)
Twjkfu, Qsbwp, Qpmjujlb, Iswbutlb

**Figure 4** Translation using the Firefox plugin

If a part of the web page (e.g., some text) is selected, LetsMT! extension adds a new menu entry "LetsMT! translate selection" that has a submenu listing all the available systems (pairs of source and target languages). The list of language pairs is obtained by contacting the LetsMT! Service. If there is no selection, LetsMT! extension adds a new menu entry "LetsMT! translate page" that also offers the same list of language pairs. When LetsMT! extension is installed, it comes with generic credentials (username and password). Users are able to provide different credentials by changing extension preferences. In the Add-ons window (one can access it by selecting Tools->Add-ons in the menu), "Extension" panel should be selected. After the LetsMT! extension is selected, the "Preferences" button will be visible. When the "Preferences" button is clicked, the user can provide different credentials. The first version of the extension was tested only in Firefox 3.* and 4.0 beta.

The usage of the LetsMT! translation plug-in for Microsoft Internet Explorer is illustrated in Figure 5



ki nacionalni korpus

Hrvatski nacionalni korpus?

cionalni korpus (HNK) usustavljena je zbirka odabranih tekstova pretežito a hrvatskoga jezika koji pokrivaju razne medije, žanrove, stilove, područja i tematiku. ıs, popraćen dodatnim lingvističkim i nelingvističkim podatcima, smješten u datobazu upa s pomoću programa za pretraživanje Bonito.

je korpus namijenjen?

i slobodno dostupan za istraživanje, obrazovanje i ostale nekomercijalne uporabe. se korisnici moraju predbilježiti kako bi dobili svoj korisnički račun i zaporku.

javljen u onom obliku koji je i dostupan za pretraživanje, a istodobno je podložan ez naknadne posebne obavijesti korisnicima. Popis (novih) izvora bit će uvijek isnicima.

neni pristup

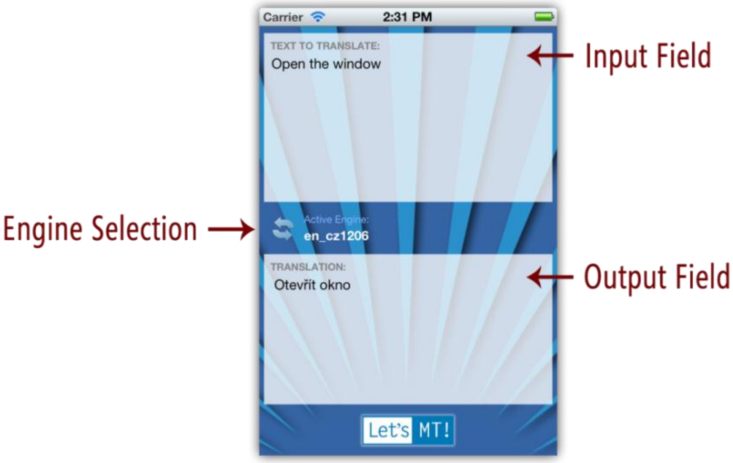
dok se korpus još sastavlja, omogućen je pristup bez posebne predbilježbe već je

**Figure 5** Translation using the LetsMT! MSIE plugin

The browser plug-ins support all the latest versions of Mozilla Firefox and Microsoft Internet Explorer browsers, with improved user experience. The translation widget uses JSONP web service facilities and does not require a redirect script. The modules with installation instructions are made available on the LetsMT! public website.

## *3.7 LetsMT! Mobile translation application*

LetsMT! can be used via mobile environment as well. Currently, it is possible to use LetsMT on mobile devices which use Android or iOS operating systems (Figure 6 and Figure 7).

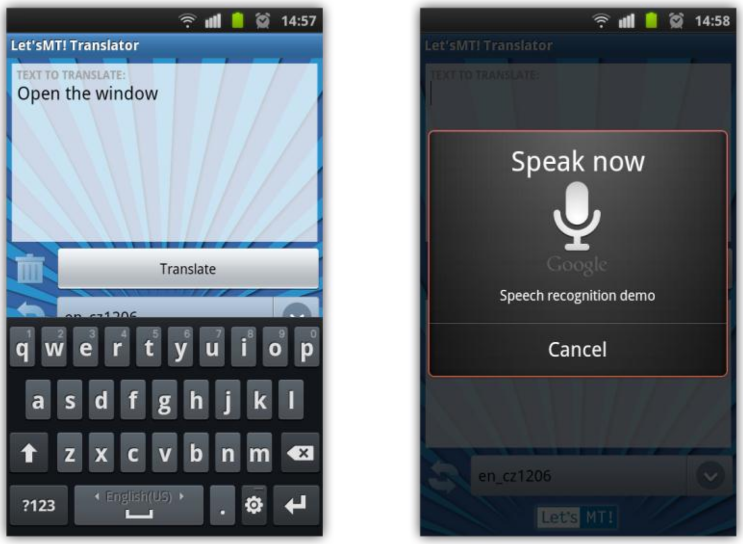


TEXT TO TRANSLATE:
Open the window

Active Engine:
en_cz1206

TRANSLATION:
Otevřít okno

**Figure 6** Main Screen of LetsMT! Translator for iOS.



TEXT TO TRANSLATE: Open the window

**Figure 7** LetsMT! on Android, both input methods, keyboard and Voice recognition

These mobile clients allow LetsMT! users to connect customized engines in situations when a desktop computer is not usually available. In addition to the text based input, the Android version offers also speech to text facility which means that you can tell a sentence, the sentence will be written into a device, translated and finally the device will tell you the translated sentence. Of course, this process is quite complicated and the quality of results is often questionable, however, speech to text is continually improving and its quality should be better by time.

Having customized MT systems deployed on mobile devices is a significant added value in comparison with other mobile MT solutions that provide only general domain translation.

## 3.8 Machine translation services for business and financial news

To assess the usability of LetsMT! in the financial and business news translation scenario, several MT systems were trained and evaluated.

Summary of the automatic evaluation scores is provided in Table 5. For each language pair an evaluation set has been randomly selected and excluded from the training data before the LetsMT! systems were trained. These evaluation sets have been translated with both the LetsMT! system and with *Google Translate* for comparison.

For each evaluation set, the evaluation scores for BLEU, METEOR and TER are calculated. It is important to mention that it is only possible to rank BLEU scores and other automatic measures for the same evaluation set since the level of BLEU scores and other measures depend heavily on a language pair and text type.

Please note that TER score measures the number of insertions, deletions, substitutions and shifts and compares it to the number of words in the input sentence. Therefore, a low TER score is better than a high score. For BLEU and METEOR, a higher score is better.

**Table 5** Summary of the system scores. Best systems for each language pair from LetsMT! and scores from using Google Translate. Best systems for each language pair are marked in bold

| System name | Evaluation set | BLEU | METEOR | TER |
|---|---|---|---|---|
| **LetsMT! English-Danish Finance IV** | **Evaluation, EN-DA finance, validated** | **72.5** | **0.493** | **29.1** |
| **Google translation English-Danish** | Evaluation, EN-DA finance, validated | 40.1 | 0.337 | 53.6 |
| **LetsMT! English - Swedish Finance v3** | **Evaluation EN-SV Finance, validated** | **65.5** | **0.459** | **32.5** |
| **Google translation English - Swedish** | Evaluation EN-SV Finance, validated | 36.8 | 0.3218 | 55.3 |
| **LetsMT! English - Dutch Finance v3** | **Evaluation EN-NL Finance, validated** | **63.0** | **0.435** | **38.7** |
| **Google translation English - Dutch** | Evaluation EN-NL Finance, validated | 37.8 | 0.336 | 57.4 |
| **LetsMT! Dutch - English Ver3(M28)** | **Evaluation EN-NL Finance, validated** | **62.1** | **0.442** | **38.9** |
| **Google translation Dutch - English** | Evaluation EN-NL Finance, validated | 36.1 | 0.358 | 53.5 |
| **LetsMT! English - Czech v3 (M28)** | **Evaluation EN-CS Finance, validated** | **59.9** | **0.411** | **41.8** |
| **Google translation English - Czech** | Evaluation EN-CS Finance, validated | 22.3 | 0.223 | 80.9 |
| **LetsMT! English - Polish Finance v3 (M28)** | **Evaluation EN-PL Finance, validated** | **53.1** | **0.359** | **57.4** |
| **Google translation English - Polish** | Evaluation EN-PL Finance, validated | 19.4 | 0.206 | 78.6 |

The Table 5 illustrates that the 6 LetsMT! systems – all trained for text within specific domains – outperform Google Translate for the chosen evaluation sets.

The evaluation scores also show that all the three evaluation measures rank the systems for each language pair in the same prioritized order. This indicates that the automatic evaluation measures are in concordance with each other when looking at a specific language-pair.

The values of the evaluation metrics also indicate that for each language pair (except English-Croatian) a system now exists that is usable for post-editing the output. The English-Danish Finance IV has the best scores, with BLEU scores higher than 70, and TER scores below 30. These scores indicate a good output quality. Furthermore the English - Swedish Finance v3, English - Dutch Finance v3 and Dutch - English Ver3 (June, 2012) systems get BLEU scores higher than 50, and TER scores below 40. All these systems look very promising for getting a large benefit when using the system alone or in combination with a CAT tool.

News Analytics event based scenario translation is used for the functional evaluation of sentences. This method is a combination of semantic analytics for financial markets and combined text extraction, event recognition and pattern matching techniques from the SMT output.

ViewerPro is a software which was used to transform natural language news messages to computer readable news events. This procedure is based on a domain specific ontology. It contains the concepts and their lexical representations relevant to a specific domain, e.g. Company Ontology describing all companies of interest to the user in terms of their synonyms, ISIN codes, ticker symbols etc.

Figure 8 shows the results of the system testing. These findings indicate that, despite the rather poor translation quality of less than 30% adequately translated sentences, still in 80% of these sentences the correct ViewerPro events were detected. Assuming a rule of thumb that natural language processing system needs to be about 75% accurate to start giving useful results, these findings clearly illustrate that a sub optimal translation can still be of use for further processing by ViewerPro.

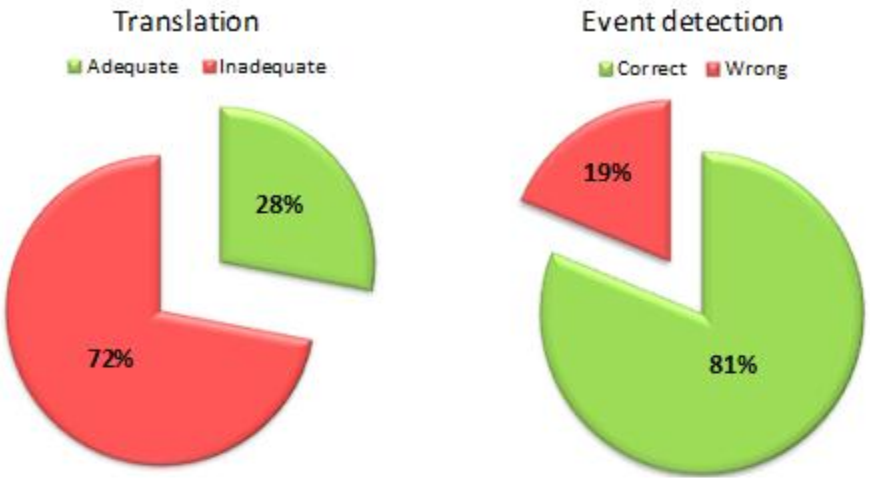


**Figure 8** Percentage of adequate translations (left) vs percentage of correct ViewerPro events (right) over the dataset of 500 random chosen messages

Another evaluation comes from benchmarking LetsMT! with *Google Translate*. Over the dataset of 500 randomly chosen messages, *Google Translate* performance was 33% wrong and 67% adequate ViewerPro events.

The resulting translations of event by ViewerPro account for over 80% of the cases. LetsMT! performs 8% adequate event translation against the benchmark *Google Translate* of 67%, which is a better result by 14%.

This makes the LetsMT! system immediately useful for the extension of ViewerPro into processing non-English data sources.

## 3.9 Facilities and evaluation of MT usage in localisation

One of the aims of LetsMT! is to implement MT in the translation workflow for LSPs (language service providers). Professional translation agencies work with specific XML-based formats such as XLIFF and TMX. In such documents, text strings contain more information (e.g., placeholders) than a pure text and it is necessary to maintain the integrity of this information during translation. This markup cannot be translated but it has to be replaced appropriately into the target language. Therefore, it was decided to develop a specific tag translation mechanism.

The tag translation feature maintains a high accuracy which allows deployment in the general translation environment and serves as a general proof of concept for the purpose of the LetsMT! project. The results of this work (software code) are integrated into the LetsMT! Platform and also made publicly available under the LGPL license, besides implementation procedure has been provided[3]. It allows professional companies to further customise or enhance the tagging functionality.

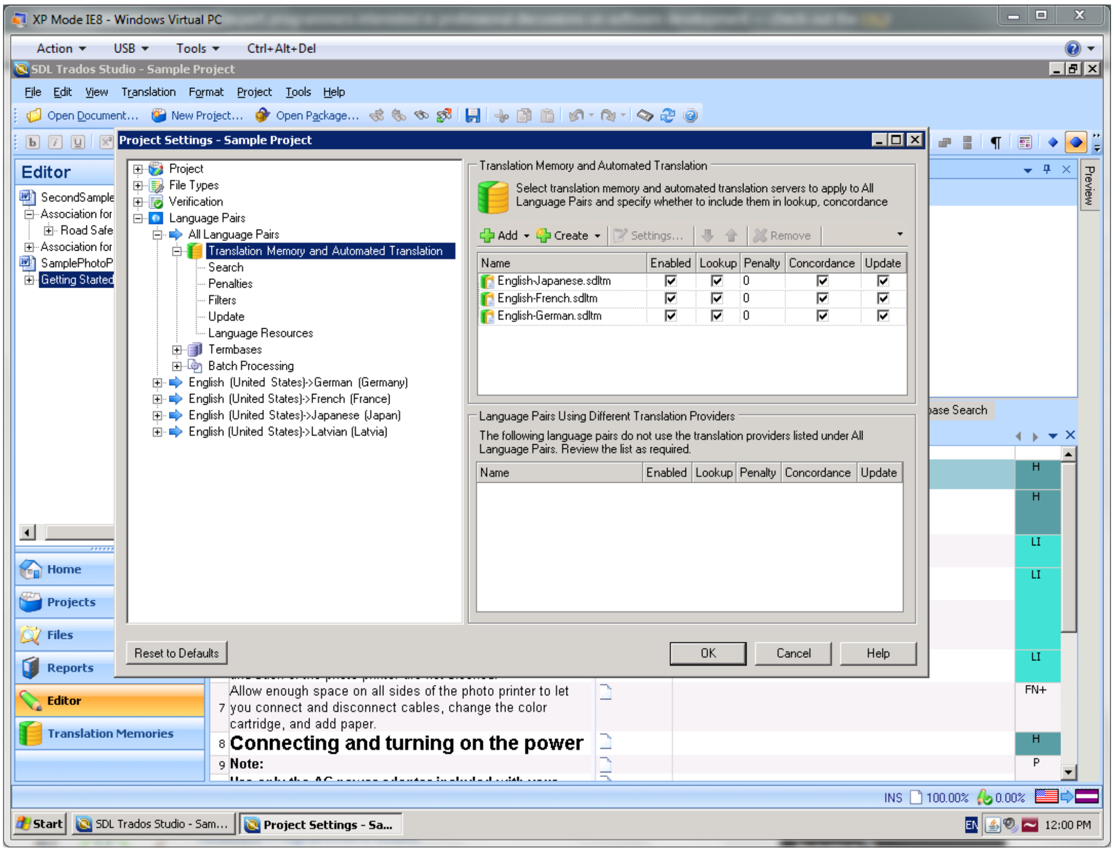
LSPs can use the LetsMT! Platform in different ways. If an LSP is using SDL Trados Studio or MemoQ CAT tools in their translation project, the easiest solution would be to use the LetsMT! plug-in. The most recent plug-in for SDL Trados Studio (Figure 9) can be downloaded from the LetsMT! site[4].

Kilgray's MemoQ has integrated the plug-in into its main application (starting from version 6). If an LSP is using a different CAT tool, it is easy to integrate LetsMT! into workflow via the LetsMT! API[5].

[3] http://www.mt-archive.info/EAMT-2011-Hudik.pdf

[4] https://www.letsmt.eu/Integration.aspx

[5] https://www.letsmt.eu/downloads/Public_translation_API.pdf



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

| Name | Enabled | Lookup | Penalty | Concordance | Update |
|---|---|---|---|---|---|
| English-Japanese.sdltm | ☑ | ☑ | 0 | ☑ | ☑ |
| English-French.sdltm | ☑ | ☑ | 0 | ☑ | ☑ |
| English-German.sdltm | ☑ | ☑ | 0 | ☑ | ☑ |

The following language pairs do not use the translation providers listed under All Language Pairs. Review the list as required.

| Name | Enabled | Lookup | Penalty | Concordance | Update |
|---|---|---|---|---|---|

7 Allow enough space on all sides of the photo printer to let you connect and disconnect cables, change the color cartridge, and add paper.

8 Connecting and turning on the power

9 Note:

**Figure 9:** SDL Trados Studio plug-in configuration – open translation memory settings

During the project six MT systems were trained in the localisation domain. Blue, METEOR and TER metrics were used for automated evaluation of the systems (see Table 6).

**Table 6** Summary of system scores. Best systems for each language pair are marked with bold

| System name | Evaluation set | BLEU | METEOR | TER |
|---|---|---|---|---|
| **LetsMT! Danish-English Uni.adm.** | **KU2b – evaluation set** | **56.30** | **0.41** | **53.92** |
| Google translation | KU2b - evaluation set | 29.40 | 0.32 | 64.15 |
| **LetsMT! English-Estonian IT, Ver1(M24)** | **EN-ET-IT-v1 eval, set** | **60.70** | **0.42** | **43.60** |
| Google translation | LetsMT IT evaluation set: English-Estonian | 20.00 | 0.21 | 81.30 |
| **LetsMT! English-Hungarian IT, Ver1(M28)** | **EN-HU-IT-v1 eval, set** | **59.50** | **0.41** | **46.10** |
| Google translation | EN-HU-IT-v1 eval, set | 19.20 | 0.22 | 81.10 |
| English-Latvian IT, Ver1,1(M28) | LetsMT! IT evaluation set: English Latvian | 60.80 | 0.43 | 39.20 |
| English-Latvian IT, Ver2(M28) | LetsMT! IT evaluation set: English-Latvian | 61.20 | 0.43 | 38.60 |
| **English-Latvian IT, Ver2.1(M28)** | **LetsMT! IT evaluation set: English-Latvian** | **69.57** | **0.48** | **31.80** |
| Google translation | LetsMT! IT evaluation set: English-Latvian | 28.00 | 0.27 | 66.60 |
| **English-Lithuanian IT, Ver2(M28)** | **LetsMT! IT evaluation set: English-Lithuanian** | **59.70** | **0.43** | **40.10** |
| Google translation | LetsMT! IT evaluation set: English-Lithuanian | 22.30 | 0.24 | 72.50 |
| **English-Polish IT, Ver2(M28)** | **EN-PL-v2 eval, set** | **72.20** | **0.50** | **38.90** |
| Google translation | EN-PL-v2 eval, set | 23.50 | 0.22 | 79.00 |
| **English-Czech IT, Ver1(M28)** | **EN-CS-IT evaluation set** | **71.10** | **0.49** | **34.14** |
| Google translation | EN-CS-IT evaluation set | 27.8 | 0.266 | 67.5 |

Concerning the amount of training data, all systems – expect the *Danish-English Uni.adm.* – are trained with reasonable large in-domain resources. The *Danish-English Uni.adm.* system has been trained on a scant 24.000 in-domain parallel segments. Adding more in-domain data might not always improve the evaluation results, as the combination of relevant vocabulary and text from the same text type also are very important for good translation quality.

According to the automatic evaluation measures all the LetsMT! systems achieved significantly better scores than the translation output by *Google Translate.*

The values of the evaluation metrics also indicate that for each language pair a system now exists that is very usable when post-editing the output. The English-Czech IT, Ver1(M28), English-Latvian IT Ver2.1(M28) and English-Polish IT, Ver2(M28) get high level scores, with BLEU scores around 70, and TER scores below 40. These systems look very promising for getting a large benefit when using the system alone or in combination with a CAT tool.

Moravia and Tilde (bigger LSPs) have tested LetsMT! platform on their real-life projects. In particular, EN-CZ, EN-PL, EN-HU, EN-ET, EN-LV and EN-LT language pairs were tested. The use of the SMT suggestions in addition to the translation memories lead to the increase of translation performance in the range from 18% to 32.9%[6].

# 4 Main dissemination activities and exploitation of results

Dissemination activities defined different target audience groups and channels to be used to communicate information about the project. Also, a unique visual identity was defined that includes a logo, colour schema, typefaces and templates that were used when presenting the LetsMT! project in an awareness rising campaign.

These channels were:

- LetsMT! project public web site;
- posters, leaflets and T-shirts;
- two specialised LetsMT! workshops for two different target audience groups: translation and localisation industry and scientific community;
- scientific publications in journals and conference proceedings;
- recorded LetsMT! presentations accessible as on-line video lectures;
- short video clips that are used as introductory video tutorials.

The initial LetsMT! project website http://project.letsmt.eu is the place where all information related to the LetsMT! project is stored and made available to the global web audience. In addition to the classic dissemination of website content, information about the project is also provided through other means, such as video recorded lectures or short video tutorials.

The LetsMT! project organised two specialised LetsMT! workshops where the project achievements were presented in more detail. The first one was prepared for participants from the localisation and translation industry and it took place during the GALA2012 conference, on 25 March 2012 in Monte Carlo. The second workshop was aimed at the scientific community and was organised as the LREC2012 post-conference whole day workshop, on 26 May 2012 in Istanbul.

It should be noted that with the GALA workshop, we have successfully managed to raise the interest amoung the translation and localisation business community for the LetsMT! machine translation services, particularly for under-resourced languages, that are entering into the focus of this industry more and more every day.

The LREC2012 post-conference workshop, together with participation at the LREC2012 main conference, could be regarded as a LetsMT!'s important event in the scientific community, since

[6] For detailed description see: http://www.mt-archive.info/EAMT-2011-Skadins.pdf

LetsMT! was presented with 2 papers at the workshop which was also organised by the project, and the editor-in-chief of the workshop proceedings was one of the LetsMT! partners, while at the LREC2012 main conference LetsMT! was presented with another paper.

Besides organizing two specialized LetsMT! workshops, partners from the project participated at 37 different global, European, regional or national conferences, workshops, awareness days etc. with successful presentations of the project results and that resulted in 13 published journal papers, conference proceedings papers or book chapters. Here we will list only the most important ones: LREC2010 and LREC2012, EAMT2011 and EAMT2012, MT Summit 2011, ACL2012, META-FORUM2010, META-FORUM 2011 and META-FORUM 2012, etc.

Some illustrations of LetsMT! presence at different industry and research events have been provided in the figures below.

**Figure 10** LetsMT! at FLaReNet Forum 2011, Venice

**Figure 11** LetsMT! at META-FORUM2011, Budapest

**Figure 12** LetsMT! organised workshop at the GALA2012 conference, Monte Carlo

**Figure 13** LetsMT! was presented at the LREC 2012 conference in Istanbul.

In addition to these dissemination activities, different portals following topics in translation, localisation, MT, SMT, LT, APIs, etc. have mentioned LetsMT! as a project deserving an interest.

We can call this phenomenon “secondary dissemination” since the project members did not have any additional instrument to influence the appearance of these references to the LetsMT! project apart from the dissemination instruments that were already planned in the dissemination plan. In this respect and since we were not able to collect an exhaustive list of the said references, an illustrative list is given in the Final Dissemination Report. However, the article in *New Scientist* mentions LetsMT! as a possible competitor to Google Translate, and this particular moment deserves a special attention here.

A prominent article about LetsMT! was published by the leading localization and translation industry magazine MultiLingual in its January/February  2012 issue.

Several leaflets and T-shirts were produced and broadly disseminated during different industry and research events.

By the end of the project, more than 80 potential LetsMT! users have requested permissions to evaluate the LetsMT! platform, and more than 40 have tried the LetsMT! platform.

# 5 The project's participants public website and contact details

List of all beneficiaries with the corresponding contact name and details

Tilde SIA, Latvia

URL: http://www.tilde.eu

University of Edinburgh, United Kingdom

URL: http://www.ed.ac.uk/

University of Zagreb, Croatia

URL: http://hnk.ffzg.hr/default_en.htm

University of Copenhagen, Denmark

URL: http://www.ku.dk/

Moravia IT a.s.,, the Czech Republic

URL: http://www.moravia.com

Uppsala University, Sweden

URL: http://www.uu.se

SemLab (Zoorobotics BV), the Netherlands

URL: http://www.semlab.nl