LetsMT!

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

www.letsmt.eu

Project No. 250456

Deliverable D1.7
Elaborated LetsMT! platform deployed

Version 1.0

29/02/2012

Document Information

Deliverable number:	D1.7
Deliverable title:	Elaborated LetsMT! platform deployed
Due date of deliverable according to DoW:	29/2/2012
Actual submission date of deliverable:	29/2/2012
Main Author(s):	TILDE
Participants:	UEDIN, FFZG, UPP, MOR
Reviewer	MOR
Work package:	WP1
Work package title:	LetsMT! platform and infrastructure
Work package leader:	TILDE
Dissemination level:	PU
Version:	V1.0
Keywords:	platform, infrastructure, deployment

History of Versions

Version	Date	Status	Name of the Author (Partner)	Contributions	Description/Approval level
0.1	07/2/2012	Draft	TILDE	-	Initial draft version of document.
0.2	15.02.2012	Draft	TILDE	-	Draft version of document.
1.0	29.02.2012	Final	TILDE	-	Final version of document

EXECUTIVE SUMMARY

The LetsMT! platform has been specified, designed, implemented, elaborated and deployed. The released version is available at the following URL: http://letsmt.eu. This document provides a description of the platform, description of the main features from the user perspective and summary conclusions.

Table of Contents

Abbreviations

Abbreviation	Term/definition
API	Application programming interface.
CAT	Computer aided translation.
DoW	LetsMT! Project Description of Work.
MT	Machine translation.
SMT	Statistical machine translation system.
TMX	Translation Memory eXchange format.
WP	Work package of LetsMT! project – corresponding to the Description of Work.
Corpora	Non-downloadable SMT training data uploaded to the Resource Repository

1 General description

The LetsMT! platform is developed according to Task 1.3 and Task 1.4 of the DoW. It provides integration of the key LetsMT! modules:

- Facilities for sharing of SMT training data;
- Facilities for training and running of SMT engines;
- Facilities for use in a news translation scenario;
- Facilities for use in a localization usage scenario.

Supporting software infrastructure has also been developed and deployed providing mechanisms for user registration, authentication and access rights management and control, as well as a Web page is provided for text translation.

Current implementation of the LetsMT! platform is based on the following project deliverables:

- D1.3 Hardware infrastructure;
- D1.6 Revised functional specification;
- D2.1 Specification of data formats allowed;
- D2.2 SMT resource repository and data processing facilities ready for integration;
- 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.

The LetsMT! platform is deployed on Amazon Cloud Services (AWS) and uses the following AWS services:

- Amazon Elastic Compute Cloud (Amazon EC2) – provides environment of virtual computers (instances) with resizable computing capacity and variety of operating systems;
- Amazon Elastic Block Storage (Amazon EBS) – provides network-attached persistent storage to Amazon EC2 instances;
- Amazon Simple Storage Service (Amazon S3) – provides a highly durable storage designed for mission-critical data storage.

LetsMT! platform is deployed and is available at the following URL: http://letsmt.eu.

Development and deployment of LetsMT! platform was organized in 2 major steps – first publicly available LetsMT! platform beta version was deployed on M17 (see deliverable "D1.4 LetsMT! platform deployed") and second, elaborated version of LetsMT! platform was deployed on M24. Description of latest release of LetsMT! functionality from user perspective is summarized in this deliverable.

The latest version of the LetsMT! platform is ready for users to upload MT training data, train custom SMT systems, and use these systems for translation needs.

2 Main functionality of the LetsMT! platform

This section provides a short description of the main features implemented in the latest version of the LetsMT! platform. This description is provided from the user's perspective to guide the user in trying the respective functionality. More detailed description of the functionality (including technical) can be found in other LetsMT! deliverables – system specification documents

(deliverables *D1.6 "Revised functional specification", "D1.3 Hardware infrastructure", "D2.1 Specification of data formats allowed", "D2.2 SMT resource repository and data processing facilities ready for integration", "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"*).

2.1 User authentication

To log in to the LetsMT! platform, click the Login hyperlink at the top of the LetsMT! webpage. Type user name/password and click Sign In.

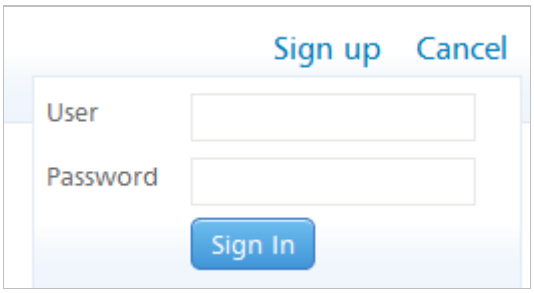


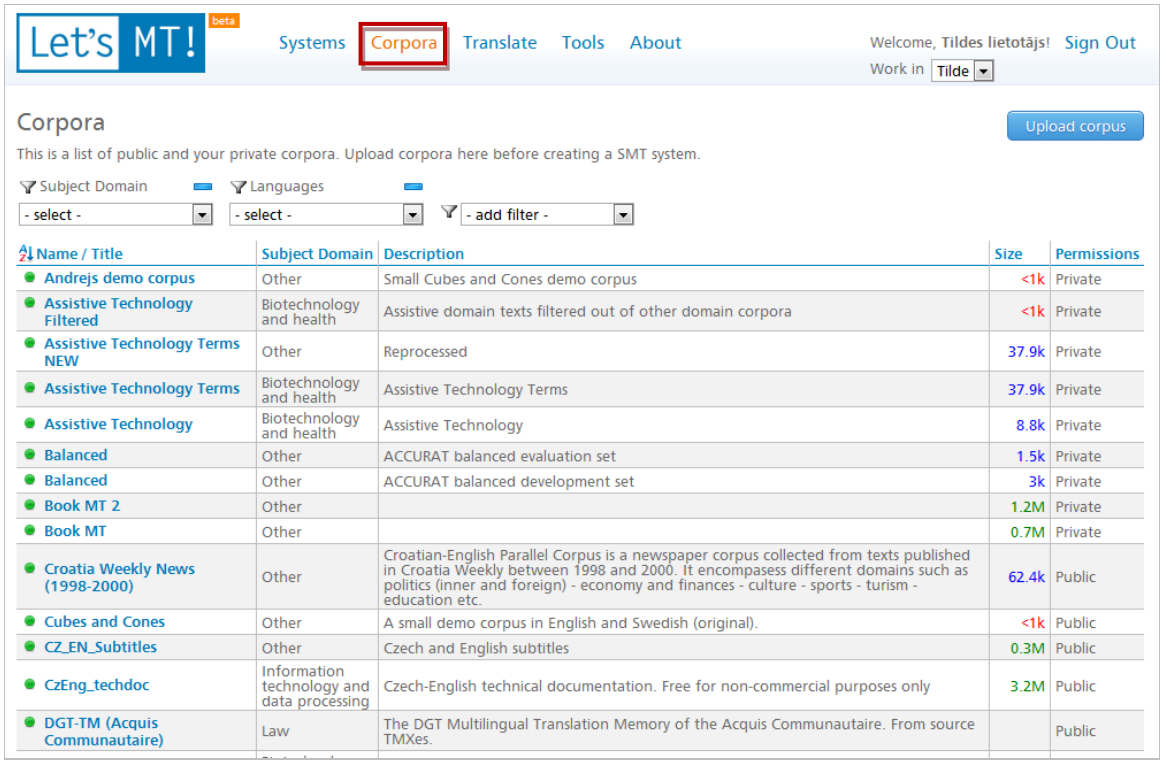
Figure 1. Login

For the details on how to get login credentials and apply for using the current version of the LetsMT! platform, please read https://letsmt.eu/Default.aspx?section=register.

2.2 Work with corpora

2.2.1 Browse corpora

To access the corpora list, click Corpora at the top of the LetsMT! webpage.



Name / Title	Subject Domain	Description	Size	Permissions
Andrejs demo corpus	Other	Small Cubes and Cones demo corpus	<1k	Private
Assistive Technology Filtered	Biotechnology and health	Assistive domain texts filtered out of other domain corpora	<1k	Private
Assistive Technology Terms NEW	Other	Reprocessed	37.9k	Private
Assistive Technology Terms	Biotechnology and health	Assistive Technology Terms	37.9k	Private
Assistive Technology	Biotechnology and health	Assistive Technology	8.8k	Private
Balanced	Other	ACCURAT balanced evaluation set	1.5k	Private
Balanced	Other	ACCURAT balanced development set	3k	Private
Book MT 2	Other		1.2M	Private
Book MT	Other		0.7M	Private
Croatia Weekly News (1998-2000)	Other	Croatian-English Parallel Corpus is a newspaper corpus collected from texts published in Croatia Weekly between 1998 and 2000. It encompasess different domains such as politics (inner and foreign) - economy and finances - culture - sports - turism - education etc.	62.4k	Public
Cubes and Cones	Other	A small demo corpus in English and Swedish (original).	<1k	Public
CZ_EN_Subtitles	Other	Czech and English subtitles	0.3M	Public
CzEng_techdoc	Information technology and data processing	Czech-English technical documentation. Free for non-commercial purposes only	3.2M	Public
DGT-TM (Acquis Communautaire)	Law	The DGT Multilingual Translation Memory of the Acquis Communautaire. From source TMXes.		Public

Figure 2. Corpora list

- The corpora list contains all Public corpora and those Private corpora (by default) with an authenticated user as the owner. If the site visitor does not have authorization, only the list of available public corpora will be displayed;
- To sort the corpora list, please click table headings;

Figure 3. Sorting corpora list

- To filter the corpora list, use the dynamic filters provided at the top of the corpora list. First, select a filter and then select a value in the filter box. The use of filters is summarized in the following two screenshots.

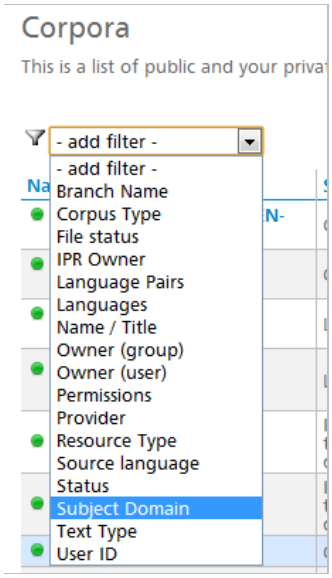


Figure 4. Select filter

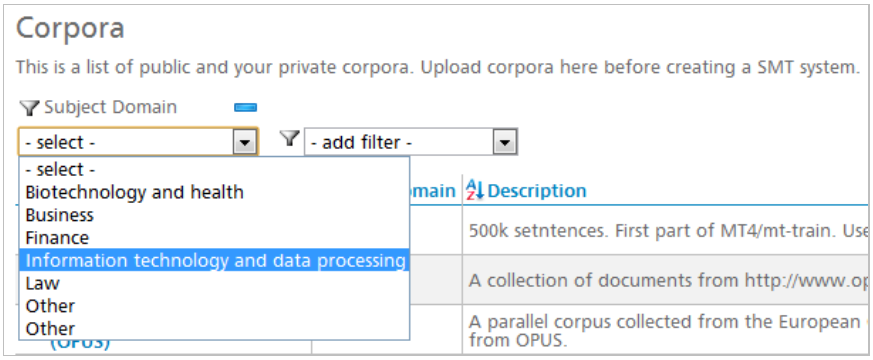


Figure 5. Select filter value

- In the corpora list, a status icon for each corpus is displayed. If the icon is green, the processing of corpus files is successfully completed and the corpus is ready to be used for SMT system training. If the icon is yellow, the processing of files is still in progress, and if the icon is red – an error has occurred during data import.

2.2.2 Upload corpora

Click the Upload corpus button at the right top of the corpora list.

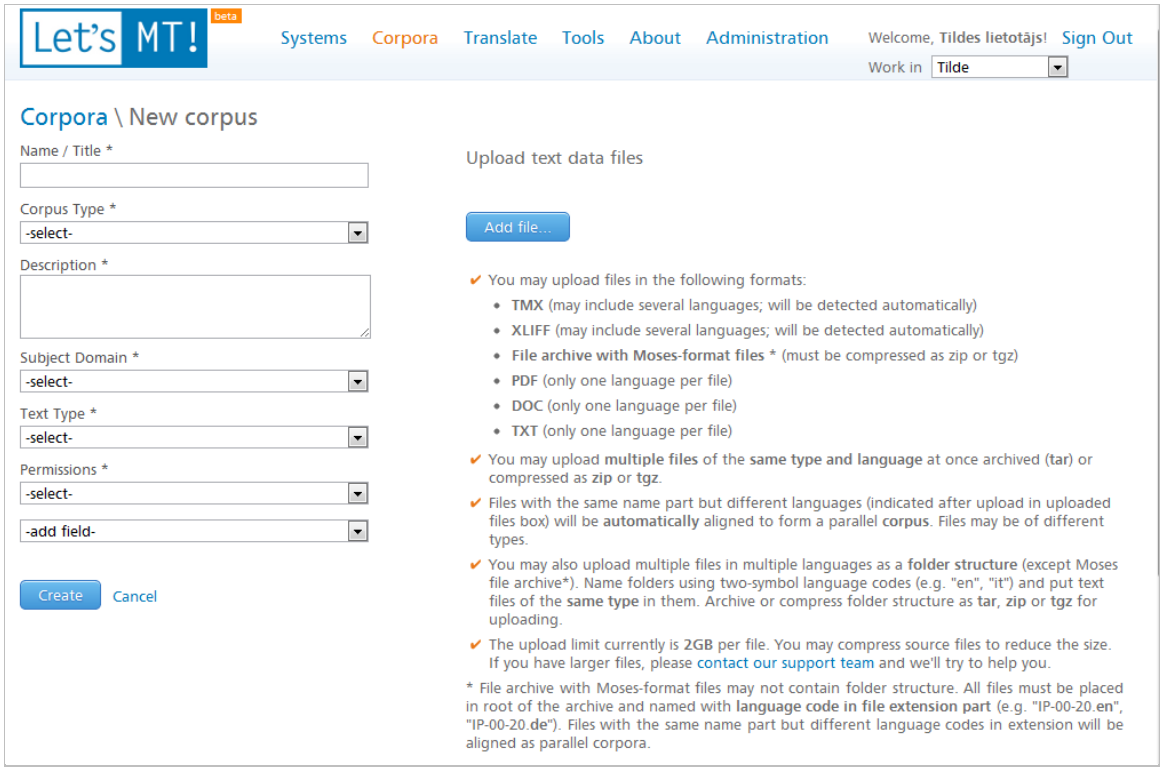


Figure 6. Upload corpora webpage

In the metadata editor of a new corpus, fill the metadata fields and upload corpus training data files. Additional corpus metadata fields can be added by using the -add field- combo-box. Metadata fields that are already in use will be greyed out and will not be available for selection.

The following considerations are important for uploading the training data file:

- You can 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 archives 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 can upload multiple files of the same type and language at a time, archived (tar) or compressed as zip or tgz.
- Files with the same name part but different languages (indicated after upload in the uploaded files box) will be automatically aligned to form a parallel corpus. They may be files of different types.
- You can also upload multiple files in multiple languages as a folder structure (except a Moses file archive*). Name the folders using two-character language codes (e.g., "en", "it") and put the files of the same type in the folders. Archive or compress the folder structure as tar, zip or tgz for uploading.
- The upload limit currently is 2 GB 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.

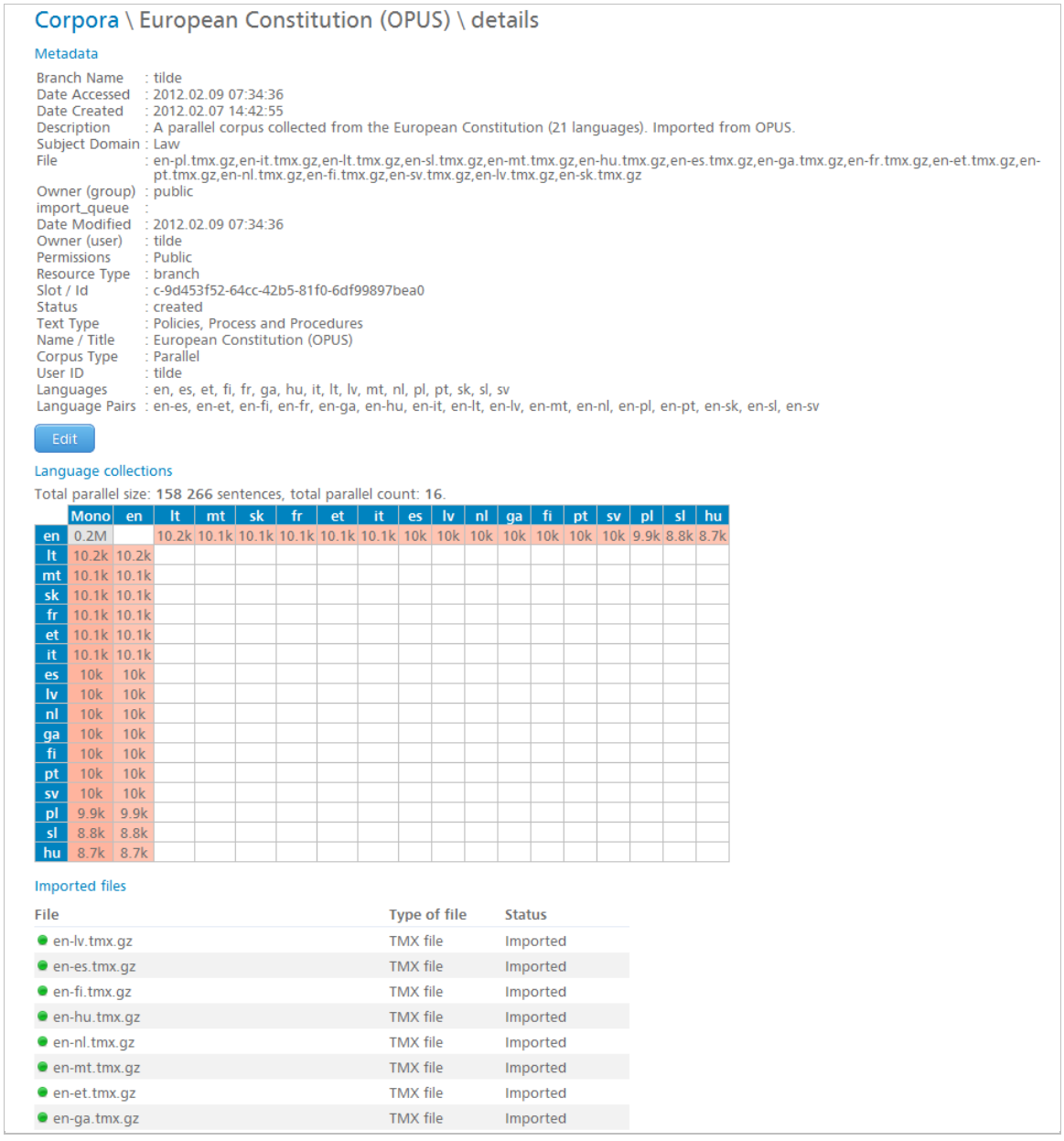
When all metadata fields are filled and the training data files have been uploaded, the final step is to click the Create button.

2.2.3 View information about corpora

In the corpora list, click the corpus name. A short summary describing the corpora will be displayed.

Figure 7. Corpora details (summary)

More detailed view of corpora information is available by clicking Details.



	Mono	en	lt	mt	sk	fr	et	it	es	lv	nl	ga	fi	pt	sv	pl	sl	hu
en	0.2M		10.2k	10.1k	10.1k	10.1k	10.1k	10.1k	10k	10k	10k	10k	10k	10k	10k	9.9k	8.8k	8.7k
lt	10.2k	10.2k																
mt	10.1k	10.1k																
sk	10.1k	10.1k																
fr	10.1k	10.1k																
et	10.1k	10.1k																
it	10.1k	10.1k																
es	10k	10k																
lv	10k	10k																
nl	10k	10k																
ga	10k	10k																
fi	10k	10k																
pt	10k	10k																
sv	10k	10k																
pl	9.9k	9.9k																
sl	8.8k	8.8k																
hu	8.7k	8.7k																

Figure 8. Corpora details (detailed view)

It takes time to process data files in the Resource Repository. The platform user can use corpora for SMT systems training only when all uploaded data files are processed and stored in internal Resource Repository data format. LetsMT! platform user can view the status summary of the processed corpora files in the corpora details view – the Imported files section. In this section, all corpora files are listed in a table; the processed files are marked with green icon, the files under processing are marked with yellow icon. If an error has occurred during data processing, a red icon is displayed.

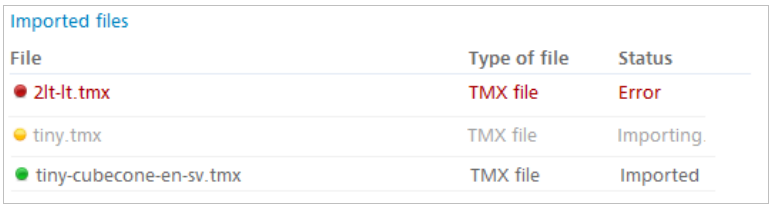


Figure 9 List of corpora files with status summary

2.2.4 Edit corpus metadata/Delete corpus

In the corpora list, click the corpus name and then click Edit.

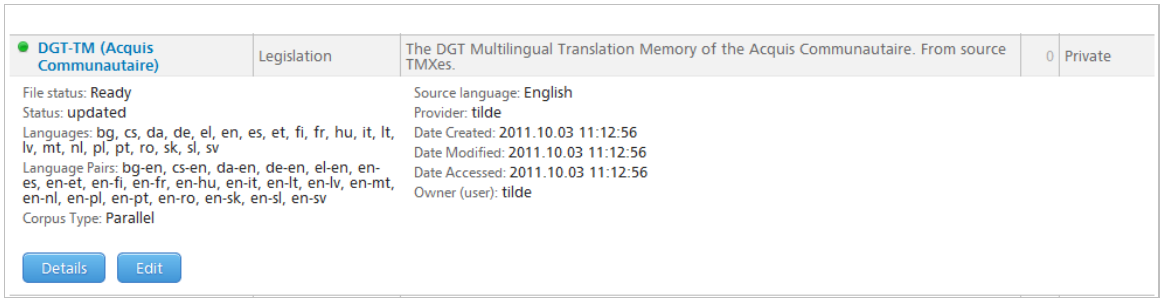


Figure 10. Editing of corpora metadata

This will open the corpora metadata editor.

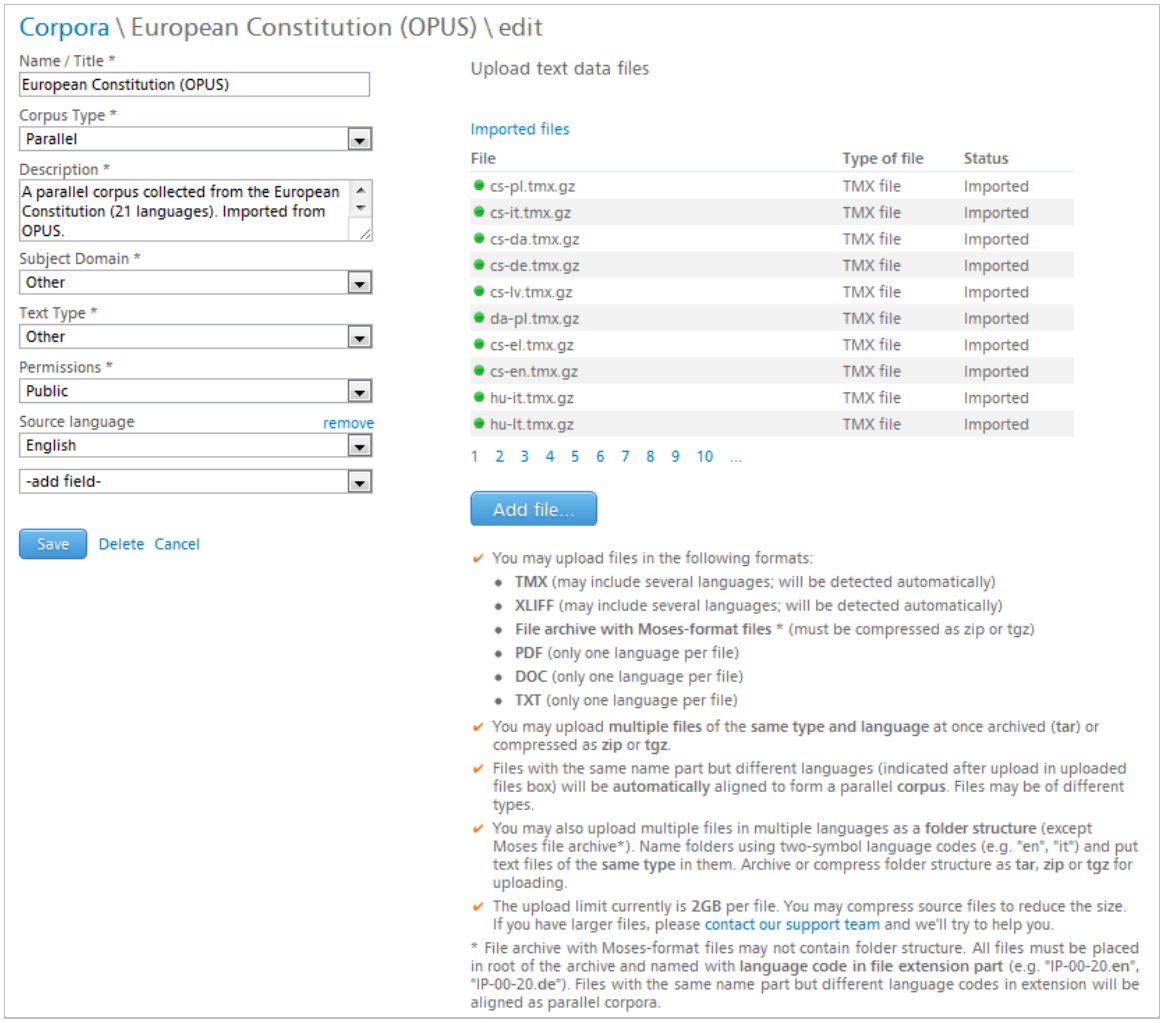


Figure 11. Corpora editing page

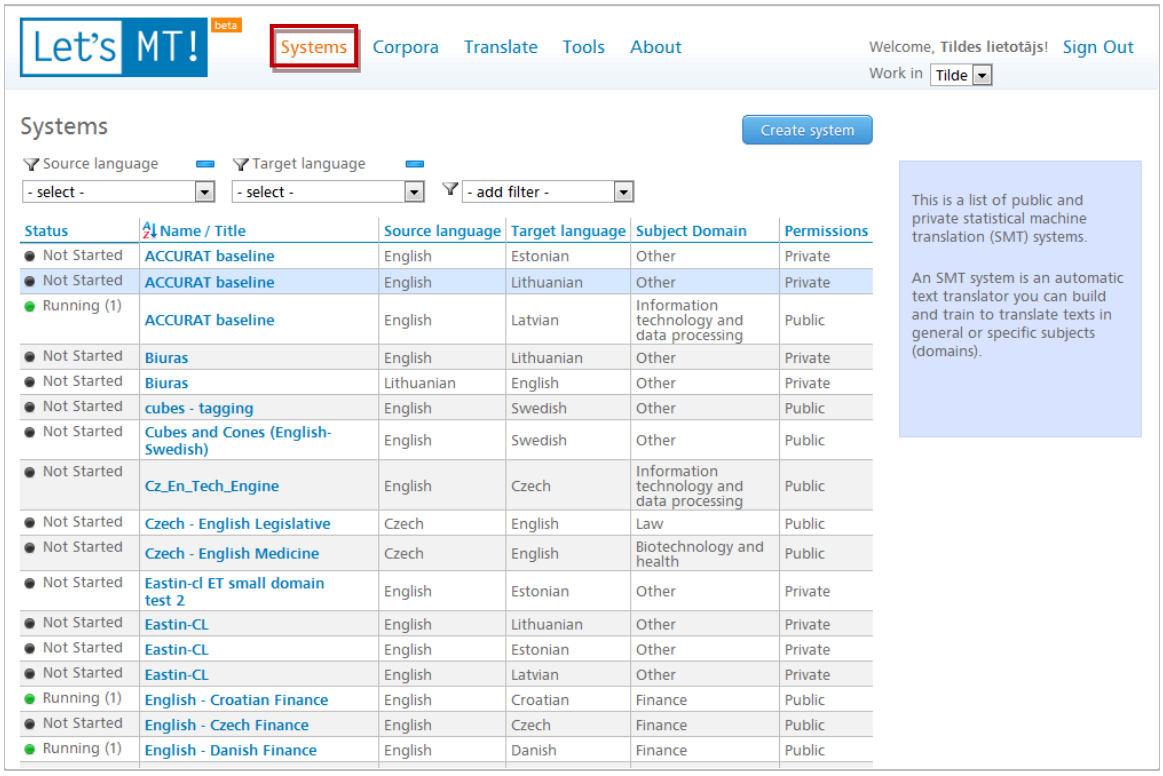
In the metadata editor, you can add and/or update corpus metadata fields and upload additional training data files.

- To save changes, click Save;
- To discard changes, click Cancel;
- To delete a corpus, click Delete.

2.3 Work with SMT systems

2.3.1 Browse SMT systems

To access the SMT systems list, click SMT systems at the top of the LetsMT! webpage.



Status	Name / Title	Source language	Target language	Subject Domain	Permissions
Not Started	ACCURAT baseline	English	Estonian	Other	Private
Not Started	ACCURAT baseline	English	Lithuanian	Other	Private
Running (1)	ACCURAT baseline	English	Latvian	Information technology and data processing	Public
Not Started	Biuras	English	Lithuanian	Other	Private
Not Started	Biuras	Lithuanian	English	Other	Private
Not Started	cubes - tagging	English	Swedish	Other	Public
Not Started	Cubes and Cones (English-Swedish)	English	Swedish	Other	Public
Not Started	Cz_En_Tech_Engine	English	Czech	Information technology and data processing	Public
Not Started	Czech - English Legislative	Czech	English	Law	Public
Not Started	Czech - English Medicine	Czech	English	Biotechnology and health	Public
Not Started	Eastin-cl ET small domain test 2	English	Estonian	Other	Private
Not Started	Eastin-CL	English	Lithuanian	Other	Private
Not Started	Eastin-CL	English	Estonian	Other	Private
Not Started	Eastin-CL	English	Latvian	Other	Private
Running (1)	English - Croatian Finance	English	Croatian	Finance	Public
Not Started	English - Czech Finance	English	Czech	Finance	Public
Running (1)	English - Danish Finance	English	Danish	Finance	Public

Figure 12. SMT systems list

- Like the corpora list, the SMT systems list contains all Public systems and those Private SMT systems with an authenticated user as the owner. If the user is not authenticated, only public SMT systems will be displayed;
- Data sorting in a SMT systems list. It is identical to the sorting of a corpora list (see paragraph Browse corpora) – click table headings;
- Data filtering in a SMT systems list. Again, it is identical as in the corpora list (2.2.1) – first select a filter and then select the filter value.
- Status attributes of SMT systems explained:
    - "Not Started" – the SMT system is trained, but it is not started. It is not possible to use it for translation. In order to start translation, you must start the system;
    - "Running" – the SMT system is trained, started and running. It is possible to use it for training tasks;
    - "Error" – the SMT system is not trained. An error occurred in training process;
    - "Not trained" – the SMT system definition is created but not trained. It is not possible to use it for translation;
    - "Training" – the SMT system training process is in progress;
    - "Starting" – the SMT system starting process is in progress.

2.3.2 Create a new SMT system

Click Create system in the upper right corner of the SMT system list page.

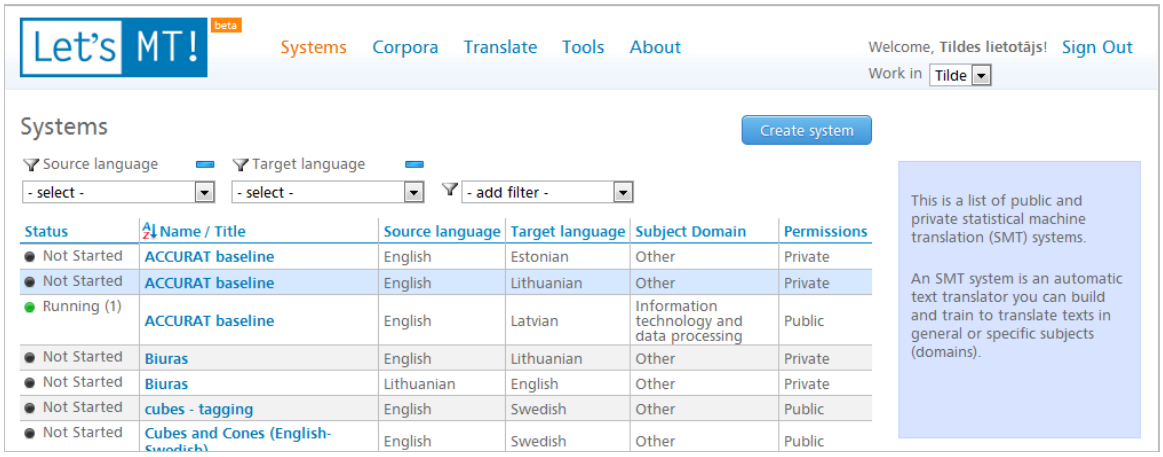


Figure 13. Create new SMT system

The creation of LetsMT! SMT system has been simplified to few easy steps. User must complete the following steps (the minimum required to create a SMT system):

1. Select target and source languages, provide a SMT system name. Step: System properties,
2. Select parallel corpora,
3. Select monolingual corpora,
4. Provide advanced SMT system training options, if needed.

In the next few paragraphs, each step is described in more detail (with screenshots).

System properties

- Fill the required metadata fields and, optionally, add some more metadata fields. In this step, source and target languages and the name of the SMT system must be specified;
- Additional metadata fields can be added by clicking the -add field- combo-box. In this combo-box, all available SMT system metadata fields will be available for selection;
- Manually added metadata fields can be removed by clicking the Remove hyperlink at the upper right corner of the metadata field.

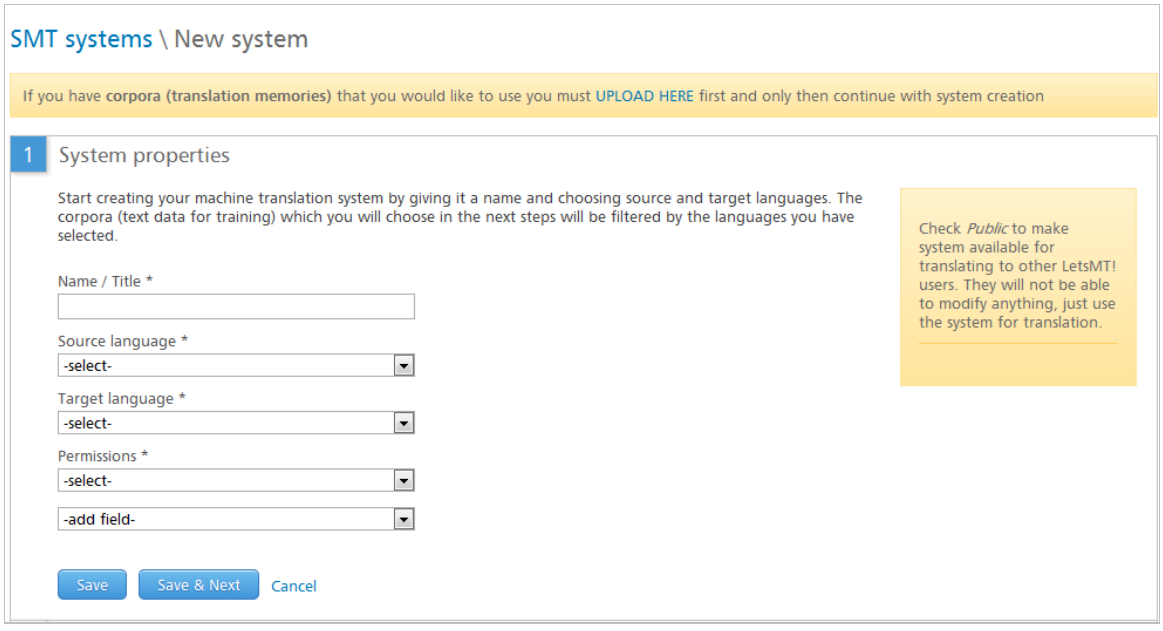


Figure 14. Step "System properties"

Parallel corpora

- Select parallel corpora to use for system training. By default, a list of corpora filtered to match the selected (in the previous step) source and target languages is shown.

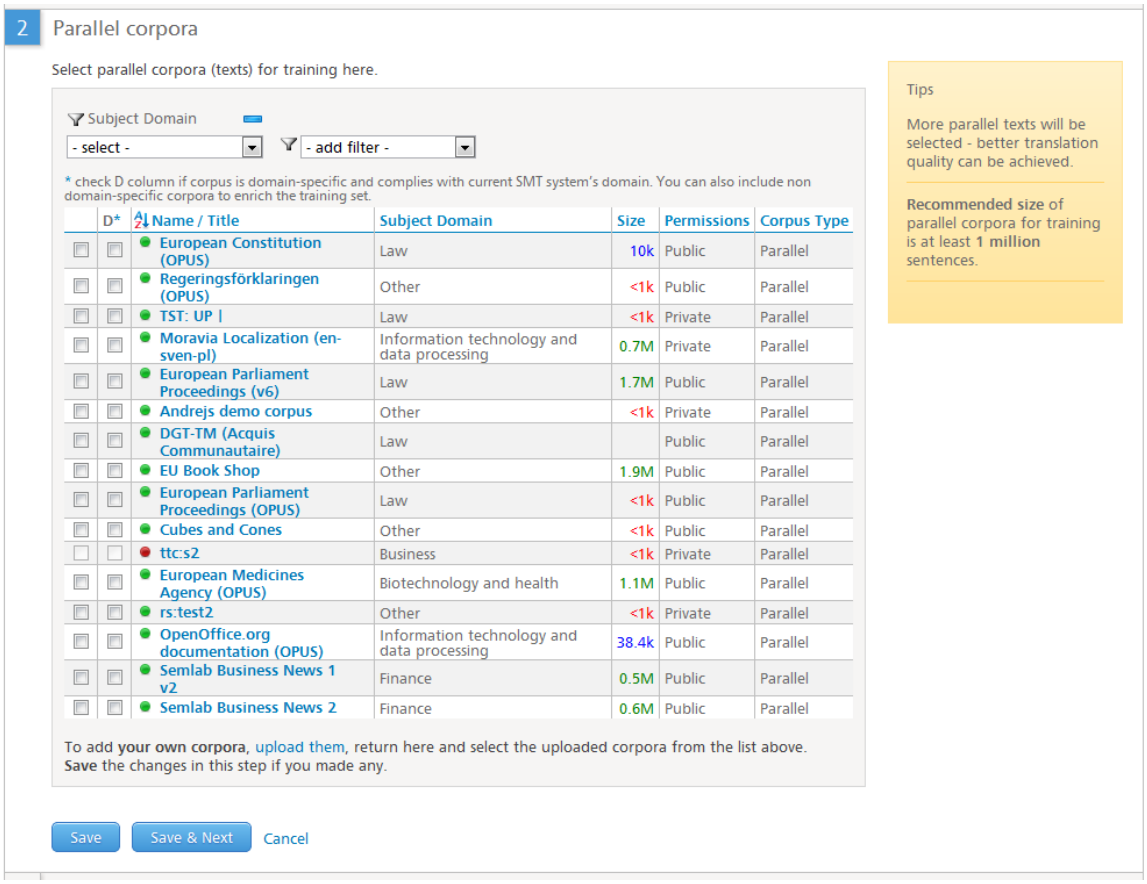


Figure 15 Step „Parallel corpora"

Monolingual corpora

- Select monolingual corpora to use for system training.

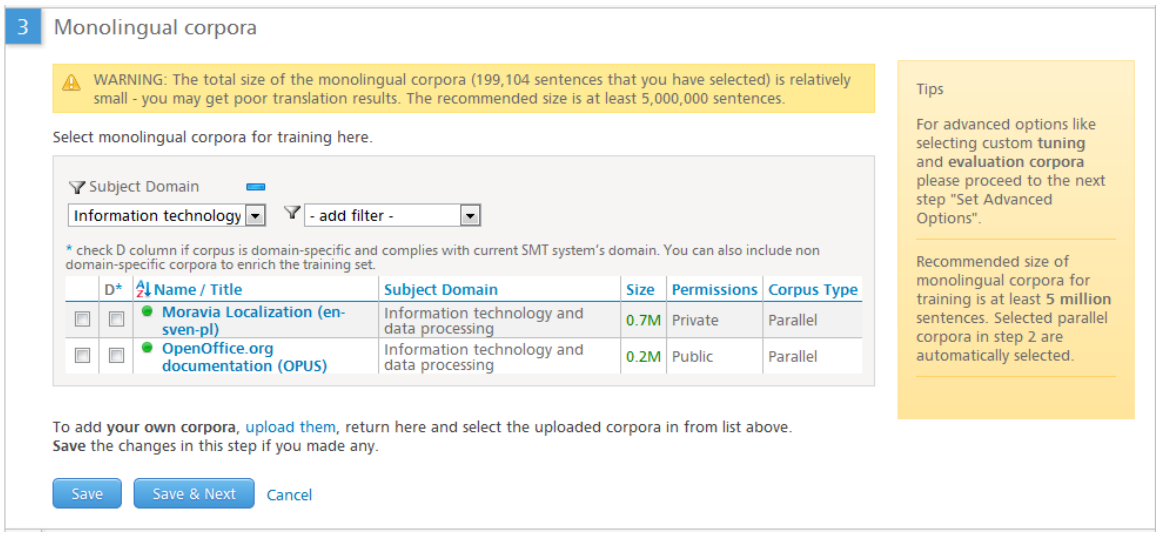


Figure 16 Step „Monolingual corpora"

Advanced options (optional)

- In this step, you can specify advanced training options. The following advanced options are available for selection: *Tune SMT system*, *Add custom content set* and *Use custom evaluation step*. Each option is explicitly explained in the UI (see Figure 17. Step „Advanced options").

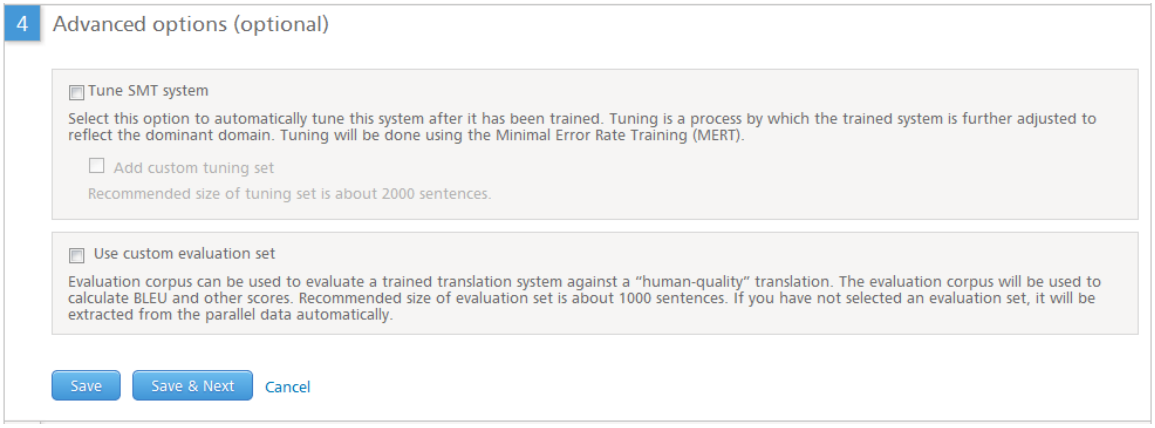


Figure 17. Step „Advanced options"

2.3.3 Train SMT system

In the SMT systems list, click the system name and then click Details.

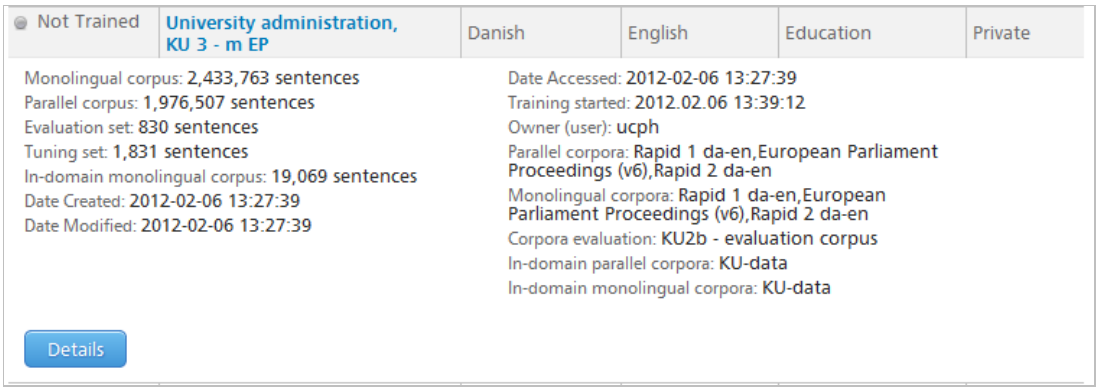


Figure 18. Short summary about SMT system

In system details view, click the Training step.

Figure 19 Training step of system details view

Then click the Start training button.

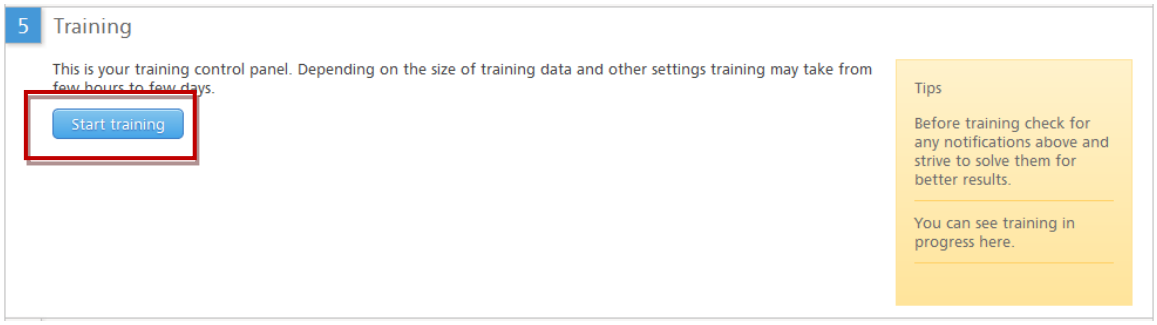


Figure 20 Start training of SMT system

System training is a time- and resources-consuming task. The process can take from some hours to several days. Once training has started, the status of training, completed and remaining steps can be tracked by clicking the View chart.

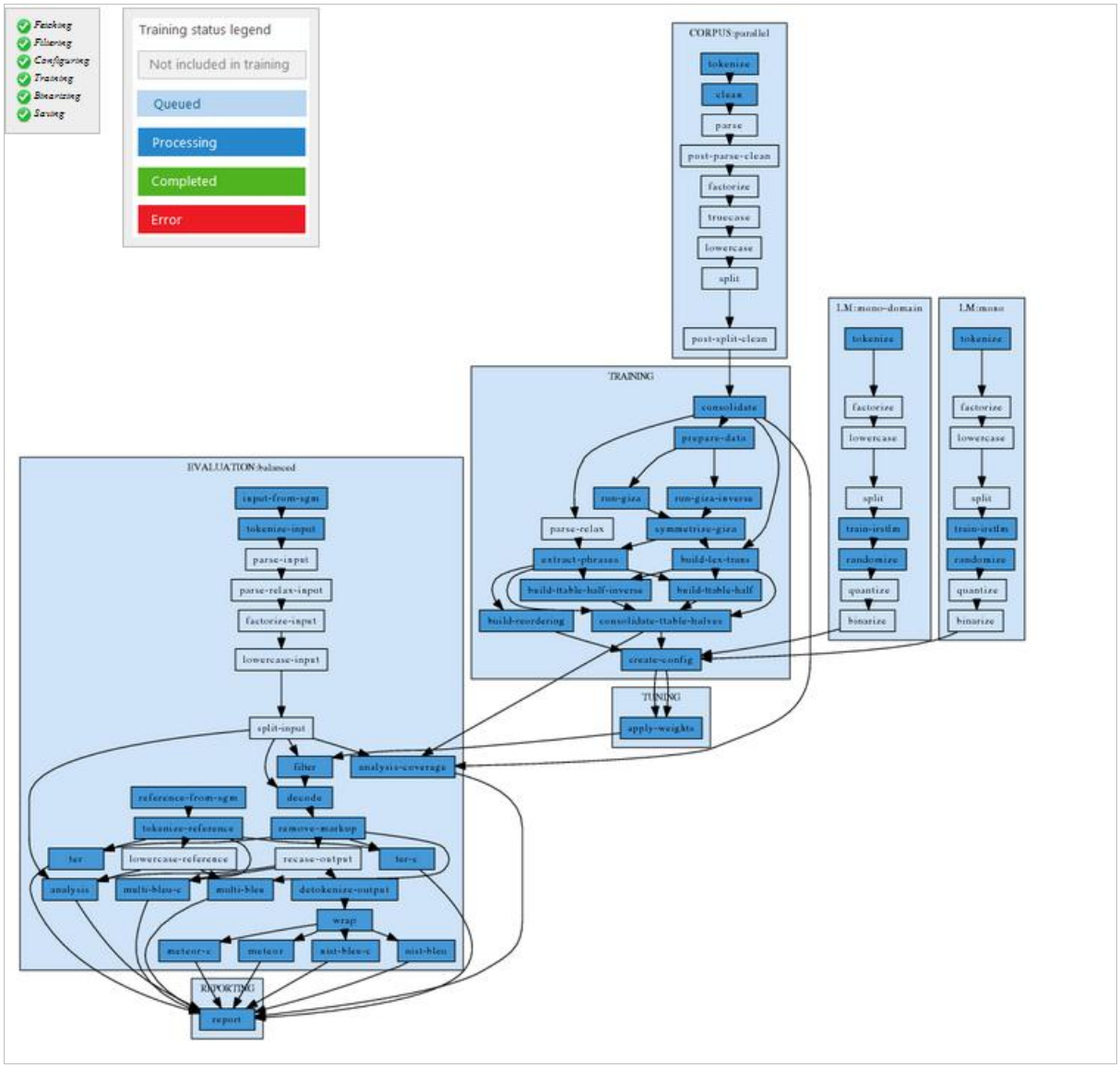


Figure 21. SMT system training steps

2.3.4 View information about SMT system

In the SMT systems list, click the system name. A short summary view of the SMT system is displayed.

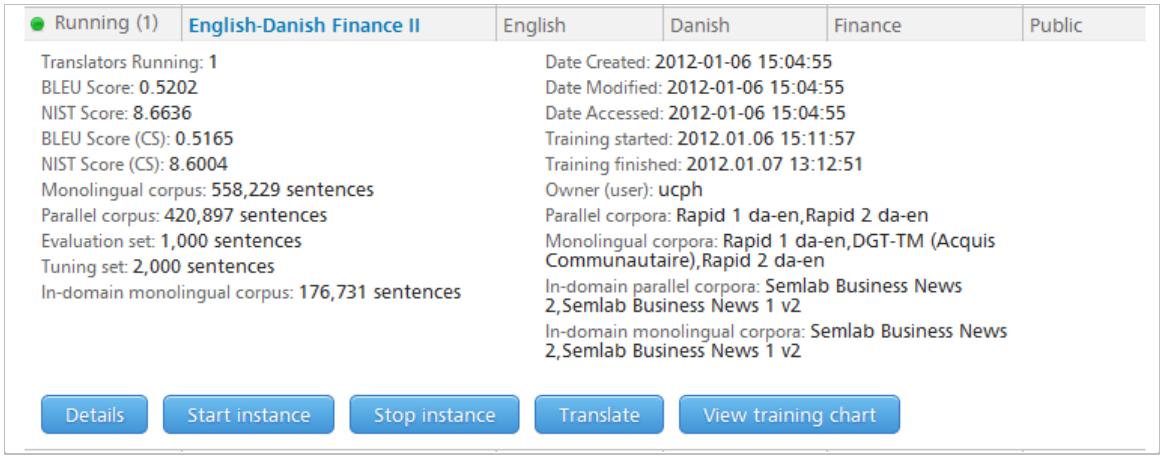


Figure 22. SMT system details - summary

In order to get a more detailed overview about the SMT system, click Details.

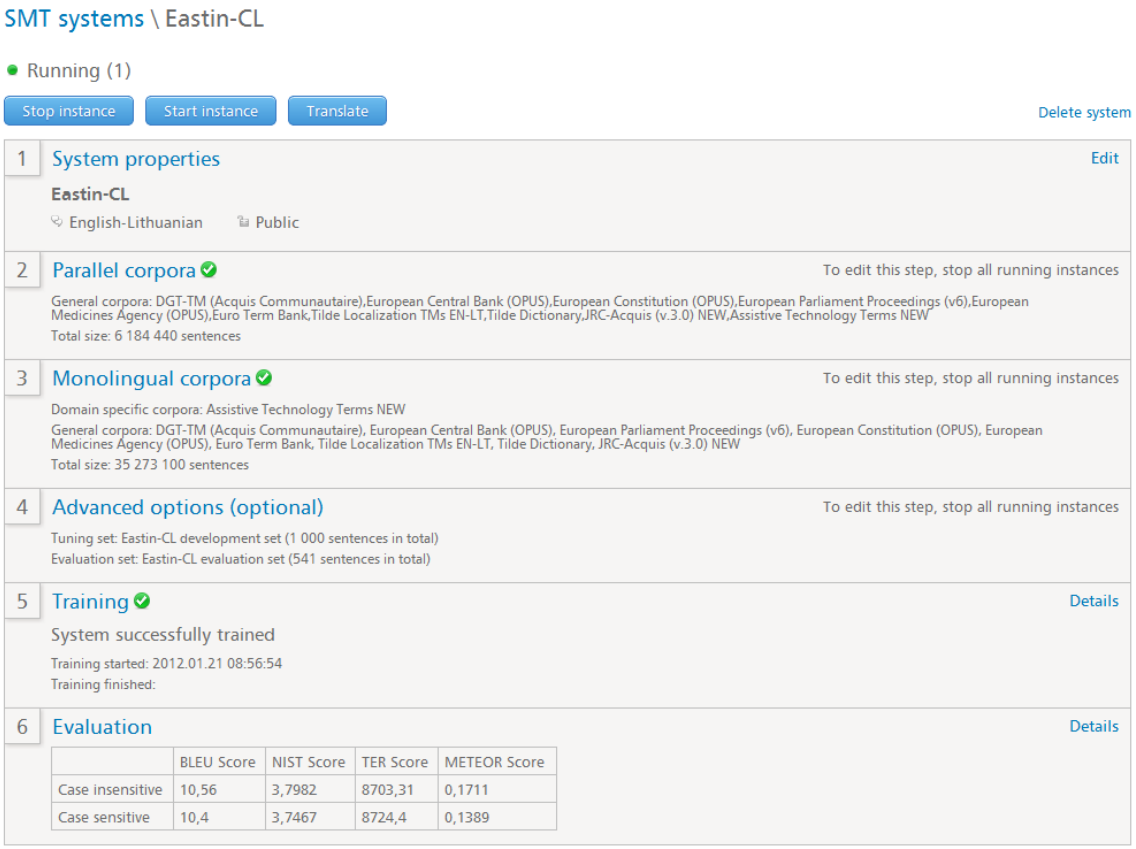


Figure 23. SMT system details

2.3.5 Edit/Delete SMT system

In the SMT systems list, click the system name and then click Details. This will display the system overview page.

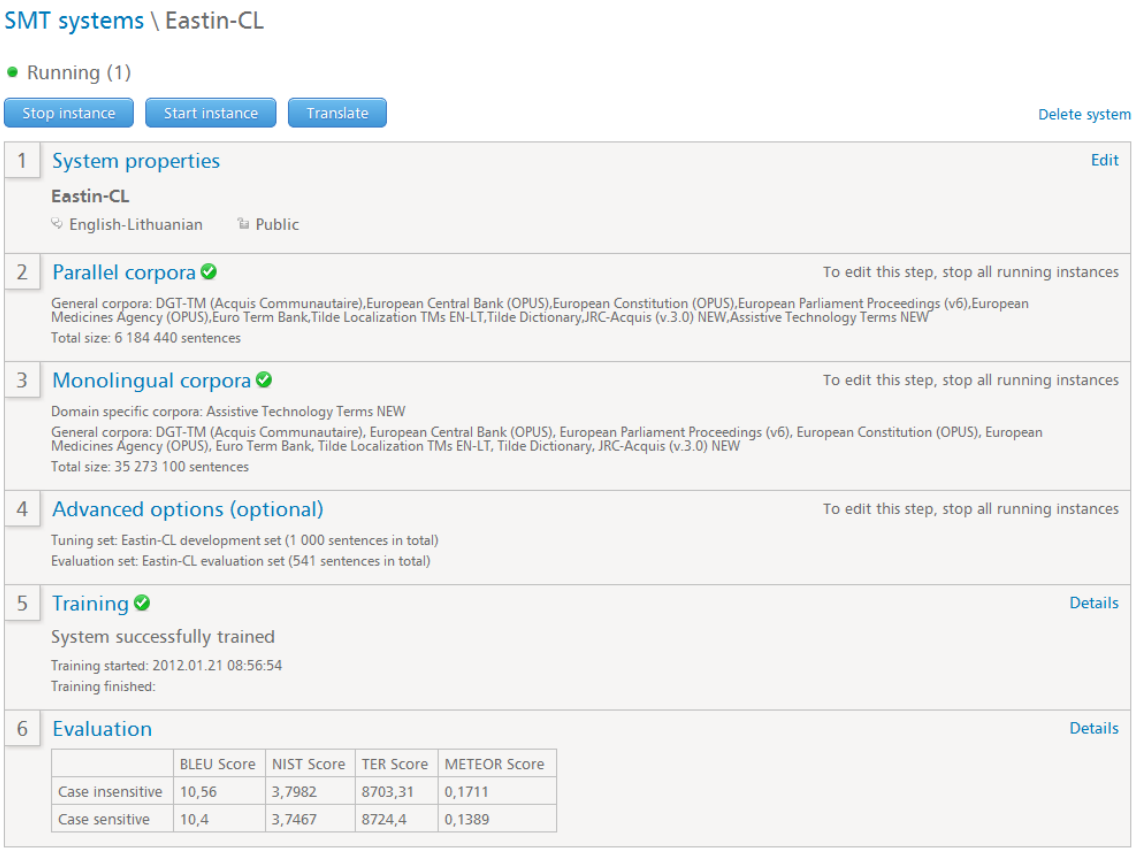


Figure 24. SMT system overview page

In order to edit a SMT system, click the Edit link in any section of the system overview, for example, System properties. Edit data and click Save.

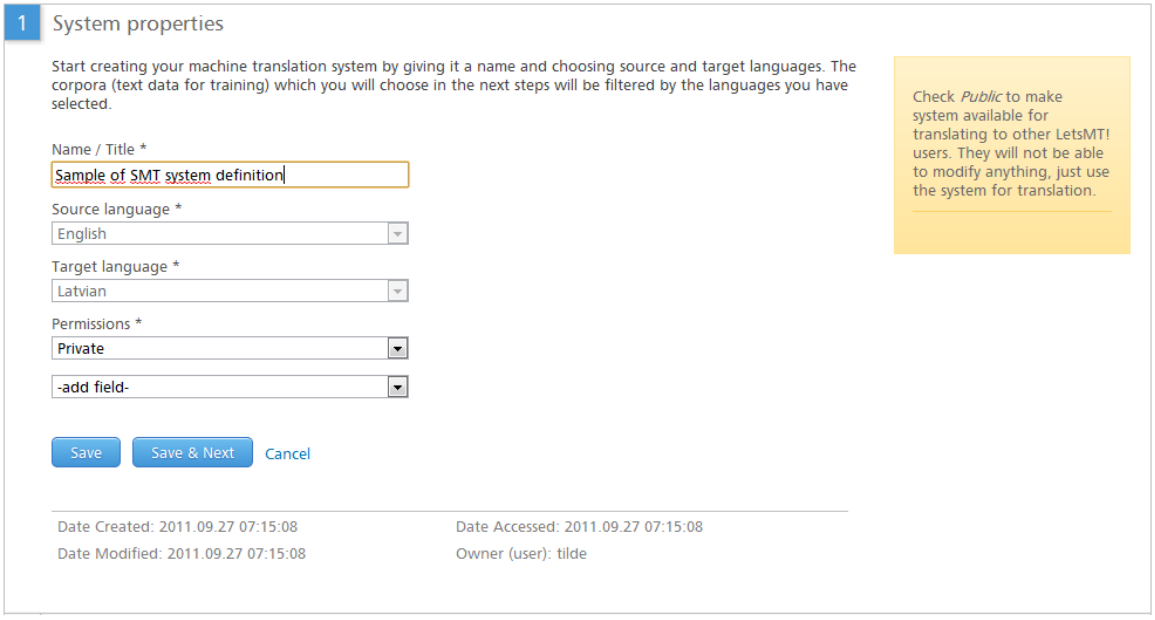


Figure 25. Edit SMT system

In order to delete a SMT system, click the Delete system link and confirm the delete action.

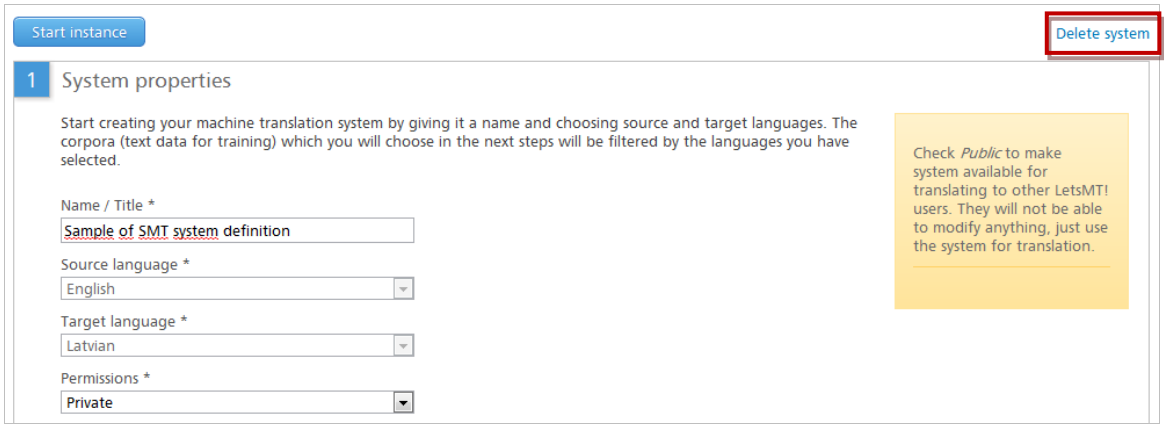


Figure 26. Delete SMT system

2.4 Translate on the public LetsMT! website

Click Translate at the top of the LetsMT! webpage. In the System combo-box, select one of the running SMT systems, enter your text and click Translate.

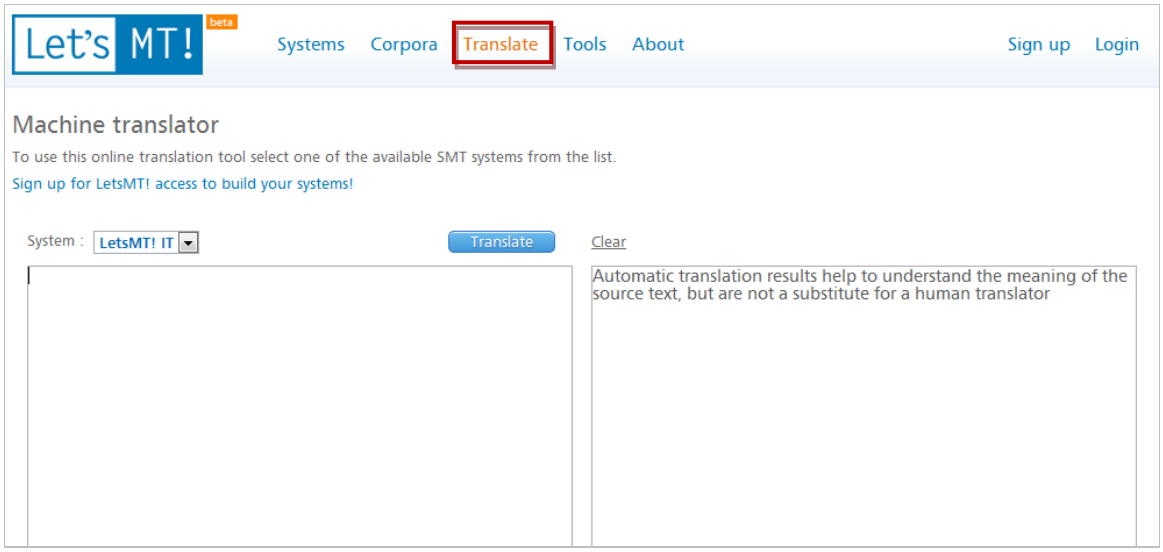


Figure 27. LetsMT! translation webpage

2.5 Translate using a widget, browser and CAT plug-ins

To access the LetsMT! platform widgets, plug-ins, API descriptions, click Tools at the top of the LetsMT! webpage.

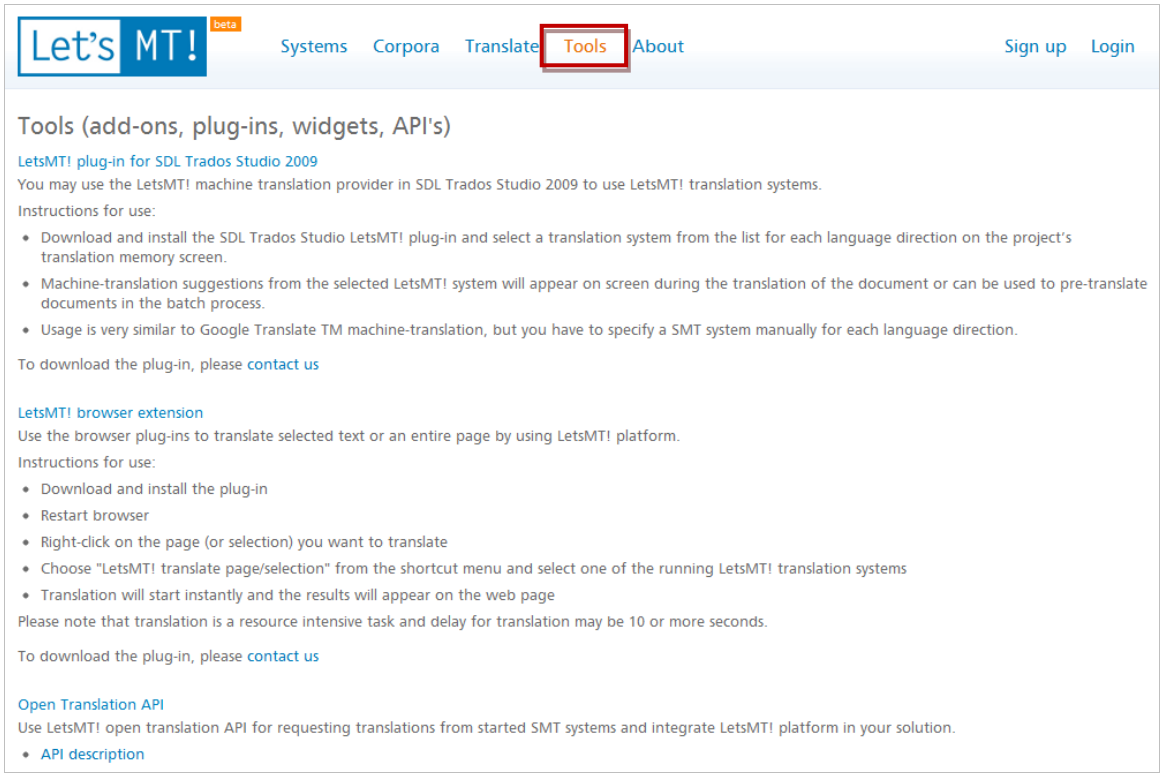


Figure 28. Integration webpage

On the displayed webpage, you can find plug-ins, widgets, API descriptions as well as short operation instructions.

2.6 Functionality for System Administrators

LetsMT! users with a System Administrator role have additional functionality available from the LetsMT! user interface (UI) (see Figure 29. Administrators menu).

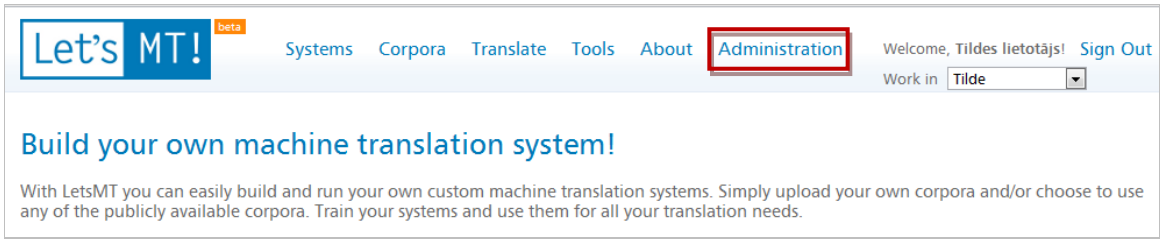


Figure 29. Administrators menu

The main features for System Administrators are:

- Users and User Group Management. System administrators can: 1) create new LetsMT! users and user groups, 2) edit information about existing users and user groups, 3) delete users or groups, 4) change user roles and user groups.
- LetsMT! infrastructure management. System administrators can: 1) view information about running tasks and resource usage on LetsMT! infrastructure, 2) scale up or scale down LetsMT! infrastructure by adding new or removing High Performance Cluster (HPC) grid nodes.

3 Conclusions and next steps

The LetsMT! platform has been specified, designed, developed and deployed according to the DoW. The current version of the platform is available at the following URL: http://letsmt.eu. The LetsMT! platform supports all key user tasks: upload of SMT training data, building of custom SMT systems, and use of systems for the translation.

The LetsMT! platform has been improved significantly compared with the first publicly available beta version (released on M17). The main areas of improvements are the following: 1) usability, 2) SMT system training performance and stability, 3) training data upload and sharing. As the result, the latest version of LetsMT! platform is a fully functional and globally available software service.