D5.4 Addendum: Final release of the CASMACAT workbench

Mercedes García-Martínez, Michael Carl, Bartolomé Mesa-Lao, Vicent Alabau, Daniel Ortíz-Martínez, Philipp Koehn

Distribution: Public

CASMACAT
Cognitive Analysis and Statistical Methods
for Advanced Computer Aided Translation

ICT Project 287576 Deliverable D5.4 Addendum

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 287576.

Project ref no.	ICT-287576
Project acronym	CASMACAT
Project full title	Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation
Instrument	STREP
Thematic Priority	ICT-2011.4.2 Language Technologies
Start date / duration	01 November 2011 / 36 Months

Distribution	Public
Contractual date of delivery	May 1st, 2014
Actual date of delivery	November 4, 2014
Date of last update	November 4, 2014
Deliverable number	D5.4 Addendum
Deliverable title	Final release of the CASMACAT workbench
Type	Report
Status & version	Draft
Number of pages	13
Contributing WP(s)	WP5
WP / Task responsible	UEDIN, CBS, UPV
Other contributors	
Internal reviewer	Philipp Koehn
Author(s)	Mercedes García-Martínez, Michael Carl, Bartolomé Mesa-Lao, Vicent Alabau, Daniel Ortíz-Martínez, Philipp Koehn
EC project officer	Aleksandra Wesolowska
Keywords	

The partners in CASMACAT are:

University of Edinburgh (UEDIN)
Copenhagen Business School (CBS)
Universitat Politècnica de València (UPVLC)
Celer Soluciones (CS)

For copies of reports, updates on project activities and other CASMACAT related information, contact:

The CASMACAT Project Co-ordinator
Philipp Koehn, University of Edinburgh
10 Crichton Street, Edinburgh, EH8 9AB, United Kingdom
pkoehn@inf.ed.ac.uk
Phone +44 (131) 650-8287 - Fax +44 (131) 650-6626

Copies of reports and other material can also be accessed via the project's homepage: http://www.casmacat.eu/

Executive Summary

This document is an extension of D5.4 as suggested in the second review report. It contains details about the implementation of the final prototype of the CASMACAT workbench and outlines the improvements of the workbench with respect of the previous deliverable 5.4.

The objective of WP5 is to integrate the translation system and user interface and to develop the CASMACAT workbench. This deliverable shows the functional components of the workbench and describes their interaction possibilities in the last CASMACAT prototype. It also describes the most recent additions to the workbench.

Contents

1 CASMACAT configurations for different user profiles

Figure 1 shows various functional properties of the CASMACAT workbench.

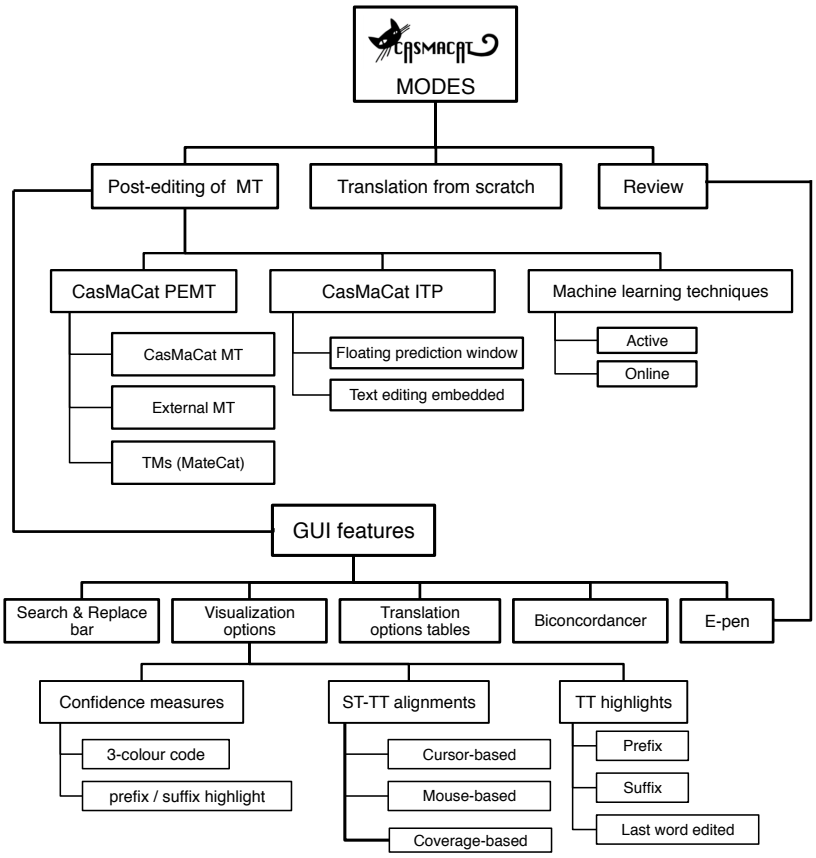


Figure 1: Different options of configuration in CASMACAT

In CASMACAT there are three general configurations that can have different modes, they are explained below:

- Post-editing MT output (PEMT):
  - Conventional PEMT:
    1. PEMT and translation memory (TM) via the MATECAT server (see D5.3 section 1.1.2).
    2. PEMT via the CASMACAT MT server (see deliverables D5.1, D5.2, D5.3, D5.4)
    3. PEMT from external MT output (see section 4).
  - Interactive Translation Prediction (ITP):
    1. inline prediction (see D5.3 section 1.1.3).
    2. floating prediction (see D5.4 section 1.3).
- Translation from scratch (see section 2).
- Translation review mode (see section 3).

These configurations can have different features active or not. These features are listed below:

- E-pen: hand-writing recognition using an electronic pen. Tested in the review mode (see D5.4 section 2).
- Learning translation modifications at runtime:
  - Online learning (see D4.3 section 1).
  - Active learning (see D4.3 section 2).
- Biconcordancer (see D5.4 section 1.4).
- Translation options (see D5.4 section 1.5).
- Visualization of features (see D5.4 section 1.1).
  - Alignment visualization:
    * Aligned words under the mouse cursor.
    * Aligned words where the caret is placed.
    * Aligned words that have been translated already (Section 5).
  - Confidence scores (color: red, orange, black, light for inconfident suffix).
  - Highlighting of:
    * edited prefix.
    * suffix to be edited.
    * last word edited.
- Search and replace (see D5.3 section 1.1.3).

1.1 Configuration of the links to translate a file in CASMACAT

The XLIFF files to be translate them in CASMACAT can be upload them directly to the tool with the desired configuration using a script. The user will be provided with a link to the go directly to the translation view of the file with all the configuration already set. The link provided will have this structure:

$translate/project\_name/source\_language{-}target\_language/id\_job{-}password/conf\_mode@server$

Where conf_mode will have the name of a javascript file with all the settings of the configuration and features initialized as the user desired.

2 Translation from scratch mode

It is possible to translate from scratch in CASMACAT. For that purpose, a new configuration has been implemented that you can activate. In this mode, the machine translation is not shown, so the user will see the target window in blank to write his translation.

3 Review mode

A new mode of using CASMACAT tool called *review* has been integrated. This mode is used for the reviewers of the translations. In the professional translation companies is necessary to proof read the translations to make sure that are good enough for the client.

A review mode has been implemented in CASMACAT tool. In order to do a review using CASMACAT the XLIFF file to be translated has to contain the target label filled with the translation to be reviewed. CASMACAT will show the translation view filling the target text with the translation to be reviewed. The user will translate normally with the post-edition mode and optionally, using the e-pen. The logging has been extended to keep the initial translation in the database and to initialize the final translation for reviewers.

Figure 2: Coverage-based word alignment: shading off translated source words and highlighting the next source word to be translated.

4 Post-editing external MT output

The user is able to post-edit a translation taken from an external MT server. This mode works similar to the review mode. The translation from the external MT has to be paste in the target label of the XLIFF file. After uploading this XLIFF file, the user will be able to post-edit the translation in the normal translation view of CASMACAT.

5 Coverage-based word alignment visualization

We added another way to use word alignment information to guide the translator in interactive translation prediction mode: by shading off source words that have already been translated, the translators focus is drawn to untranslated words. See Figure 2 for a screenshot of this visualization option.

Note that in addition to translated words that are shaded, the next word expected to be translated according to the interactive translation prediction is highlighted in a light shade of orange.

6 E-pen logging

CASMACAT provides a mode for hand-written recognition using an e-pen, more details were explained in deliverable 5.4 section 2 (T5.3 E-pen Interaction). The logging function in CASMACAT has been extended with new information about e-pen mode. The next items related with the HTR mode have been added to the logging information:

1. htrResult: informs about the result of the hand writing recognition.
2. htrUpdate: is the partial recognition meanwhile you are writing, it is not the final recognition.
3. htrTextChange: is executed when the target text changes due to the recognitino of hand writing or a gesture.
4. htrNBestClick: this event is shown when a word is selected from the list of words suggested.
5. htrGesture: is executed when a gesture is recognized.
6. htrStart: informs about when a recognition starts.
7. htrAddStroke: keeps a number of coordinates when the electronic pen is lifted.
8. htrEnd: informs about when the user has finished writing.

Besides these new items, the logging has been extended adding the new option of edition called *epen* in the text event. The text event with the attribute edition epen is obtained when the target text changes due to the hand writing recognition.

7 Upload function

A function to upload the data from a log file coming from a translation file process to the CASMACAT database has been implemented in the tool. All the information in the log file and different events that were produced during the translation process are uploaded again to the CASMACAT database.

In order to upload a log file to the database, it has to be copied in the directory *uploads* and run a script with the name of the file as a parameter. It is not possible to upload it directly from the interface due to the big size of the log files. It is possible to list the new files that have been uploaded from the list of documents view (more information in deliverable 5.3 section 1.1.5), it is also possible to translate or review them again using CASMACAT tool.

It was already possible to download a log file from the database of CASMACAT. The different advantages of having an upload function can be listed as:

- A file translated in a different server can be merged to a new server.
- The user can modify the logging of a file from the log file in a easier way and add desired options and he is able to upload it again.
- An improved log file can be uploaded to the database and translate or replay it again. In the replay mode it is possible to fix the gaze to word and it is possible to download the log file again.

8 Infusion of InputLog Data

Inputlog [2] is a windows-based logging tool that logs all types of input modes: keyboard, mouse & speech recognition. In contrast to CASMACAT, Inputlog is not aplication dependent. That is, it can log keyboard activities independent of the application which receives the input. Inputlog knows which application is on focus, and stores this information together with the actual key pressed and the time stamp of a keystroke (or mouse movement) in its log file. For instance in a browser application, Inputlog knows which website is on focus and associates keystrokes to the current website in its IDFX log files. In this way web searches can be tracked and reconstructed. On the one hand, InputLog is thus more universally deployable, in different windows-based applications. On the other hand, Inputlog has no possibility to know where on the screen and where in a text the typed characters occur. That is, for tracking post-editing behaviour, Inputlog would produce insufficient information. From Inputlog we know which keystrokes were pressed, but not necessarily which characters are typed (for instance with a Chinese or Hindi input device) or which characters are deleted and we also do not know where in a text these operations would take place. However, Inputlog is a valuable complement to CASMACAT, as post-editors often resort to external resources (such as google seach, linguee, or online dictionaries) and in such cases we can trace the searches produced.

We conducted a set of experiments in which both, CASMACAT and Inputlog run in parallel. A tool was deviced (InfuseIDFX.pl) that integrates both log files after the sessions are recorded. The 'InfuseIDFX.pl' script first synchronizes the common keystrokes in the CASMACAT and in the Inputlog log (IDFX) files. In a second step, Inputlog keystrokes that are produced outside the CASMACAT GUI are 'infused' into the CASMACAT log file. The script is part of the TPR-DB and can be downloaded from the TPR-DB website,`http://bridge.cbs.dk/platform/?q=CRITT_TPR-db`, and more specifically in the bin folder under `https://130.226.34.13/svn/tpr-db/`.

9 Extension of the TPR-DB

Work on the TPR-DB in the context of the TDA summer workshop resulted in a number of extensions to the TPR-DB. These include handling of the external features that are now infused by the new InputLog script into the CASMACAT logging file, and additional features in the already existing tables. The features will be further discussed in D1.4

10 Home Edition

We developed a version of the CASMACAT Workbench that can be installed by translators on their home computers. This so-called CASMACAT Home Edition includes all the software integrated into the workbench and successfully tested in field trials or external field trials. The Home Edition also integrates the Moses and Thot training pipelines that allow training of machine translation systems on the translators' own data.

10.1 Installation

Given that CASMACAT has been developed as a web application with a back end running on Linux workstations, there is a challenge to create a version of the workbench that can be run on typical home computers that run Windows or maybe MacOS. We addressed the challenge by packaging up the software for installation in a Virtual Machine with a web-based administrative interface.

The full installation instructions are on the CASMACAT web site[1]. The required steps are:

- installation of free software to run virtual machines on the computer (Virtualbox[2]
- installation of a free open source Linux distribution (Ubuntu[3]).
- download and execution of an installation script that sets up all the required software and dependencies

Figure 3 displays a partial screenshot of the installation progress. The screenshot includes a glimpse of the administrative interface viewed on a browser running within the virtual machine. The administrative interface is also accessible directly from the translator's computer (technically called the "host machine"). This where all interactions with the Home Edition installation will take place after installation.

The installation of the CASMACAT Home Edition has been tested on Windows, MacOS, and Linux computers.

10.2 Training of Customized Machine Translation Systems

The basic installation of the CASMACAT Home Edition ships with a very simple French–English machine translation system, which purely exists for demonstration purposes. We expect that a typical user of the CASMACAT Home Edition will want to build a customized machine translation system optimized for a given translation task.

[1] `http://www.casmacat.eu/index.php?n=Installation.HomePage`
[2] `https://www.virtualbox.org/wiki/Downloads`
[3] `http://www.ubuntu.com/download/desktop`

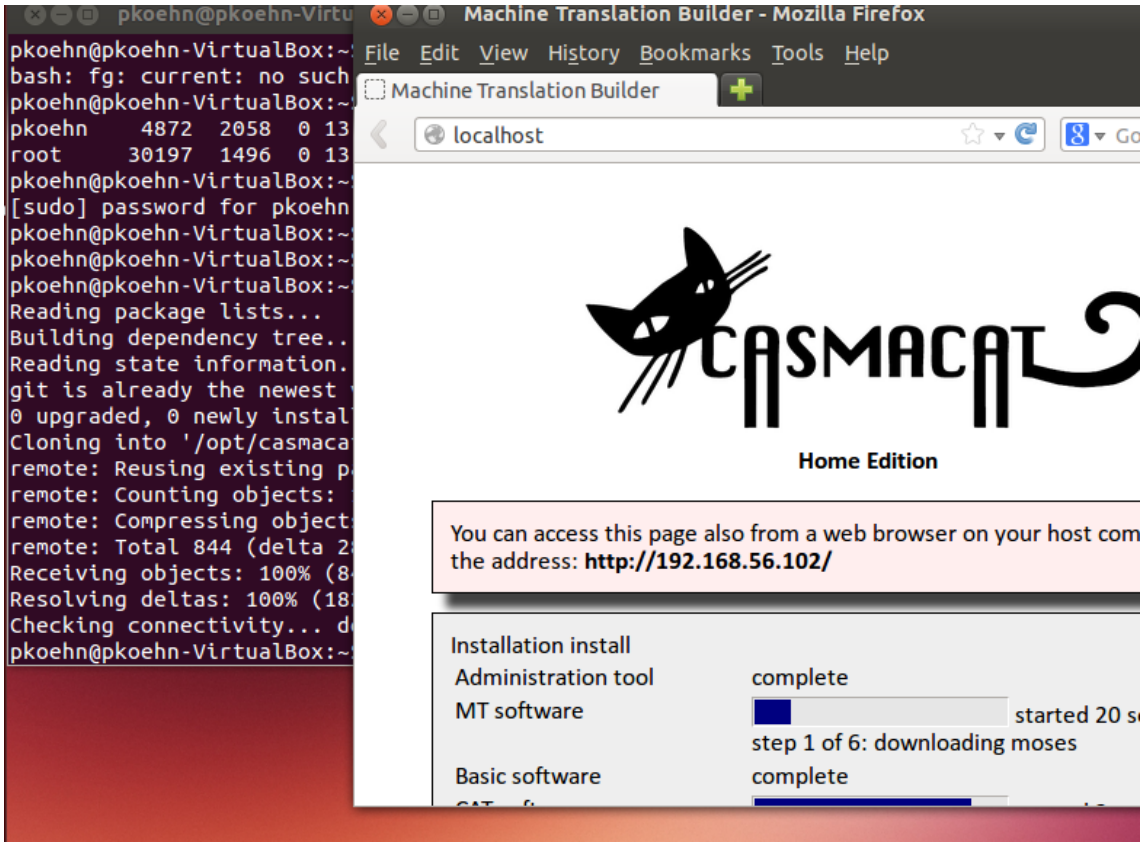


Figure 3: Installation of the CASMACAT Home Edition in a virtual machine.

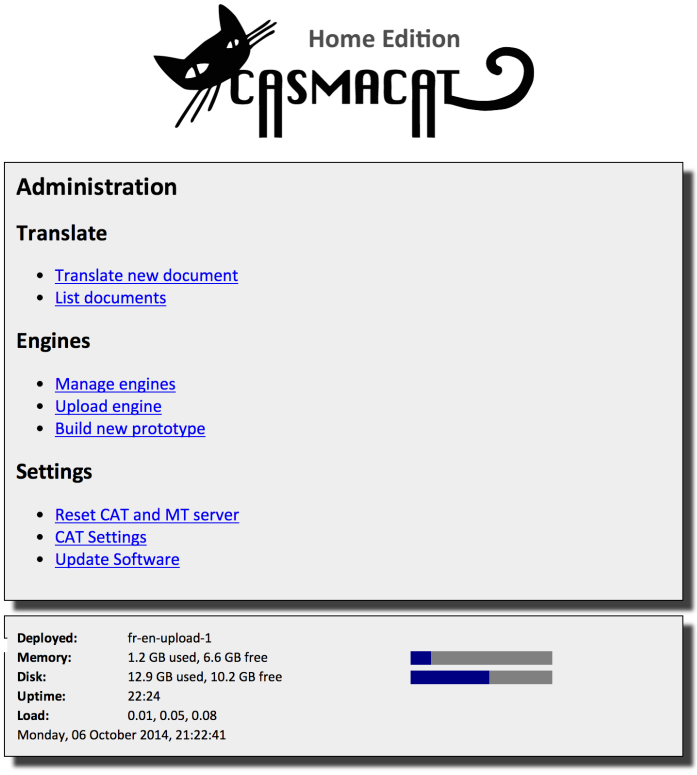


Figure 4: Main menu of the CASMACAT Home Edition administrative interface, viewed through a web browser.

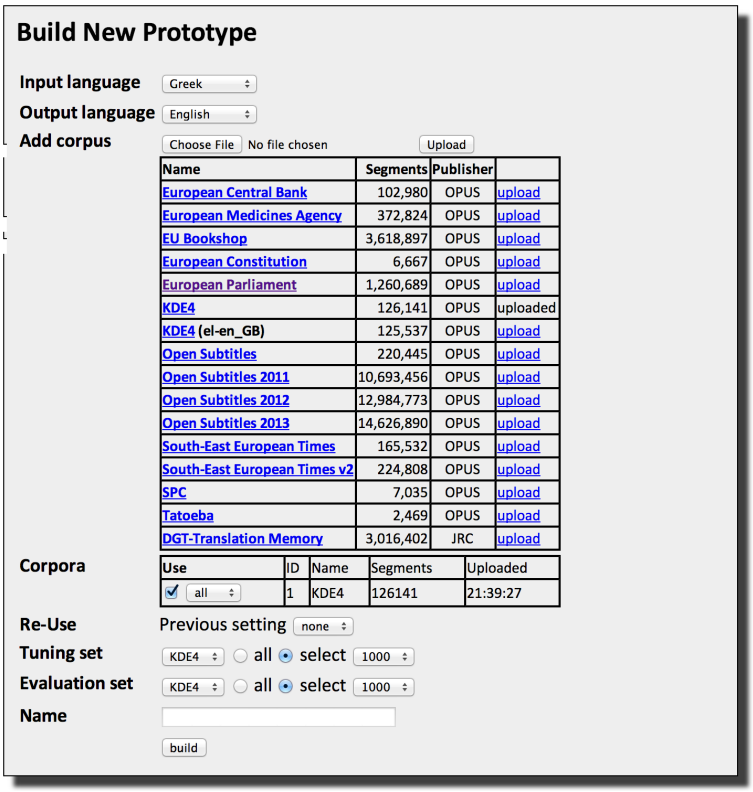


Figure 5: Specifying a setup for training a new machine translation system.

10.2.1 Training Data

Due to established use of translation memory system to aid translators, the typical translator will have accumulated a historic collection of translated text — which is exactly what is needed to train a machine translation system. We expect that this data is available in the form of XLIFF (the standard format of the CASMACAT workbench).

However, such personal translation memory data may be insufficient to train a machine translation system of respectable quality. In this case, the translator may want to use additional publicly available corpora. We integrated into the workbench a web service that queries public repositories of such data, and currently connected it with the OPUS project[4] and a repository hosted at the CASMACAT web site. For many European languages a diversity of training data is thus accessible.

10.2.2 Settings for Training

Once the translator selects training data, also development sets and test sets have to be chosen. These may be subsets of the data selected for training (which obviously will be excluded from the translation model training).

In a basic use case, the translator sets the language pair, uploads her translation memory, and accepts all defaults for training. See Figure 5 for a screen shot of the administrative interface view for model training. In this case, the translator sets out to build a Greek–English system on KDE (open source software) data.

[4] `http://opus.lingfil.uu.se/`

Training will take several hours and maybe more than a day, depending on the size of the corpus. It uses the Experimental Management System (EMS)[5] [1] developed as part of the Moses statistical machine translation toolkit for manage the training process. EMS has been extended to support training of Thot machine translation systems. Training process is reported in the status window of the administrative interface and can be further inspected using the web interface to EMS.

Note that training can potentially use all features available in the Moses or Thot machine translation systems. The web interface currently offers only basic but reasonable settings. It would be very straightforward to give more options in future versions of the CASMACAT Home Edition.

10.3 Managing Machine Translation Engines

In the terminology of the CASMACAT Home Edition, machine translation model training results in a *prototype*. Technically, this is a training run based on a given configuration specification, resulting in a collection of machine translation model components that was tuned and tested on specified test sets. Its performance is measure on a test set. The translator may chose at any time to change the training conditions and build another prototype, which may share components with a prior prototypes. For instance, if a different tuning set is chosen, then the language model and translation model will be re-used, but the model weights will be changed.

Once, the user is satisfied with the test performance of the built prototype, it can be converted into a machine translation *engine*. An engine — according to our definition — is the set of all relevant model files and settings in a self-contained package. Engines can be downloaded from the CASMACAT Home Edition, shared with other translators, who can upload it into their CASMACAT Home Edition installation.

Figure 6 is a screen shot of the administrative interface view that allows the management of machine translation engines. Here, several prototypes have been build for various language pairs (English–French, English–Spanish, French–English, Spanish–English). Some of the prototypes have been converted into engines. One of the engines, here the French–English "(x1) Toy" engine has been selected for deployment, meaning that it runs on the backend of the CASMACAT workbench.

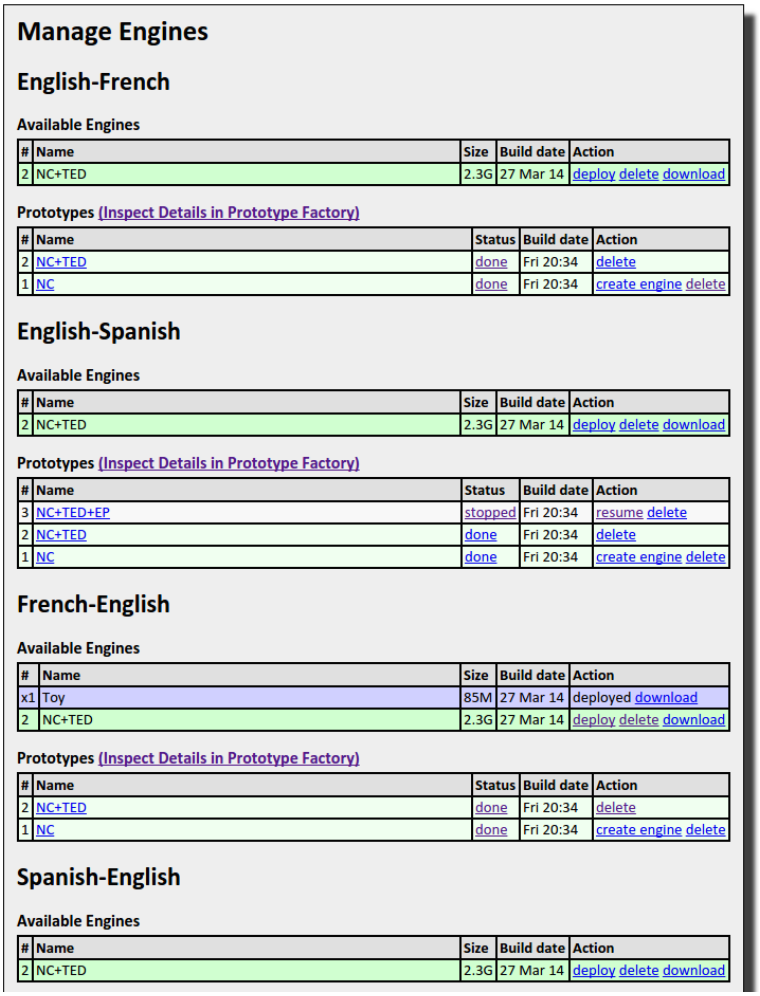
The administrative interface allows the deployment of any available engine, the creation of engines from prototypes, the deletion of engines or prototypes, and the download of engines. Ongoing training runs can be interrupted or resumed. A link to "Details in Prototype Factory" connects to the web interface of the Experimental Management System.

10.4 Using the Workbench

There is nothing surprising about using the workbench within the CASMACAT Home Edition. Once an engine has been selected, the translator can upload a document to be translated and access the usual translate view of the workbench. All data is stored locally in the virtual machine on the translators home computer.

The translator may set the specific functionality of the workbench under "CAT Settings" (Figure 7).

[5] t http://www.statmt.org/moses/?n=FactoredTraining.EMS



Manage Engines

English-French

Available Engines

#	Name	Size	Build date	Action
2	NC+TED	2.3G	27 Mar 14	deploy delete download

Prototypes (Inspect Details in Prototype Factory)

#	Name	Status	Build date	Action
2	NC+TED	done	Fri 20:34	delete
1	NC	done	Fri 20:34	create engine delete

English-Spanish

Available Engines

#	Name	Size	Build date	Action
2	NC+TED	2.3G	27 Mar 14	deploy delete download

Prototypes (Inspect Details in Prototype Factory)

#	Name	Status	Build date	Action
3	NC+TED+EP	stopped	Fri 20:34	resume delete
2	NC+TED	done	Fri 20:34	delete
1	NC	done	Fri 20:34	create engine delete

French-English

Available Engines

#	Name	Size	Build date	Action
x1	Toy	85M	27 Mar 14	deployed download
2	NC+TED	2.3G	27 Mar 14	deploy delete download

Prototypes (Inspect Details in Prototype Factory)

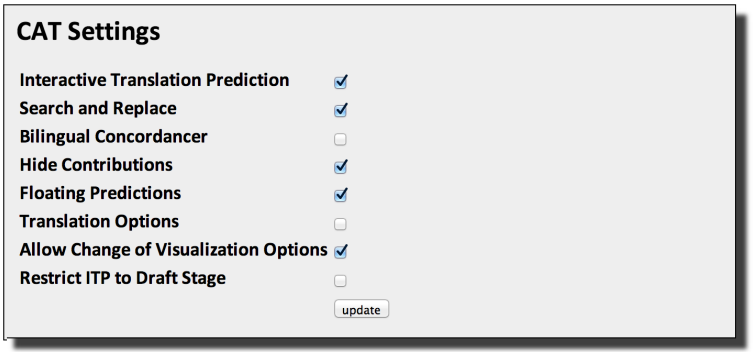
#	Name	Status	Build date	Action
2	NC+TED	done	Fri 20:34	delete
1	NC	done	Fri 20:34	create engine delete

Spanish-English

Available Engines

#	Name	Size	Build date	Action
2	NC+TED	2.3G	27 Mar 14	deploy delete download

Figure 6: Managing prototypes and engines in the CASMACAT Home Edition.



CAT Settings

- Interactive Translation Prediction ☑
- Search and Replace ☑
- Bilingual Concordancer ☐
- Hide Contributions ☑
- Floating Predictions ☑
- Translation Options ☐
- Allow Change of Visualization Options ☑
- Restrict ITP to Draft Stage ☐

update

Figure 7: Select functionality of the CASMACAT Home Edition.

References

[1] Philipp Koehn. An experimental management system. *The Prague Bulletin of Mathematical Linguistics*, 94:87–96, 2010.

[2] Leijten, M., & Van Waes, L. Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes. *Written Communication*, 30(3):358392, 2013.