



D4.2.2: REPORT ON ONLINE MT SYSTEM

**Pablo Nieto Caride, Giuseppe Deriard Nolasco, Pedro L. Díez Orzas,
Félix Fernández, Consuelo Aldana, Pablo Badía (Linguaserve)
Ankit Srivastava, Declan Groves (Dublin City University)
Thomas Ruedesheim (Lucy Software)
Román Díez, Alberto Crespo (Spanish Tax Office)**

Distribution: Public

MultilingualWeb-LT (LT-Web)
Language Technology in the Web

FP7-ICT-2011-7

Project no: 287815

Document Information

Deliverable number:	4.2.2
Deliverable title:	Report on Online MT System
Dissemination level:	PU
Contractual date of delivery:	25 th June 2013
Actual date of delivery:	3th July 2013
Author(s):	Pablo Nieto Caride, Giuseppe Deriard Nolasco, Pedro L. Díez Orzas, Félix Fernández, Consuelo Aldana, Pablo Badía, Ankit Srivastava, Declan Groves, Thomas Ruedesheim, Román Díez, Alberto Crespo
Participants:	Linguaserve, Dublin City University, Lucy Software, and the Spanish Tax Office
Internal Reviewer:	Linguaserve
Workpackage:	WP4
Task Responsible:	Giuseppe Deriard
Workpackage Leader:	Pedro L. Díez Orzas

Revision History

Revision	Date	Author	Organization	Description
0	03/2013	Pablo Nieto Caride	Linguaserve	Template
1	04/2013	Pablo Nieto Caride	Linguaserve	Compiling information and draft creation
2	05/2013	Pablo Nieto Caride	Linguaserve	First version
3	05/2013	Pedro L. Díez Orzas	Linguaserve	Revised structure and completed section 6
4	06/2013	Pablo Nieto Caride Ankit Srivastava Thomas Ruedesheim	Linguaserve DCU Lucy Soft.	Final version
5	06/2013	Pablo Nieto Caride	Linguaserve	Revised final version
6	06/2013	All	All	Revised final version
7	07/2013	Pedro L. Díez Orzas	Linguaserve	Revised and approved final version
8	08/2013	Pablo Nieto Caride	Linguaserve	Summary and video

1. CONTENTS

Document Information	2
Revision History	2
1. Contents	3
2. Executive Summary	6
3. Introduction	7
4. ITS 2.0 compliant Real-time Multilingual Publishing System	8
4.1 System functionality.....	8
4.1.1 ATLAS Real-time.....	8
4.1.2 RBMT: Lucy Software.....	13
4.1.3 SMT: MaTrex.....	16
4.2 System performance	19
4.2.1 ATLAS Real-time.....	19
4.2.2 RBMT: Lucy Software.....	21
4.2.3 SMT: MaTrEx.....	23
4.3 Security.....	24
4.3.1 Telefonica Cache and Apache Server.....	24
4.3.2 Spanish Tax Agency CMS	24
4.3.3 ATLAS Real-time Servers.....	24
4.3.4 Apache Server	24
4.3.5 ATLAS Real-time.....	24
4.3.6 Database	25
4.3.7 MT Systems.....	25
4.3.8 ATLAS File Management Tool and CAT Tool	25
4.4 Translation memory update.....	25
4.5 Environments	27
4.5.1 Pre-production.....	27
4.5.2 Production	27

5.	Localisation project in Real-time System with ITS 2.0	29
5.1	Project description	30
5.1.1	Preparation tasks	30
5.1.2	Execution in Pre-production	31
5.1.3	Maintenance tasks in Production.....	32
5.2	Post-edition applying ITS 2.0.....	33
5.2.1	ITS 2.0 PE contextual information	33
5.2.2	ITS 2.0 PE activation rules	34
5.3	Evaluation of linguistic resources and PE Efforts	37
5.3.1	PE Effort Evaluation	37
5.3.2	RBMT: Lucy Software	38
5.3.3	SMT: MaTrEx.....	38
6.	Business report and Exploitation.....	39
6.1	SWOT analysis	41
6.2	Opportunities rise from needs	43
6.2.1	Small and Medium companies.....	43
6.2.2	E-commerce	43
6.2.3	Web 2.0 and WEB 3.0	44
6.2.4	Online Public sector	44
6.2.5	Complex back-office web sites	44
6.2.6	New jobs for young translators	44
6.3	Final recommendations and best practices	45
6.3.1	Oriented to Real-time Multilingual Publication Systems (RTMPS)	45
6.3.2	Oriented to annotators.....	45
6.3.3	Oriented to Post-editing	47
6.3.4	Oriented to customer or consumer	47
6.3.5	Oriented to QA.....	47
6.4	Extensions: Readiness	47

6.4.1	Definition	47
6.4.2	Implementation	48
6.5	Beyond 2013: future work	50
6.5.1	Source and target language tools	50
6.5.2	LINGUASERVE Challenges	50
6.5.3	Lucy Challenges.....	51
6.5.4	DCU Challenges.....	51
7.	Glossary of terms and acronyms	52
8.	References	54

2. EXECUTIVE SUMMARY

The present document constitutes a detailed report of the testing and evaluation of the Online MT System Showcase (D4.2.1) and final recommendations and best practices. It contains the different final tests to check functionality, behaviour, speed and security, and it also describes the linguistic material to be used in the implementation: glossaries, post edited RBMT and the results of the evaluation of the linguistic resources in RBMT and SMT for the goal of processing LT-Web metadata in a Real-time Multilingual Publishing System (RTMPS) for the Spanish Tax Agency (www.agenciatributaria.es).

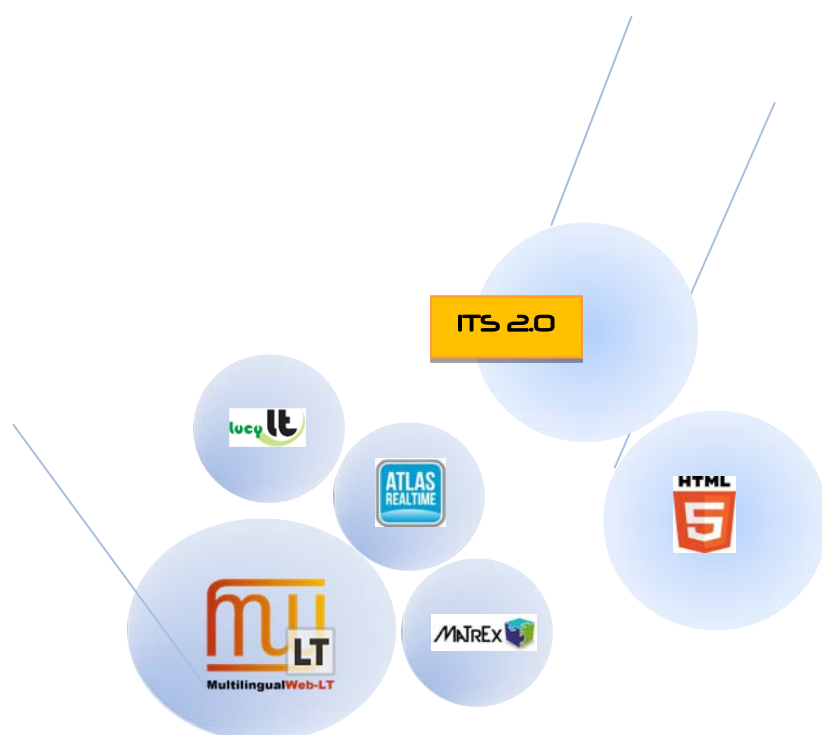


Figure 2.1: Logos

3. INTRODUCTION

Recent advances in machine translation are emphasising the importance of language technology for the Web. An ever increasing necessity of a proper and useful metadata for content processing is in the spotlight nowadays. The metadata set offered by ITS 2.0 intends to address the “black box” problem of the three gaps in the chain of content processing and consumption on the Web, which are:

1. An online machine translation service might make mistakes like translation of fixed terminology or named entities. Language technology does not know about metadata in the source content, e.g. “What parts of the input should be translated?”
2. In the database from which the translated text has been generated, the information about translatability might have been available. However, the machine translation system does not acknowledge that kind of “hidden Web” information. There is no description of the processes available, which was the basis for generating “surface Web” pages.
3. The last gap is about clearly defined identification of metadata: Information like gaps 1 and 2 is not described in a standardized manner. The same is true for resources used by language technology applications (e.g. “what lexicon is used for machine translation?”) and the applications themselves (e.g. “general purpose machine translation versus application tailored towards a specific domain”). This hinders the ad hoc creation of language technology applications in the Web, i.e. the re-combination of resources and application modules.

These gaps are especially important when the localization system only has the html content as input.

After the integration of an ITS 2.0 processor in different online MT systems, as described in D4.1.4, and its usage in the present Showcase, as explained in D4.2.1, a report to demonstrate a certain solution for the before mentioned gaps was needed and thus the present document, which includes three different sorts of reports related to a technical, linguistic and business analysis, made accordingly with the outcome of the different tests performed out of the Showcase.

Finally the report presents the gained understanding regarding the solution to the processing of web content inline mark-up and to leverage content metadata both when localizing content and when exploiting real-time web solutions and MT engines.

4. ITS 2.0 COMPLIANT REAL-TIME MULTILINGUAL PUBLISHING SYSTEM

In the first place the description of the ITS 2.0 compliant Real-time Multilingual Publishing System is presented, comprising information such as the system's functionalities and different configurations, performance tests, security configurations, the translation memory update process, to integrate in an end-client the ATLAS Real-time, Lucy LT and MaTrEx solution modified to be ITS 2.0 compliant.

4.1 System functionality

4.1.1 ATLAS REAL-TIME

ATLAS Real-time is a real-time solution for multilingual publications through the Internet (RTMPS). ATLAS Real-time allows the user to navigate a website in a completely transparent way in the selected languages. The system is fully scalable and configurable, allowing multi-server redundant configurations, and can be integrated in the client's Web site.

When ATLAS Real-time receives a translation request, it downloads the original document, sends it to the MT System and finally delivers the target language translated document to the user's browser.

The overall system is composed by four modules (applications):

4.1.1.1 ATLAS REAL-TIME

It is a Web application that manages the translation requests sent by the client's Web server. The system receives a translation request, downloads the original document, checks whether it was translated before or not so as to retrieve it from the cache system or to send it to the MT, and finally serves the translated document in the user's browser. ATLAS Real-time also has an administration interface that allows administrators or project managers to manage and configure projects, that is, specific tailored configurations for the different clients' needs.

4.1.1.2 MT SYSTEM

It is a rule-based or statistical Machine Translation System that can use plain text translation memories.

4.1.1.3 ATLAS FILEMANAGEMENT TOOL (FMT)

It is an application that is responsible for managing the transfer of files from the MT System to the Content Editor and for updating the MT's translation memories and glossaries.

4.1.1.4 CONTENT EDITOR

It is an application that allows the edition of the translations performed by the MT System.

Here is a diagram showing the basic workflow of the overall system:

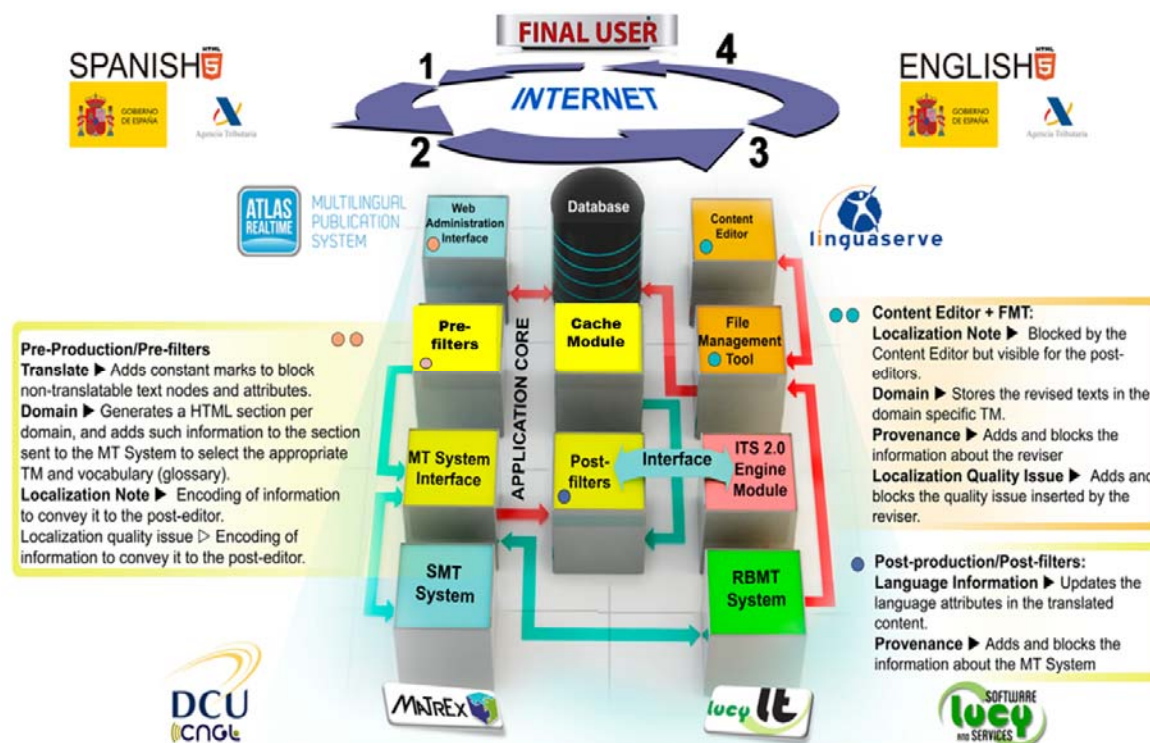


Figure 4.1: ATLAS Real-time modules configuration

4.1.1.5 CONFIGURATION

The necessary steps to configure a web site in ATLAS real-time are:

- 1) Access to the administration and configuration interface. Since the credentials cannot be provided due to a privacy restrictions, the following link to a video is given offering an explanation in detail with respect to the ATLAS RT configuration (<http://youtu.be/PXZfZ7sBloE>).

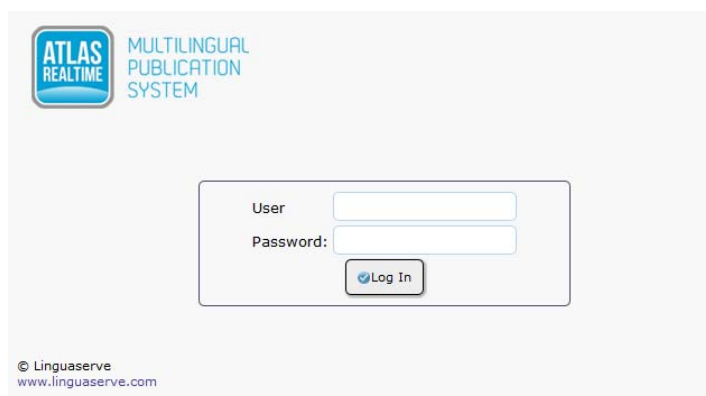


Figure 4.2: ATLAS Real-time login screen

- 2) Once inside, you will need “Administrator” or “Project Manager” permission to create and manage projects:

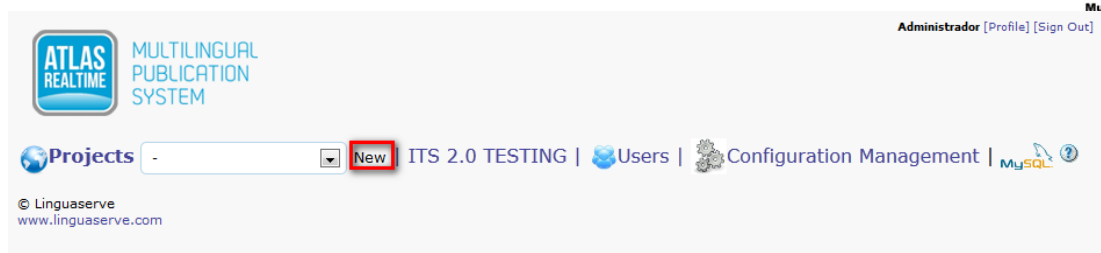


Figure 4.3: ATLAS Real-time administration interface I

- 3) Once the main information is defined, you need to select other parameters such as the language pairs, the MT engine, whether you want to use the cache system or not, or if the project will process ITS 2.0 among others:

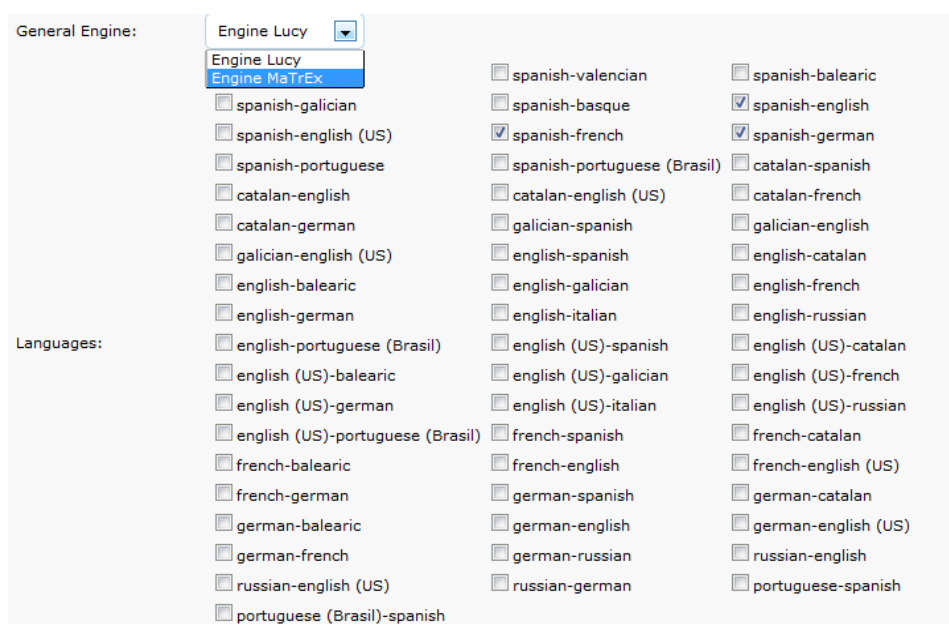


Figure 4.4: ATLAS Real-time modules configuration I

It is possible to select either Lucy MT or DCU MT. The translation requests would be processed and translated by the selected MT System, in a transparent way for the user.



Figure 4.5: ATLAS Real-time modules configuration II

- 4) After that, it is imperative to set up the domains that apply to the project:

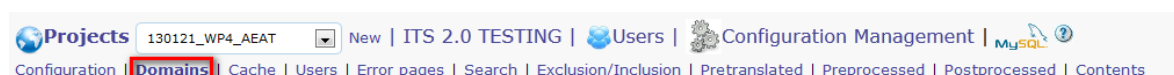


Figure 4.6: ATLAS Real-time administration interface II

For the domains is at least required to specify the patterns of the request that will be sent to the server, the original web site, and the parsed links in the target language, using regular expressions, and the language pair. With this information the system, when receiving a new translation request, will know where to download the content from and how to rewrite the links.

✗	http://its2-aeat.linguaserve.net/en/AEAT.internet(?!Inicio_en_GB)/?.*	http://its2demo.aeat.es/AEAT.internet(?!Inicio_es_ES)/?.*	http://its2demo.aeat.es/AEAT.internet/en_gb(?!Inicio_en_GB)/?.*	es - en
✗	http://its2-aeat.linguaserve.net/en/AEAT.internet/Inicio_en_GB/?.*	http://its2demo.aeat.es/AEAT.internet/Inicio_es_ES/?.*	http://its2demo.aeat.es/AEAT.internet/en_gb/Inicio_en_GB/?.*	
✗	http://its2-aeat.linguaserve.net/en/static_files/?.*	http://its2demo.aeat.es/static_files/?.*	http://its2demo.aeat.es/static_files/?.*	
✗	http://its2-aeat.linguaserve.net/fr/AEAT.internet(?!Inicio_fr_FR)/?.*	http://its2demo.aeat.es/AEAT.internet(?!Inicio_es_ES)/?.*	http://its2demo.aeat.es/AEAT.internet/fr_fr(?!Inicio_fr_FR)/?.*	es - fr
✗	http://its2-aeat.linguaserve.net/fr/AEAT.internet/Inicio_fr_FR/?.*	http://its2demo.aeat.es/AEAT.internet/Inicio_es_ES/?.*	http://its2demo.aeat.es/AEAT.internet/fr_fr/Inicio_fr_FR/?.*	
✗	http://its2-aeat.linguaserve.net/fr/static_files/?.*	http://its2demo.aeat.es/static_files/?.*	http://its2demo.aeat.es/static_files/?.*	
✗	http://its2-aeat.linguaserve.net/de/AEAT.internet(?!Inicio_de_DE)/?.*	http://its2demo.aeat.es/AEAT.internet(?!Inicio_es_ES)/?.*	http://its2demo.aeat.es/AEAT.internet/de_de(?!Inicio_de_DE)/?.*	es - de
✗	http://its2-aeat.linguaserve.net/de/AEAT.internet/Inicio_de_DE/?.*	http://its2demo.aeat.es/AEAT.internet/Inicio_es_ES/?.*	http://its2demo.aeat.es/AEAT.internet/de_de/Inicio_de_DE/?.*	
✗	http://its2-aeat.linguaserve.net/de/static_files/?.*	http://its2demo.aeat.es/static_files/?.*	http://its2demo.aeat.es/static_files/?.*	

Figure 4.7: ATLAS Real-time domains configuration

- 5) The last step is to set up the “Pre-translated content” such as pre-translated images, how the original content must be processed (Pre-processed), and how the translated content must be processed (Post-processed):

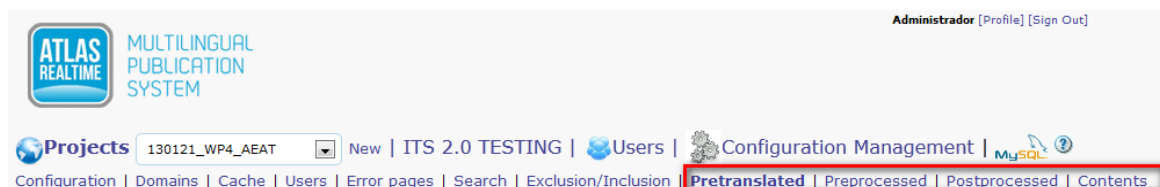


Figure 4.8: ATLAS Real-time administration interface III

Pre-translated:

✗	[en] ##DL##\${1}_en_gb.\$ {2}	
✗	[fr] ##DL##\${1}_fr_fr.\$ {2}	
✗	[de] ##DL##\${1}_de_de.\$ {2}	

Figure 4.9: ATLAS Real-time pre-translated configuration

Pre-processed:

  <code><meta name="(description keywords)" content=</code>	 META_TAG_PREPRO	  [en] <code><img name="{1}" alt=</code>
		  [fr] <code><img name="{1}" alt=</code>
		  [de] <code><img name="{1}" alt=</code>
  <code><(input area)</code>	 TAG_PREPRO	  [en] <code><img id="{1}_las_translation"</code>
		  [fr] <code><img id="{1}_las_translation"</code>
		  [de] <code><img id="{1}_las_translation"</code>

Figure 4.10: ATLAS Real-time modules pre-processed configuration

Post-processed:

  <code></code>		  [en] <code></code>
		  [fr] <code></code>
		  [de] <code></code>
  <code><(body <title >)</code>	 provenance	  [en] <code>{1} its-org="Lucysoftware" its-tool="MTEngine 6.9" its-rev-person="English revisor 1" its-rev-org=" ...</code>
		  [fr] <code>{1} its-org="Lucysoftware" its-tool="MTEngine 6.9" its-rev-person="French revisor 1" its-rev-org=" ...</code>
		  [de] <code>{1} its-org="Lucysoftware" its-tool="MTEngine 6.9" its-rev-person="German revisor 1" its-rev-org=" ...</code>
  <code><script type="application/its+xml">.*?</script></code>	 remove_script_rules	  [en]
		  [fr]
		  [de]
  <code>(its-loc-quality-issue-enabled=".*")</code>	 lq_deactivated	  [en] <code>{1} its-loc-quality-issue-enabled="no"</code>
		  [fr] <code>{1} its-loc-quality-issue-enabled="no"</code>
		  [de] <code>{1} its-loc-quality-issue-enabled="no"</code>
  <code><img id="(area input)_las_translation"</code>	 TAG_POSTPRO	  [en] <code><{1}</code>
		  [fr] <code><{1}</code>
		  [de] <code><{1}</code>
  <code><img name="(keywords description)" alt=</code>	 META_TAG_POSTPRO	  [en] <code><meta name="{1}" content=</code>
		  [fr] <code><meta name="{1}" content=</code>
		  [de] <code><meta name="{1}" content=</code>

Figure 4.11: ATLAS Real-time modules post-processed configuration

4.1.1.6 FUNCTIONALITIES

While the overall process of serving translation requests to the user remains the same, with the modifications introduced in the system to be ITS 2.0 compliant, it will allow to:

- Translate HTML5 documents from an ITS-conformant Web CMS.
- Control precisely which sentences or sentence fragments should or should not be translated and which is the source language.
- Choose, in a paragraph, sentence or at word level, the appropriate training corpora or glossary (depending on the MT System) that should be used on the translation by the MT Systems in order to disambiguate.
- Convey information about the localization process to editors.
- Communicate the identity of agents that have been involved in the revision and the translation of the content, and to allow translation quality reviewers, to evaluate how the performance of these agents affects the quality of the translation.

4.1.1.7 TRANSLATION REQUEST WORKFLOW FOR ITS 2.0

This illustrates what happens from the moment the system receives a translation request until it serves the translated content in the user's browser:

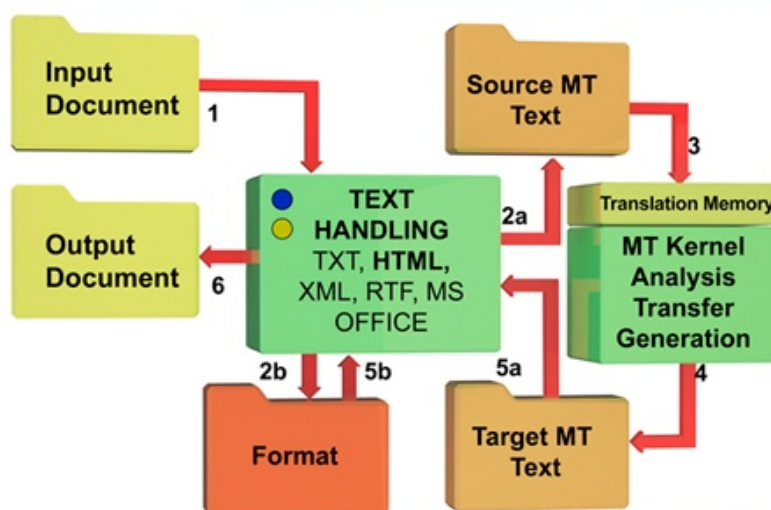
- The user makes a translation request and the client's Web Server redirects the request to ATLAS Real-time.

- ATLAS Real-time receives a translation request, downloads the original document, and checks whether it has been translated before or not so as to retrieve it from the Cache System or to send it to the MT System.
- ATLAS Real-time also pre-processes the file if the client's project configuration says so.
 - **Translate:** The system adds constant marks to block non-translatable text nodes and attributes.
 - **Domain:** The system generates an HTML section per domain, and adds such information to the section sent to the MT System to select the appropriate TM and vocabulary (glossary).
 - **Localization Note:** Encoding of information to convey it to the post-editor.
 - **Localization quality issue:** Encoding of information to convey it to the post-editor.
- If the content of the original document, or piece of the document, remains unchanged since the last translation request processed for that very document, the system retrieves the already translated content from the system's database.
- If the content of the original document is new or was not previously processed, then it's sent to the MT System.
- The MT System, after parsing the content to generate the list of translatable segments, checks if any of those segments is already stored in the TM (Translation Memory) to retrieve them and subsequently translates the remaining segments; here starts a sub-process related to the post-edition.
 - For all of those who are not stored in the TM, the MT System creates a pair of input and output files (LMI and LMO), which are processed and conveyed to the Content Editor by the FMT.
 - Once in the Content Editor after the content of these input and output files is edited and validated, new LMI and LMO files are created and used by the FMT to update the TM.
 - **Localization Note:** Blocked by the Content Editor but visible for the post-editors.
 - **Domain:** Stores the revised texts in the domain specific TM.
 - **Provenance:** The system adds and blocks the information about the reviser.
 - **Localization Quality Issue:** The system adds and blocks the quality issue inserted by the reviser.
- The MT System sends the translated content back to ATLAS Real-time.
- ATLAS Real-time also post-processes the file if the client's project configuration says so.
 - **Language Information:** The system updates the language attributes in the translated content.
 - **Provenance:** The system adds and blocks the information about the MT System.
- Finally the translated document is shown in the user's browser.

4.1.2 RBMT: LUCY SOFTWARE

The Lucy LT Engine is a rule-based MT engine bundled with a text handling module for various document formats. It also contains a basic translation memory (TM) module for pure text.

- **Text Handling – Pre/Trans Domain** ► Set translation parameter "Subject Areas" according to global Domain mapping.
- Text Handling - PreTranslate** ► Hide nodes that should not be translated.



- **Text Handling – Pre/Post Language Info** ► Hide nodes not matching the current translation direction (source). Update language information after translation (target).

Figure 4.12: Lucy's RMBT system configuration

A set of translation parameters controls the whole processing of documents. The processing units are documents, which means that parameters have a document-wide scope. The parameters can be general parameters or parameters focused on the language pair, some of them are:

- **TRANSLATION_DIRECTION:** The source and target languages activated for the current processing request.
- **TRANSLOGFILES:** Save unk, lmi, lmo and/or log files of a text translation in the directory specified by the engine parameter TRANSLOGDIR.
- **LEXICON:** Pathname of the lexicon without filename extension. If a relative path is used, it will be relative to the engine root folder.
- **SUBJECT_AREAS:** If a word has alternative translations, the one with the specified subject area is taken.
- **MEMORY_MODULES:** List of module names used for translation (should be separated by '|').
- **PPM_USE:** Activates / deactivates Perl Pattern Matcher.

De- and Reformatting of HTML documents are implemented as a proprietary pseudo-parser that is capable of handling a configurable set of HTML5 tags.

4.1.2.1 CONFIGURATION

Lucy MT on its own can process ITS 2.0 data categories, but when integrated in the RTMPS, Linguaserve's ATLAS Real-time takes care of handling the ITS 2.0 data category due to the client's restriction of dividing the document structure into sections, which makes it impossible to connect the System to an ITS 2.0 compliant MT Engine, because the HTML5 integrity sent to the MT System is not guaranteed.

Given this, a general configuration with Lucy MT is needed. For that purpose it is required to set up the above

mentioned parameters, ATLAS Real-time can do it in real-time depending on the translation request.

- **TRANSLATION_DIRECTION:** Depending on the language pair, it can have any of the following values:
 - SPANISH-ENGLISH
 - SPANISH-FRENCH
 - SPANISH-GERMAN
- **TRANSLOGFILES:** 2 = save lmi and lmo.
- **SUBJECT_AREAS:** Depending on the domain data category, after the corresponding mapping, it can have one or more of the following values:
 - GV - General Vocabulary
 - CSV - Common Social Vocabulary
 - ECON - Economy & Trade
 - ECON-POL - Political Economics
 - ECON-USER - Economy, User Entries
 - LAW - Law & Legal Science
 - POL - Government, Public Authorities, International Relations
- **MEMORY_MODULES:** (modulo_agencia_es_en|modulo_agencia_es_fr|modulo_agencia_es_de).
- **PPM_USE:** The use of filters is required in order to block the translation of the SPT. The value should be set to "1". The filters should have the following regular expressions:
 - @ANT\[.*?\]@
 - \[@@.*?@@\]

Other parameters:

Since the system translates on the fly, there is no need to mark unknown words or similar that can be helpful when revising.

- **MARK_UNKNOWNS:** Words which are not found in the lexicon. The value should be set to "0".
- **MARK_ALTERNATIVES:** Marks words for which at least one alternative translation within the selected subject area is found in the lexicon. The value should be set to "0".

Lucy MT should be up and running. The configuration console should look like this:

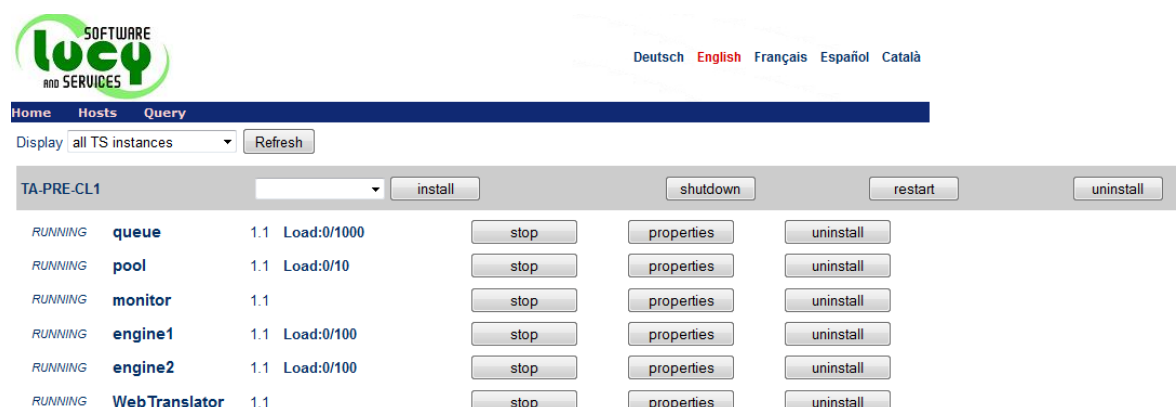


Figure 4.13: Lucy's administration interface

4.1.2.2 FUNCTIONALITIES

The ITS 2.0 version 11 of Lucy MT offers the next functionalities:

Data categories may be used locally (attributes) or globally in external XML rule files or inline rules (inside XML scripts). The specifications for handling the data categories (default values, overriding and inheritance) are contained in the ITS 2.0 specification document.

The handling of the standard HTML “lang” attribute is made in the following way: Given a translation direction for the translation task, the system translates only text content that is not marked by a “lang” attribute or that is marked as being in the current source language. The “lang” attribute is inherited to sub-elements and may be overwritten. If text content is marked, the mark-up value will be changed to the current target language after translation took place.

- Automatic processing of Translate
 - o *Global rules and local “translate” attribute may override the default.*
 - o *No-translate nodes are masked as data and will not be processed by MT Kernel.*
- Automatic processing of Domain
 - o *Global rules define a domain property.*
 - o *Domain property can be mapped to tool-specific namespaces.*
 - o *Lucy LT engine accepts one domain per document (“Subject Area”).*
 - o *Subject Area used for word sense disambiguation.*

4.1.2.3 WORKFLOW

The general workflow is as follows:

1. An input document is passed to the Text Handling module of LT Engine
2. A format-specific text handling routine separates textual content of the input from format information and stores them separately (De-formatting).
3. The pure textual content is passed over to the MT kernel for translation.
 1. The filters prevent the marked segment and words from being translated.
 2. The MT Kernel translates using the subject area mapped from the original domain.
4. The MT kernel writes the target text into a file.
5. The text handling module merges the format information with the target text (Re-formatting) and produces the output document.
6. The output document is returned to the client.

4.1.3 SMT: MATREX

The Dublin City University (DCU) Machine Translation (MT) system, also known as **MaTrEx** (Machine Translation using Examples) is a hybrid multi-engine MT system (implementing statistical (phrase-based / tree-based) and example-based paradigms). For the MLW-LT project, MaTrEx uses the phrase-based statistical MT (PB-SMT) engine based on the open source log-linear phrase-based decoder Moses.

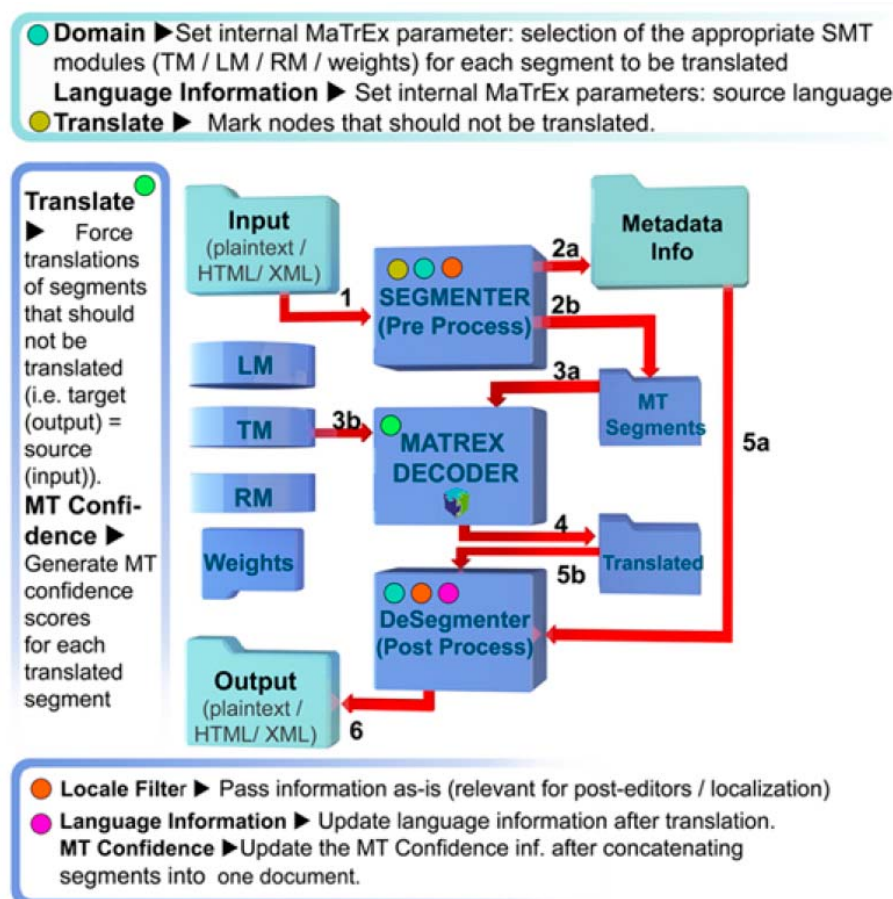


Figure 4.14: DCU's MaTrEx modules configuration

The architecture of the MaTrEx system consists of an input document splitter module (pre-processing), a translation module (MaTrEx), and an output document merging module (post-processing).

The pre-processing singles out the translatable content to be sent to the MaTrEx decoder, while the post-processing concatenates all sub-segments together into one document.

4.1.3.1 CONFIGURATION

A **PB-SMT** system in contrast to a rule-based MT system is data-driven and is trained on a large corpus of sentence-aligned bilingual text (text written in source language (input) and their corresponding human translations in target language (output)). PB-SMT systems extract knowledge in the form of phrase pairs (sub-sentential alignments) from the sentence-aligned bilingual text (training data). A number of probabilistic attributes (features) for each of these phrase pairs are defined. Some of the features used in MaTrEx include source – target translation conditional probabilities (**translation model**), target language probabilities (**language model**), and source – target relative order conditional probabilities (**reordering model**). Such features are then combined in a log-linear model, the **coefficients** of which are optimized on an objective function measuring translation quality.

The input to the MaTrEx system consists of:

1. Input text (source language)
2. Source Language Code
3. Target Language Code
4. Translation Model, Language Model, Reordering Model, Coefficients of the Log-linear Model (Default settings / set by source and target language code).

The output of the MaTrEx system consists of:

1. Output text (target language);
2. MT confidence score denoting the probability of the particular translation.

MaTrEx can be accessed and tested at <http://srv-cngl.computing.dcu.ie/mlwlt/>.

4.1.3.2 FUNCTIONALITIES

While the underlying MaTrEx system remains unchanged, a sequence of pre-processing and post-processing scripts is added to the framework to make it ITS-compatible for the MLW-LT project.

The original MaTrEx system only translated single sentences. It is now capable of translating complete web pages (HTML and XML).

The pre-processing module is written in PERL and takes as input a document or a segment and splits it into sub-segments based on the ITS metadata tags. Each individual sub-segment is then passed to the decoder for translation.

The post-processing module is also written in PERL. It merges all the translated sub-segments and composes the document with all the required ITS 2.0 metadata tags to be generated.

4.1.3.3 WORKFLOW

Once the MaTrEx engine has been trained, in order to translate (also known as decoding), the following workflow is followed:

1. An input text (in source language) is segmented into a number of phrases (consecutive sequence of words);
 - a. **Translate:** Mark nodes that should not be translated.
 - b. **Domain:** Set internal MaTrEx parameter: selection of the appropriate SMT modules (TM / LM / RM / weights) for each segment to be translated.
 - c. **Language Information:** Set internal MaTrEx parameters: source language.
2. Each phrase is looked up in the translation model to compute its translation (in target language) and the phrases are scored using the log-linear features;
 - a. **Locale Filter:** Pass information as-is (relevant for post-editors / localization).
 - b. **Translate:** Force translations of segments that should not be translated (i.e. target (output) =

source (input)).

- c. **MT Confidence:** Generate MT confidence scores for each translated segment.
3. The target language sentence is composed and output along with an MT confidence score.
 - a. **Language Information:** Update language information after translation.
 - b. **MT Confidence:** Update the MT Confidence information after concatenating segments into one document.

4.2 System performance

This sub-section presents a series of tables and graphics with data extracted from various performance tests. The sample of files is a random selection of different kinds of pages with different number of words and size, from big to small.

4.2.1 ATLAS REAL-TIME

Latency:

- Without ITS 2.0 Processing (2 x Intel Xeon E5450 2.99 GHz, 4 GB RAM)

<i>Lucy MT (2 engines)</i>	<i>Download</i>		<i>Translate</i>		<i>Total</i>	
	<i>With cache</i>	<i>Without cache</i>	<i>With cache</i>	<i>Without cache</i>	<i>With cache</i>	<i>Without cache</i>
Press releases (1200-1300 words)	1 ms	52 ms	5 ms	1,513 ms	6 ms	1,565 ms
History (1400-1500 words)	1 ms	59 ms	5 ms	1,583 ms	6 ms	1,642 ms
General terms of use (1000-1100 words)	1 ms	52 ms	5 ms	1,359 ms	6 ms	1,411 ms
Information returns (50-100 words)	1 ms	54 ms	5 ms	1,364 ms	6 ms	1,418 ms
Organization Chart (200-300 words)	1 ms	59 ms	5 ms	1,268 ms	6 ms	1,327 ms

Figure 4.15: ATLAS Real speed test using the RBMT and without using ITS 2.0 module

DCU MT

	Download		Translate		Total	
	With cache	Without cache	With cache	Without cache	With cache	Without cache
Press releases (1200-1300 words)	1 ms	52 ms	5 ms	11,138 ms	6 ms	11,202 ms
History (1400-1500 words)	1 ms	59 ms	5 ms	26,314 ms	6 ms	26,385 ms
General terms of use (1000-1100 words)	1 ms	52 ms	5 ms	28,411 ms	6 ms	28,475 ms
Information returns (50-100 words)	1 ms	54 ms	5 ms	10,526 ms	6 ms	10,592 ms
Organization Chart (200-300 words)	1 ms	59 ms	5 ms	10,457 ms	6 ms	10,528 ms

Figure 4.16: ATLAS Real speed test using the SMT and without using the ITS 2.0 module

- With ITS 2.0 Processing (2 x Intel Xeon E5450 2.99 GHz, 4 GB RAM)

Lucy MT (2 engines)

	Download		Translate		Total	
	With cache	Without cache	With cache	Without cache	With cache	Without cache
Press releases (1200-1300 words)	1 ms	75 ms	12 ms	1,537 ms	13 ms	1,612 ms
History (1400-1500 words)	1 ms	51 ms	6 ms	1,513 ms	7 ms	1,564 ms
General terms of use (1000-1100 words)	1 ms	53 ms	6 ms	1,711 ms	7 ms	1,764 ms
Information returns (50-100 words)	1 ms	78 ms	6 ms	1,328 ms	7 ms	1,406 ms

Organization Chart (200-300 words)	1 ms	70 ms	5 ms	1,387 ms	6 ms	1,457 ms
--	------	-------	------	----------	------	----------

Figure 4.17: ATLAS Real speed test using the RBMT and the ITS 2.0 module

DCU MT	Download		Translate		Total	
	With cache	Without cache	With cache	Without cache	With cache	Without cache
Press releases (1200-1300 words)	1 ms	75 ms	12 ms	14,266 ms	13 ms	14,367 ms
History (1400-1500 words)	1 ms	51 ms	6 ms	29,470 ms	7 ms	29,535 ms
General terms of use (1000-1100 words)	1 ms	53 ms	6 ms	31,679 ms	7 ms	31,764 ms
Information returns (50-100 words)	1 ms	78 ms	6 ms	13,558 ms	7 ms	13,650 ms
Organization Chart (200-300 words)	1 ms	70 ms	5 ms	13,401 ms	6 ms	13,477 ms

Figure 4.18: ATLAS Real speed test using the SMT and the ITS 2.0 module

Telefónica Cache System

To take advantage of the user's navigation, the system counts with a high level cache System. Apart from the ATLAS Real-time own cache system, the environment also uses a higher one so that once a translated page has been requested, the next requests will serve the content directly without being processed until the TTL (Time To Live) expires. This cache system gets an average response time of 50 ms.

4.2.2 RBMT: LUCY SOFTWARE

Latency:

- Without ITS 2.0 Processing (2 x Intel Xeon E5345 2.33 GHz, 3 GB RAM)

Pre-filter	Translate	Post-filter	Total
------------	-----------	-------------	-------

Press releases (1200-1300 words)	203 ms	1,079 ms	46 ms	1,359 ms
History (1400-1500 words)	329 ms	3,125 ms	63 ms	3,547 ms
General terms of use (1000-1100 words)	250 ms	3,422 ms	47 ms	3,750 ms
Information returns (50-100 words)	203 ms	735 ms	46 ms	1,015 ms
Organization Chart (200-300 words)	547 ms	578 ms	47 ms	1,203 ms

Figure 4.19: Lucy RBMT system speed test without using the ITS 2.0 module

- **With ITS 2.0 Processing**

	Pre-filter	Translate	Post-filter	Total
Press releases (1200-1300 words)	109 ms	1,016 ms	31 ms	1,188 ms
History (1400-1500 words)	141 ms	2,922 ms	47 ms	3,140 ms
General terms of use (1000-1100 words)	126 ms	2,937 ms	31 ms	3,141 ms
Information returns (50-100 words)	94 ms	469 ms	16 ms	609 ms
Organization Chart (200-300 words)	63 ms	532 ms	15 ms	656 ms

Figure 4.20: Lucy RBMT system speed test using the ITS 2.0 module

4.2.3 SMT: MATREX

Quality:

The quality of the translation depends on how well trained is the SMT, but this is out of scope in this work package (WP4) because is subjected to WP5.

Latency:

- **Without ITS 2.0 Processing (2 x Intel Xeon E5450 2.99 GHz, 4 GB RAM)**

	Segmenter	Translate	De-segmenter	Total
Press releases (1200-1300 words)	44 ms	11,085 ms	9 ms	11,138 ms
History (1400-1500 words)	44 ms	26,262 ms	8 ms	26,314 ms
General terms of use (1000-1100 words)	44 ms	28,358 ms	9 ms	28,411 ms
Information returns (50-100 words)	44 ms	10,473 ms	9 ms	10,526 ms
Organization Chart (200-300 words)	36 ms	10,413 ms	8 ms	10,457 ms

Figure 4.21: DCU MaTrEx SMT system speed test without using the ITS 2.0 module

- **With ITS 2.0 Processing (2 x Intel Xeon E5450 2.99 GHz, 4 GB RAM)**

	Segmenter	Translate	De-segmenter	Total
Press releases (1200-1300 words)	1,508 ms	11,085 ms	1,673 ms	14,266 ms
History (1400-1500 words)	1,512 ms	26,262 ms	1,696 ms	29,470 ms

General terms of use (1000-1100 words)	1,612 ms	28,358 ms	1,709 ms	31,679 ms
Information returns (50-100 words)	1,492 ms	10,473 ms	1,593 ms	13,558 ms
Organization Chart (200-300 words)	1,472 ms	10,413 ms	1,576 ms	13,401 ms

Figure 4.22: DCU MaTrEx SMT system speed test using the ITS 2.0 module

4.3 Security

To ensure the security of the RTMPS, a series of security measures have been implemented:

4.3.1 TELEFONICA CACHE AND APACHE SERVER

The Apache that serves the original and translated web pages is set up to the Telecommunications leading company in Spain, with that said, it is clear that the highest security levels are ensured. The Web Server uses a cache System that serves the web pages in an eye blink.

4.3.2 SPANISH TAX AGENCY CMS

The Tax Agency's CMS is located in the internal network of the Tax Agency (DMZ); the server is not accessible from outside the network. Given the importance of this Government Agency, the high security measures provided by such Agency ensure the safety of the System.

4.3.3 ATLAS REAL-TIME SERVERS

The RTMPS is inside a network protected by a firewall, the system is set up in a multi-server cluster configuration allowing high availability and redundancy.

4.3.4 APACHE SERVER

The Apache Server has been configured to hide the version in order to avoid from hackers from using the known vulnerabilities of the server for malicious acts, besides it has the SSL enabled and the trace method disabled.

4.3.5 ATLAS REAL-TIME

ATLAS Real-time runs under HTTPS and the access to the main console is protected by user and an MD5 encrypted password, besides the version of PHP is hidden as well.

4.3.6 DATABASE

The database is protected under a root user with an MD5 encrypted password, and the database objects have assigned the correct privileges.

4.3.7 MT SYSTEMS

The requests made by ATLAS Real-time to Lucy MT are made from inside the server, thus the safety is ensured. When it comes to MaTrEx, the server has to connect via WebService with DCU's server <http://srv-cngl.computing.dcu.ie>, the security thus relies on both servers and the communications between them.

4.3.8 ATLAS FILE MANAGEMENT TOOL AND CAT TOOL

The communications between the ATLAS Real-time servers, where the FMT resides and the server where the CAT Tool is located, are internal, so the standard security measures are sufficient to ensure security.

4.4 Translation memory update

Each time the RBMT System translates a segment that is not loaded in its Translation Memory, it generates a pair of input and output files (LMI and LMO) with the original segments and the machine translated ones respectively.

These files are plain text files, so any segment that belongs to an element with ITS 2.0 metadata inserted in it will be previously converted to SPT by the system, like this:

From

```
Declaración Tributaria Especial(<span its-loc-note="Stands for
'Declaración Tributaria Especial', use acronym in target language" its-loc-
note-type="description">DTE</span>)
```

To

```
Declaración Tributaria Especial(&[@@locNote=Stands for 'Declaración
Tributaria Especial', use acronym in target
language&locNoteType=description@@])
```

The FMT when conveying the file to CAT Tool server, process the SPT and transform it to the proprietary tag to be protected by the CAT tool.

```
Declaración Tributaria Especial (<las_its2 locNote="Stands for 'Declaración
Tributaria Especial', use acronym in target language"
locNoteType="description" /> DTE)
```

In the CAT tool, the tag will be protected and the document properly segmented because of the next configurations in the CAT Tool:

[Tags]

```
Inline=0 (<las_its2&\>)|(HTML_HASH=.*)
```

[SegmentationParaRules]

Rule1=0 <<>>\n

The filter also blocks the file's hash code, also included in both LMI and LMO.

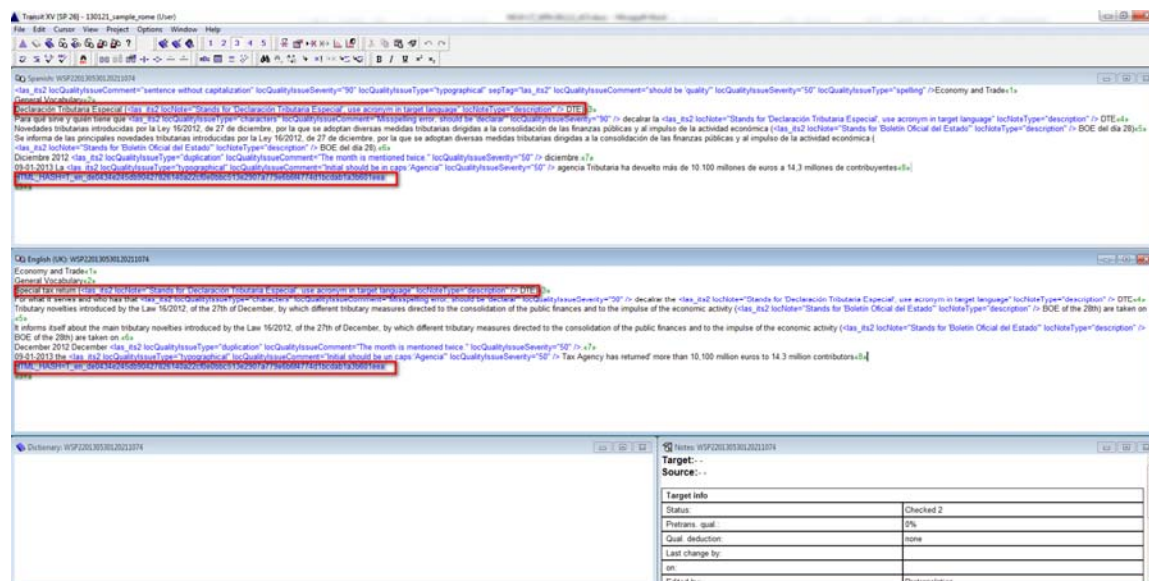


Figure 4.23: CAT Tool

After the revision is completed opposite process begins. The proprietary tag is converted into SPT, and the pair of files is uploaded to the memory via Lucy's TM Console commands:

```
1  lucy off
2  rm -rf /tmp/lucy
3  set LMI_FILE=1
4  set LMO_FILE=2
5  set SRC_LANG=es
6  set TRG_LANG=en
7  set SOURCE=1
8
9  echo LMI $LMI_FILE...
10 echo LMO $LMO_FILE...
11 echo Source Lang $SRC_LANG...
12 echo Target Lang $TRG_LANG...
13 echo Module $MODULE...
14
15 if $SRC_LANG $TRG_LANG (
16   get do
17   ) else (
18   get do
19   )
20
21 160
22 set LPT1=$SRC_LANG
23 set LPT2=$TRG_LANG
24 get do
25
26 160
27 set LPT1=$SRC_LANG
28 set LPT2=$TRG_LANG
29 get do
30
31 160
32 rm -rf /tmp/lucy
33 "D:\LucySoftware\TMConsole\bin\lucy.exe" "D:\LucySoftware\TMConsole\bin\lucy.exe" archive $LMI_FILE $LMO_FILE $SRC_LANG $TRG_LANG $MODULE_nuevo
34
35 rm -rf /tmp/lucy
36 "D:\LucySoftware\TMConsole\bin\lucy.exe" "D:\LucySoftware\TMConsole\bin\lucy.exe" merge $MODULE_nuevo $SRC_LANG $TRG_LANG $MODULE_nuevo $MODULE_nuevo
37
38 rm -rf /tmp/lucy
39 "D:\LucySoftware\TMConsole\bin\lucy.exe" "D:\LucySoftware\TMConsole\bin\lucy.exe" delete $MODULE_nuevo $SRC_LANG
40 "D:\LucySoftware\TMConsole\bin\lucy.exe" "D:\LucySoftware\TMConsole\bin\lucy.exe" delete $MODULE_nuevo $TRG_LANG
41 "D:\LucySoftware\TMConsole\bin\lucy.exe" "D:\LucySoftware\TMConsole\bin\lucy.exe" delete $MODULE_nuevo $SRC_LANG $TRG_LANG
42 "D:\LucySoftware\TMConsole\bin\lucy.exe" "D:\LucySoftware\TMConsole\bin\lucy.exe" delete $MODULE_nuevo $SRC_LANG $TRG_LANG
43 "D:\LucySoftware\TMConsole\bin\lucy.exe" "D:\LucySoftware\TMConsole\bin\lucy.exe" delete $MODULE_nuevo $SRC_LANG $TRG_LANG
44 "D:\LucySoftware\TMConsole\bin\lucy.exe" "D:\LucySoftware\TMConsole\bin\lucy.exe" delete $MODULE_nuevo $SRC_LANG $TRG_LANG
45 "D:\LucySoftware\TMConsole\bin\lucy.exe" "D:\LucySoftware\TMConsole\bin\lucy.exe" delete $MODULE_nuevo $SRC_LANG $TRG_LANG
46 "D:\LucySoftware\TMConsole\bin\lucy.exe" "D:\LucySoftware\TMConsole\bin\lucy.exe" delete $MODULE_nuevo $SRC_LANG $TRG_LANG
47 "D:\LucySoftware\TMConsole\bin\lucy.exe" "D:\LucySoftware\TMConsole\bin\lucy.exe" delete $MODULE_nuevo $SRC_LANG $TRG_LANG
48 "D:\LucySoftware\TMConsole\bin\lucy.exe" "D:\LucySoftware\TMConsole\bin\lucy.exe" delete $MODULE_nuevo $SRC_LANG $TRG_LANG
49 "D:\LucySoftware\TMConsole\bin\lucy.exe" "D:\LucySoftware\TMConsole\bin\lucy.exe" delete $MODULE_nuevo $SRC_LANG $TRG_LANG
50 "D:\LucySoftware\TMConsole\bin\lucy.exe" "D:\LucySoftware\TMConsole\bin\lucy.exe" delete $MODULE_nuevo $SRC_LANG $TRG_LANG
```

Figure 4.23: Script to update translation memories

Finally, the segments are loaded into the Translation Memory (TMX):

<tu tuid="2694">

<tuv lang="es"><seg>01-02-2013 Declaración Tributaria Especial
([@locNote=Stands for 'Declaración Tributaria Especial', use acronym in

target language&&locNoteType=description@@] DTE)</seg></tuv>

<tuv lang="en"><seg>01-02-2013 Special Tax Return ([@@locNote=Stands for 'Declaración Tributaria Especial', use acronym in target language&&locNoteType=description@@] Declaración Tributaria Especial, DTE in Spanish)</seg></tuv>

</tu>

4.5 Environments

For the showcase there are two environments available:

4.5.1 PRE-PRODUCTION

The Pre-production environment has five servers:

- 1) Two servers with the following components or modules (located in Linguaserve's Data centre):
 - a. One ATLAS Real-time application per server.
 - b. One Lucy RBMT with 2 engines per server.
 - c. One FMT Client per server.
 - d. One DataBase Management System in one of the servers.
- 2) One server (located in Linguaserve's Data centre) with:
 - a. The FMT server.
 - b. The Content Editor.
- 3) One server (located at DCU facilities) with the MaTrEx SMT.
- 4) One server (located in Tax Agency's Data centre) with the original contents and the CMS.

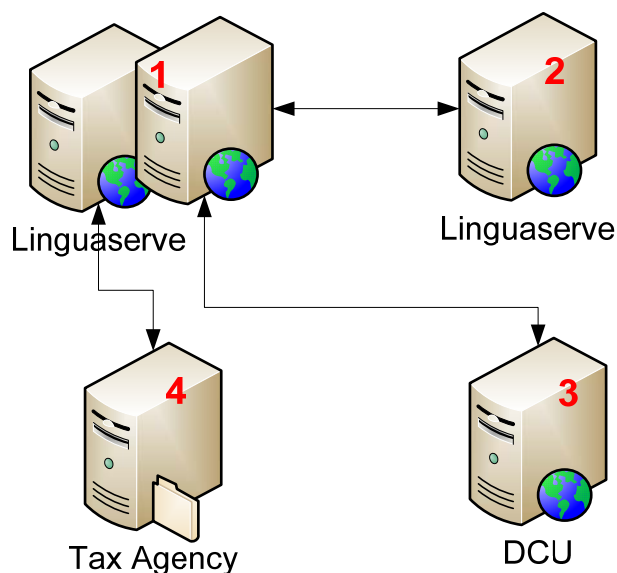


Figure 4.24: Online MT System's Pre-production environment of the Spanish Tax Agency

4.5.2 PRODUCTION

The Production environment counts with five servers too (The system is not connected to DCU's MaTrEx on this environment):

- 1) Two servers with the following components or modules (located in Linguaserve's Data centre):
 - a. One ATLAS Real-time application per server.
 - b. One Lucy RBMT with 2 engines per server.
 - c. One FMT Client per server.
- 2) One server (located in Linguaserve's Data centre) with:
 - a. The FMT server.
 - b. The Content Editor.
- 3) One server (located in Linguaserve's Data centre) with DataBase Management System.
- 4) One server (located in Tax Agency's Data centre) with the original contents and the CMS.

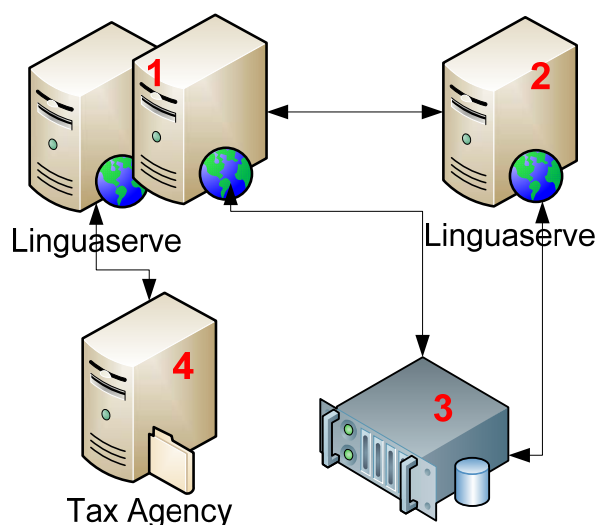


Figure 4.25: Online MT System's Production environment of the Spanish Tax Agency

4.5.2.1 INSTALLATION

The production environment does not have the ITS 2.0 compliant version of the system installed yet. The installation of the new version is pending of the client's approval. Once the client has given the green light, only a few software updates would be necessary. The following steps are required to that end:

- 1) Update of the CMS version, adding the conversion and validation system of HTML5 + ITS 2.0.
- 2) Add the annotation plug-in to CMS.
- 3) Copy the remaining content of the Production environment to the Pre-production environment, convert all the new pages of the website to HTML5 and annotate the new ones.
- 4) Update ATLAS Real-time with the ITS 2.0 module.
- 5) Update the RBMT configuration and filters.
- 6) Upload new necessary ITS 2.0 configurations to ATLAS Real-time.
- 7) Update the FMT scripts to convert from SPT to Linguaserve proprietary tags and vice versa.
- 8) Add the new filter to the Content Editor configuration.

5. LOCALISATION PROJECT IN REAL-TIME SYSTEM WITH ITS 2.0

The objective is the development of the real application of the Real-Time Multilingual Publishing and Translation Systems adapted in the Work Package 4 for a selection of Internationalization Tag Set 2.0 metadata. The sub-objectives are:

A real test of the Systems adapted to ITS 2.0 described above: ATLAS RT, Lucy MT and MaTrex.

Acquire experience and test the migration of an existing system to HTML5 with ITS 2.0 metadata inserted, with a real client: www.agenciatributaria.es.

Test the post-editing methodology and adopt best practices in the use of ITS 2.0.

The team:

During the project, a multidisciplinary team has been set up:

- 1 Technical Manager: Technical support
- 1 Linguistic Team Coordinator: Project management and guidelines
- 1 Terminologist: Terminology management
- 3 Post-editors:
 - 1 Spanish-English PE
 - 1 Spanish-French PE
 - 1 Spanish-German PE

Regarding the post-editors training, the following aspects were considered:

- *Competences development.*

The team involved in the post-editing tasks was formed of young and experienced translators, as, according to the inquiries of the EDI-TA project, the skills needed for these tasks are best found in them. They acquired the ability to locate machine translation errors and instrumental competences to understand the MT's functioning, the correct use of reference material and the comprehension of ITS 2.0 annotations. In addition to this, of course, having linguistic skills, which include an excellent knowledge of the source and target languages, plus communicative competence.

- *Understanding what IS PE and what IS NOT.*

These post-editors have been taught the differences between translating and post-editing. They have received PE guidelines compiled during the EDI-TA project to follow and have understood the importance of the ITS 2.0 annotations (Localization Notes and Localization Quality Issues in particular). The PE team coordinator activated all of them as the project needed the best possible quality.

PE guidelines

• Grammar should be accurate.
• Fix syntactic errors as long as they interfere with the message transferred.
• Ensure that key terminology is correctly translated.
• Spend time researching terms.
• Fix any omissions as long as they interfere with the message transferred.
• Fix morphological errors.

• Fix misspelling errors.
• Fix punctuation errors.
• Edit any offensive, inappropriate or culturally unacceptable information.
• All basic rules regarding spelling still apply.
• Textual standards (cohesion, coherence, standard word order etc.) are not so important.
• Ignore stylistic problems: infinitives vs. nouns in titles, personal vs. Impersonal, active vs. passive, prepositions and articles, word order.
• For tagged formats, ensure all tags are present and in the correct positions.

- *Gaining familiarity with PE guidelines.*

Post-editors gained familiarity with the PE guidelines (established by EDI-TA) during the post-editing, but applying these rules on the go can hinder task efficiency and compromise productivity. The subjectivity of the decisions according to these guidelines made the post-editors come across some dead ends which they claim would have been easier to confront using common sense.

- *Extensive PE hands-on practice on PE MT output.*

Two of the post-editors in this project had experience with post-editing and particularly with Lucy Software's MT output, which saved up part of the training process.

5.1 Project description

5.1.1 PREPARATION TASKS

In order for the post-editing to become faster, there was a previous lexical management task. Of course, the creation and periodical maintenance of a lexicon repository which bears the bilingual vocabulary of the contents that will go through the MT system is essential.

5.1.1.1 GLOSSARY

The creation of the glossary had its starting point in the linguistic analysis of the texts that were going to be translated for a correct vocabulary selection.

Specialized terms

The terminology characteristic of the subject area of the Spanish Tax Agency gains a special importance in the lexicon analysis, as terminology is the one in charge of transferring the contents with precision.

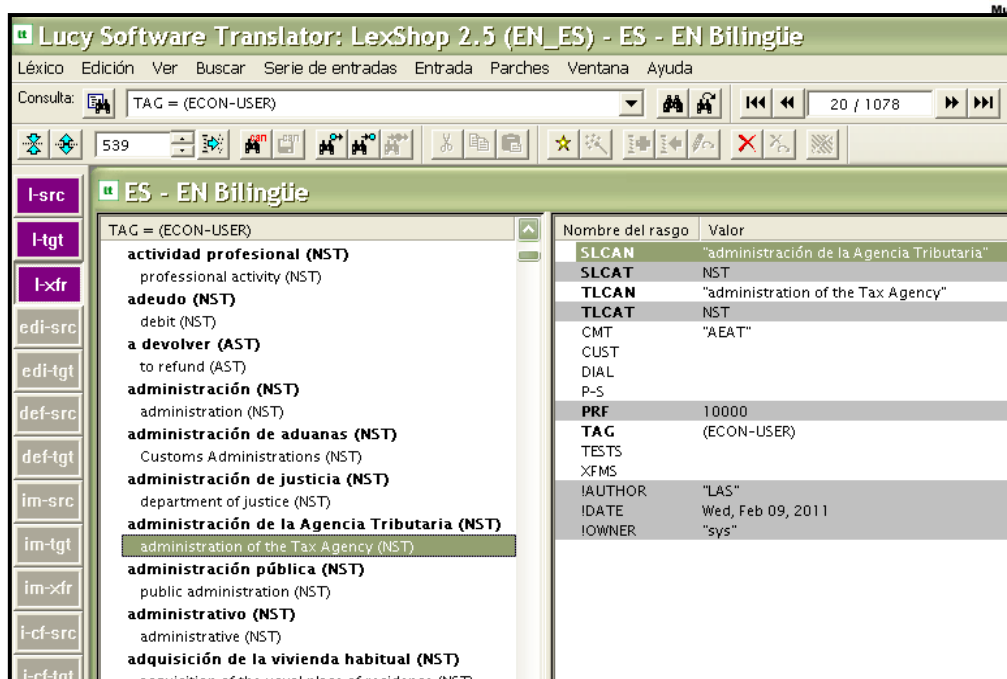


Figure 4.25: Sample of thelexicon categorised in the MT system's dictionary management tool

5.1.1.2 WEBSITE CATEGORIZATION

In order to increase the glossary's effectiveness, each of the entries selected for this project has been tagged with a Domain annotation. In this way, the application chooses the tagged lexicon instead of other similar forms whose meaning is not correct in this context. This action solves most of the terminology problems as the lexicon is approved both formally and semantically in this tool, minimizing the intervention of the post-editor.

Furthermore, this classification is used to link the semantic content of the pages and sections of the Tax Agency's website with the lexicon compiled.

1. General	GV - General Vocabulary
2. Treasury	ECON - Economy & Trade
1.1. Taxation	ECON - Economy & Trade_ ECON-USER - Economy, User Entries
2.1. Taxes	ECON - Economy & Trade_ ECON-USER - Economy, User Entries
2.2.1. Residents	ECON - Economy & Trade_ ECON-USER - Economy, User Entries
2.2.2. Non-residents	ECON - Economy & Trade_ ECON-USER - Economy, User Entries
2.2.3. Companies and professionals	ECON - Economy & Trade_ ECON-USER - Economy, User Entries
2.2.4. Non-resident companies and professionals	ECON - Economy & Trade_ ECON-USER - Economy, User Entries
3.1. Customs	ECON - Economy & Trade_ ECON-USER - Economy, User Entries
4.1. Fraud	ECON - Economy & Trade_ ECON-USER - Economy, User Entries
3. Public Administration	ECON - Economy & Trade
1.1. Electronic administration	ECON - Economy & Trade_ GV - General Vocabulary
3.1.1. Electronic procedures	ECON - Economy & Trade_ GV - General Vocabulary
2.1. International financing	POL - Government, Public Authorities, International Relations
3.1. Public institutions	POL - Government, Public Authorities, International Relations
4.1. Tax policy	POL - Government, Public Authorities, International Relations
5.1. Spanish Tax Scheme	ECON - Economy & Trade_ ECON-POL - Political Economics
3.5.1. National scheme	ECON - Economy & Trade_ ECON-POL - Political Economics
3.5.2. Regional scheme	ECON - Economy & Trade_ ECON-POL - Political Economics
3.5.3. Local scheme	ECON - Economy & Trade_ ECON-POL - Political Economics

Figure 4.26: Subject areas used in the categorisation of the pages and sections of the Spanish Tax Agency's website

5.1.2 EXECUTION IN PRE-PRODUCTION

5.1.2.1 PROJECT DESCRIPTION

- *Source text related processes.*

Parts of the source text and its annotations were previously read by the post-editors and the team coordinator to get to know the kind of text. A very important conclusion reached during the EDI-TA project was that it was necessary to have the source text present during the post-editing in order to check sentences that have been very badly translated and, at a long-term scope, learn how to predict the MT's behaviour.

- *Using Glossaries*

As explained above and below, the use of glossaries has helped improve the MT system in order to gradually improve MT output and speed up the post-editing process and improve its results.

- *Machine translation-related processes.*

The MT output has been read to prove that it would become useful to check the source text in case a restructuring of a sentence was needed or, as a last resort, retranslate it completely in order to save time and keep productivity up.

In some cases, if it was noted that some element of the source text was causing the MT a problem, they would insert a Localization Quality Issue.

- *Target text production and evaluation.*

When producing and evaluating the target post-edited text, the team took into consideration the importance of the client's quality acceptance and tried to balance it with the PE guidelines while dealing with terminology issues, guaranteeing style and terminology consistence and coherence, and following the localization notes by the author. The Localization Notes are an important tool for this purpose, as they are a direct link between the author and the post-editor.

5.1.3 MAINTANANCE TASKS IN PRODUCTION

- Maintenance Glossaries

As it's mentioned above, terminology in a large and long-term post-editing project is an issue that is dealt with little by little, and one of the reasons of this is the maintenance tasks. The EDI-TA project came to the conclusion that terminology was one of the largest issues in post-editing. Solving this would increase the quality of the output and, as a result, the productivity.

The post-editors in this project, apart from post-editing, have gradually been compiling terms that the MT was systematically mistranslating. They paired them up with their correct translation and then they reported them to the terminologist, who reported them to Lucy Software for their implementation in the MT system. As a long-term result, the MT's output will become more and more consistent and will provide a much better translation, so the post-editing process would become less comprehensive, faster (which means productivity) and more accurate and consistent.

5.2 Post-edition applying ITS 2.0

Post-editors also inserted ITS 2.0 annotations. In particular, Localization Quality Issues. These were inserted every time a post-editor considered that a decision made in the source language could have helped improve the MT output.

ES		EN
1	Congregaciones	<i><las_its2 locQualityIssueType="characters" locQualityIssueComment="Al formar parte de una tabla, se ha debido introducir un salto de línea que provoca un error de segmentación, lo cual afecta al orden de las palabras y provocará un error en la memoria de postedición" locQualityIssueSeverity="80" />Congregations</i>
2	Religiosas	Religious

**As it is part of a chart, a line break must have been inserted that causes a segmentation error, which affects the order of the words and will cause an error in the post-editing memory*

These annotations had to be inserted at the beginning of the segment because they do not identify how many words they refer to. It was also important that the post-editors did not insert any full stops inside the annotation's description, as the MT system would split the segment into two and both parts of the annotation would become plain text. They also had to avoid leaving parenthesis enclosing the annotation (in the case that they wanted to tag the content of a parenthesis).

It was found that Transit, one of the CAT system used during the PE process, alerted the post-editor about new tags found in the target test which did not exist in the source text. In spite of that, the software allowed the tags to be saved and exported once the PE was over.

5.2.1 ITS 2.0 PE CONTEXTUAL INFORMATION

Localization Notes can be used to give the post-editors contextual information in order to make decision making faster and better. It could be an annotation tagging a whole page to indicate the context, to give information on the section of the website or on the style and intention of a certain text.

<p its-loc-note="Estos párrafos forman parte del aviso de seguridad de la web de Agencia Tributaria, ámbito informático además de económico y legal" its-loc-note-type="alert">Los diversos intentos de engaño hacen referencia a supuestos reembolsos de impuestos, mediante envíos de comunicaciones masivas por correo electrónico en los que se suplanta la identidad e imagen de la Agencia Estatal de Administración Tributaria, o bien la identidad de sus directivos. Con frecuencia, los intentos de fraude hacen referencia a una supuesta información que imposibilita el envío de la devolución de la campaña de Renta y suplanta la imagen de la Agencia Tributaria.

En estos envíos de correos fraudulentos, conocidos como "phishing", se hace referencia a un reembolso de impuestos inexistente en la que el receptor del e-mail sale beneficiado. Para poder disponer del dinero hay que acceder a una dirección web en la que se deben aportar datos de cuentas bancarias y tarjetas de crédito.</p>

**These paragraphs are part of the security warning of the Tax Agency's website, field of informatics apart from economic and judicial*

Apart from this case, the use of ITS 2.0 annotations such as domain, provenance and language is mainly useful for PE project management and for the MT, as post-editors never get to see these annotations in the PE process. A future improvement could consist in the inclusion of the domain annotation as visible information for the post-editor, as it could become useful during the PE process.

5.2.2 ITS 2.0 PE ACTIVATION RULES

There are certain Activation Rules, defined in EDI-TA, which are used to allow or forbid specific PE behaviours and actions during the PE process. They are classified in various scope groups: language independent rules, language specific rules and text related rules; and some of these rules have MT issues associated to them which disappear thanks to ITS 2.0 annotations.

5.2.2.1 LANGUAGE INDEPENDENT RULES

The language independent rules, established in the EDI-TA project, targeted an issue that Localization Notes are able to solve (terminology clarification in the source language), although there are some issues which are impossible to prevent and must be part of the post-editing work (morphological, phrase structure, linear order, syntax...).

Preventing wrong terminology:

PE without ITS 2.0	PE with ITS 2.0
ES	
<i>Parámetros de liquidación por tipo de entidad.</i>	<i>Parámetros de its-loc-note="En la estadística hace referencia al cálculo de la deuda tributaria, es el conjunto de operaciones que cuantifican la cuantía del impuesto" its-loc-note-type="alert">liquidación por tipo de entidad.</i>
EN	
<i>Settlement parameters by type of organisation.</i>	<i>Assessment parameters by type of organisation.</i>

**In this statistic it's referring to the calculation of the tax debt, it's all the operations that quantify the tax amount*

PE without ITS 2.0	PE with ITS 2.0
ES	
<i>Nota relativa al desarrollo del procedimiento del mecanismo de pagos a proveedores de las EE.LL. (Real Decreto-ley 4/2012, de 24 de febrero) a partir del día 8 de mayo de 2012</i>	<i>Nota relativa al desarrollo del procedimiento del mecanismo de pagos a proveedores de las its-loc-note="Siglas de 'Entidades Locales', traducir y no usar siglas" its-loc-note-type="alert">EE.LL. (Real Decreto-ley 4/2012, de 24 de febrero) a partir del día 8 de mayo de 2012</i>
FR	
<i>Note relative au développement du procédé du mécanisme de paiements aux fournisseurs des EE.LL. (R��el D��cret-loi 4/2012, du 24 f��vrier) �� partir du 8 mai 2012</i>	<i>Note relative au d��veloppement du proc��d�� du m��canisme de paiements aux fournisseurs des Entit��s Locales (R��el D��cret-loi 4/2012, du 24 f��vrier) �� partir du 8 mai 2012</i>

**Acronym for 'Entidades Locales', translate and do not use an acronym*

PE without ITS 2.0	PE with ITS 2.0
--------------------	-----------------

ES	
Programa PADRE	Programa PADRE
DE	
VATER Programm	PADRE Programm

**This is the acronym for 'Programa de Ayuda a la Declaración de la Renta', do not translate*

5.2.2.2 LANGUAGE SPECIFIC RULES

Although the ITS 2.0 annotations are not meant to solve specific language-dependent problems in a specific target language, there are certain clues an editor can include to keep style and terminology consistent and correct throughout the translations. Terminology examples were seen above, here are some stylistic solutions.

Preventing stylistic problems:

PE without ITS 2.0	PE with ITS 2.0
ES	
Sede electrónica, todos los trámites online.	Sede electrónica , todos los trámites online.
FR	
Siège électronique, toutes les démarches en ligne.	Bureau électronique, toutes les démarches en ligne.

**Translate as if it said 'oficina electrónica' and be consistent*

PE without ITS 2.0	PE with ITS 2.0
ES	
si se presenta declaración individual o conjunta monoparental y el sexo del declarante es varón.	Si se presenta declaración individual o conjunta monoparental y el sexo del declarante es varón.
EN	
for single parent individual or joint tax returns wherein the taxpayer is male.	For single parent individual or joint tax returns wherein the registered taxpayer is male.

**Translate as 'declarante registrado' because it is a taxpayer who has already presented a tax return in previous financial years and is already 'registered'*

PE without ITS 2.0	PE with ITS 2.0
ES	

<i>En caso de aplicación del régimen especial de los bienes usados, objetos de arte, antigüedades y objetos de colección, la mención «régimen especial de los bienes usados», «régimen especial de los objetos de arte» o «régimen especial de las antigüedades y objetos de colección».</i>	<i>En caso de aplicación del régimen especial de los bienes usados, objetos de arte, antigüedades y objetos de colección, la mención «régimen especial de los its-loc-note="Traducir como 'de segunda mano'" its-loc-note-type="alert">bienes usados, «régimen especial de los objetos de arte» o «régimen especial de las antigüedades y objetos de colección».</i>
EN	
<i>If the special scheme of second-hand goods, works of art, antiques and collectibles is applied, the reference "special scheme of used goods", "special scheme of works of art" or "special scheme of antiques and collectibles".</i>	<i>If the special scheme of second-hand goods, works of art, antiques and collectibles is applied, the reference "special scheme of second-hand goods", "special scheme of works of art" or "special scheme of antiques and collectibles".</i>

**Translate as if it said 'de segunda mano'*

One of the text related rules activated for the Spanish Tax Agency's website is spending time on terminology research. Localization Notes helped saved research time by explaining certain terms and acronyms, as it is explained above.

The text of the Spanish Tax Agency's website has two very different kinds of segments: the very long ones with several subordinate clauses, and the short titles which do not form a proper and complete sentence. Both cases are very inconvenient for the MT, but there is nothing an ITS 2.0 annotation can do to solve this.

5.3 Evaluation of linguistic resources and PE Efforts

5.3.1 PE EFFORT EVALUATION

5.3.1.1 PE SPEED RATE

In certain kinds of texts (simpler syntax, general vocabulary...) a post-editor can reach +800 post-edited words per hour, according to information by EDI-TA. But the post-editing rates for the Spanish Tax Agency's website is under that cipher due to what has been explained in the text related rules (above) and to the presence of long in-line annotations which, in spite of helping achieve a better post-edited text, slow down the process.

5.3.1.2 SENTENCES SPLIT BY CODE

Although Localization Notes are mostly helpful throughout the post-editing process, they also make sentences slightly less understandable and more cryptic as they have tags with the explanations inserted directly between their words. In cases in which there is more than one annotation per phrase, the post-editor may miss the visual continuity of the sentence, spend too much time rereading it or even leave syntax mistakes from the MT uncorrected.

El valor de adquisición de los bienes y derechos objeto de la declaración especial será válido a efectos de IRPF, its-loc-note="Siglas de 'Impuesto sobre Sociedades', traducir y no usar siglas" its-loc-note-type="alert">IS e its-loc-note="Siglas de 'Impuesto sobre la Renta de No Residentes', traducir y no usar siglas" its-loc-note-type="description">IRNR, a partir de la fecha de presentación de la declaración y realización del ingreso correspondiente.

This problem would not exist if the information in the annotation appeared differently. For example, an annotated item could appear in a different colour in order to inform the post-editor of an existing note regarding that item, and then the post-editor could hover the mouse over the item to reveal the information. Another useful system would be having the annotations from a specific segment displayed in an extra window on the post-editing tool or using abbreviated icons.

5.3.1.3 SLOW CODE ENTRY

In the event of finding an issue related to the source text or to the translation engine during the post-edition, the post-editor would have to insert an ITS 2.0 Localization Quality Issue annotation manually (more conveniently by copy-pasting the code). By doing this, the post-editor might make a typing mistake, or might ruin the html code. Maybe she/he takes too long to insert the localization quality issue, or to even copy and paste it from another text file, and ceases to add annotations for being in a hurry or just out of frustration.

The system could become faster and more efficient if there was an option in the post-editing tool that allowed the user to automatically add the html code with the issue's information.

5.3.1.4 SLOW TERMINOLOGY COMPILING

As the post-editors were compiling terminology while they post-edited, productivity was lower and they sometimes lost track of what they were post-editing at that moment. But, as it has been explained above, this compiling has long-term results and, as this experiment only covers the first steps of the web's post-editing complete project, the results are not verifiable yet.

5.3.2 RBMT: LUCY SOFTWARE

5.3.2.1 SPANISH TO ENGLISH

From the post-editor's experience with other MT systems and this one in particular, the quality of Lucy Software's output was not a big problem. Most of the medium sized sentences only needed some linear ordering, article deletion or preposition correction, unless they contained ambiguous elements, in which case there was further typing to do. The larger sentences, which included subordinate clauses were mostly disordered and needed comprehensive restructuring. The problems which took the most to solve were terminology-related, as most of them involved terminology research.

Post-editing rate: 620 words / hour

5.3.2.2 SPANISH TO FRENCH

For the post-editor of French, the output was acceptable. Both terminology and syntax are not much of a problem for the MT, although terminology research was needed. The Localization Notes were only helpful for understanding certain concepts and unknown acronyms.

Post-editing rate: 685 words / hour

5.3.2.3 SPANISH TO GERMAN

The greatest problem with the German output was the great difference between Spanish and German syntax and word formation. Long subordinate clauses required complete retranslating. Terminology research was also an issue for the German post-editing. The MT got most of the terminology wrong. The Localization Notes were not helpful enough.

Post-editing rate: 600 words / hour

5.3.3 SMT: MATREX

This does not apply as the MT used was Lucy Software's and it is *rule based*. For detailed info see:

http://www.w3.org/TR/2013/WD-mlw-metadata-us-impl-20130307/#Training_for_Statistical_Machine_Translation

6. BUSINESS REPORT AND EXPLOITATION

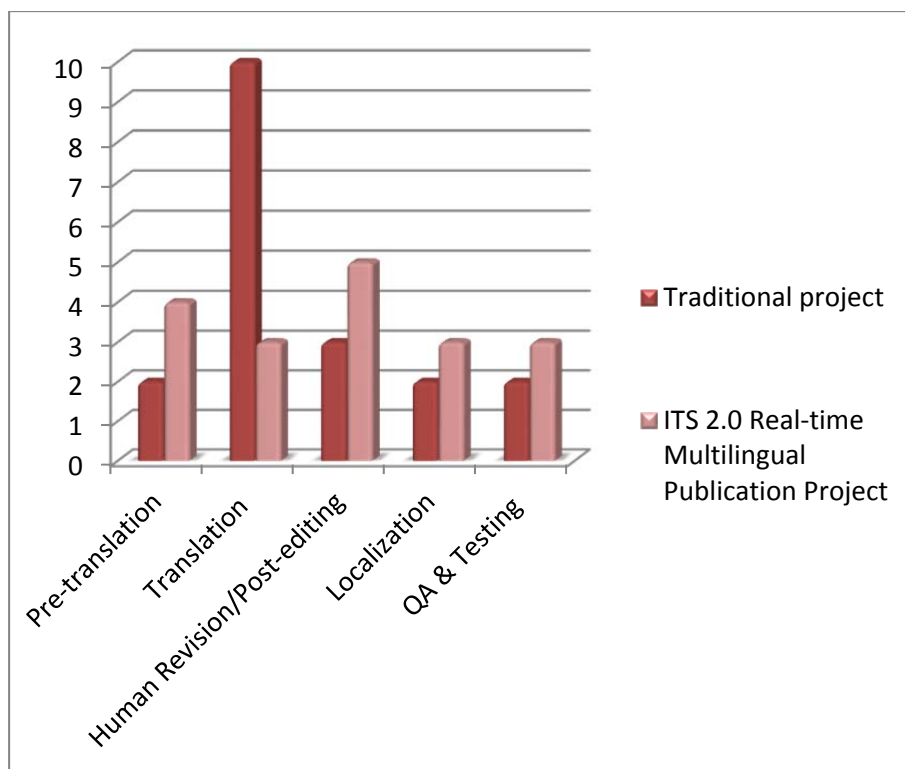
New needs require new approaches. This showcase demonstrates how an Online MT System can automatically:

1. Translate HTML5 documents from an ITS-conformant Web CMS.
2. Control a process depending on its progress, convey information to editors and indicate the state of the content to the user.
3. Provide direct communication context between the client and the post-editor and report quality issues or indications.

This use case demonstration illustrates how ITS, via a Real-time Translation System connected to different MT Service Providers, allows to:

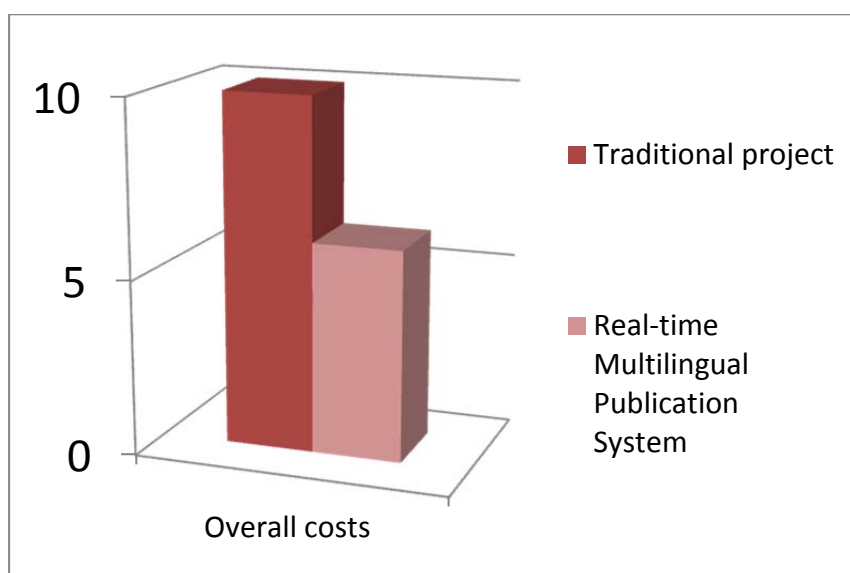
1. Communicate instructions about language, domain and translation, and convey information about the translation to a content editor.
2. Communicate instructions and information to a posteditor regarding the state of a process, and also to inform the user of its status.
3. Communicate the identity of agents that have been involved in the revision and the translation of the content, and to allow translation quality reviewers, to evaluate how the performance of these agents affects the quality of the translation.

This reduction mainly affects the translation tasks, which need other tasks and responsibilities to be increased due to the nature of this kind of solutions. For instance, the effort in project preparation and setting is higher. The cost of MT-Post-editing is higher than the revision of human translation, as well as the technical work and Q&A and Testing. All these increases are absorbed by the translation cost reduction, with a net result of overall project cost reduction at least 40%.



Other tasks like project management vary from implantation to maintenance phases. In the initial phase, which includes implantation, the project management effort is higher, while later on, in the long term, maintenance efforts are lower.

In conclusion, using RTMPS with ITS 2.0 can reduce a multilingual web project at least in 40%:



But of course, a full dedicated implementation of a RTMPS for a client has some constraints concerning size and volume for the project to be profitable. RTMPS platforms shared between several clients makes SMEs with lower project size and volume viable and profitable.

6.1 SWOT analysis

In this section we will analyse the strength, weaknesses, opportunities and threats regarding different business perspectives.

Strengths
Translation Cost Reduction: In Real-Time Multilingual Publishing Systems, using Machine Translation and post-editing reduces costs drastically. The post-editing percentage depends on the quality of the MT output and the nature of the texts. Depending on post-editing percentage, cost reduction increases. For example: MT + 100% post-edited might have a Translation cost reduction of circa -30%. It allows developing the concept of “Quality on-demand”. ITS 2.0 reduces efforts in tasks dependent on the MT system used. For example: instead of codifying in each different MT systems the lexicon excluded from translation, this information is coded with ITS 2.0 and used by the all the different MT Systems.
Management cost reduction: Although the implantation and setup have higher costs, they are much lower during the subsequent maintenance periods (-60%-80%). ITS 2.0 allows a more effective fine-grained coordination and management thanks to having the client’s requirements and indications reach post-editors by using metadata.
Increase of Translation Control: Control by the client over the quality or the customization of the translation is external to the HTML5 content (glossaries, MT rules, RTMPS parameters...). ITS 2.0 allows a more dynamic and precise control internally to the HTML5 content.
Non-invasive technology: clients do not need to install any application.
Delivery time: Real-time or fast post-edition.

Weaknesses
Viability: Differences on MT system outputs. Highly dependent on the language combination. E.g. Difficulty for post-editing differs from diverse language pairs: those typologically close between each other (ES<>FR, ES<>PT, ES<>CA, ES<>GL) are easier, those farther (ES<>EN, ES<>EU, ES<>DE, or FR>RU) slightly difficult, and the farthest (ES<>ZH, ES<>JA) certainly more difficult.
Recent MT approaches (Hybrids, vertical sectors, user customization...) will help to overcome this weakness. Quick Multilingual Publishing Systems can be used with human translation, being just “Quick” instead of being “Real-time”.

Opportunities
Volume: Affordable for high or very high volume web sites. E.g. sites with more than 1 million words/language.
Update: Affordable for high or very high frequency and/or volume update web sites. E.g. sites with several updates per day - Profitability Websites with a very high update frequency.

Technological complexity: Easier for Complex back-office web sites to operate in the final HTML5, and not with the systems that produce it. E.g. virtual offices that “paint” the HTML content from different information systems.

SMEs: In-house installations are rather for sites with over 1 million words and frequently updated, but with shared RTMPS infrastructures, Small and Medium companies and organizations benefit from the low price and low technical effort and complexity.

Recruitment of young translators for new jobs. E.g. set up and project preparation, post-editing, terminology, MT customization...

New methodologies and methods. E.g. EDI-TA methodology and training.

New Language Technologies: Pre-editing and post-editing tools.

New standards: Interoperability and readiness information.

Threats

Real-time performance: Performance of the Multilingual Publishing System and the online Machine Translation all together must stay within the parameters for Internet publication (less than 500 milliseconds). Any additional feature or MT system needs to take this into account.

Security levels: RTMPS must consider security features, especially in the shared application service provider mode for SMEs.

Robustness: In-house hosting needs stable and robust systems that provide a service level agreement >99.9%.

Control: The client might lose control of the translation on the HTML. This can be solved using ITS 2.0 adequately.

Methodologies and tools: ITS 2.0 and HTML5 compliant pre-editing and post-editing methodologies and tools are needed.

Annotation of source content: This is probably the greatest threat, convincing content creators about the benefits of annotating the source content and providing them with suitable tools for automatic, semiautomatic, and manual annotation in HTML5.

6.2 Opportunities rise from needs

As we can see, this use case shows real and promising opportunities for different sectors to move forward in the Multilingual Information and Communication Technological innovation for economic growth and quality of life in Europe. Apart from more obvious applications of the presented use case (like corporate news and publications), we can see other brilliant opportunities.

6.2.1 SMALL AND MEDIUM COMPANIES

Standardization reduces SMEs' barriers in technology reuse and provides ready-to-use methods for innovation. In this use case, in-house "exclusive" installation of Real-time Multilingual Publishing and Translation Systems are thought for big web sites and requires a budget and technical staff, but ASP (application service provider) "shared" installation brings the opportunity for Small and Medium Enterprises and Organizations to access cutting edge technological solutions:

- At a reasonable affordable price
- Without much technological internal expertise required
- Without substantial modification of their technology.
- With control over the process within the content thanks to ITS 2.0

It would allow SMEs to compete in a multilingual community like Europe with other big companies and other markets outside Europe, bringing the unified European e-market closer to smaller organizations. It also makes it easier to comply with some multilingual national and international laws which make it necessary to have different languages, now available at a reasonable price and technical effort (like in Spain with its co-official languages).

6.2.2 E-COMMERCE

E-Commerce web sites are especially susceptible to the benefits of ITS 2.0 and Real-time Multilingual Publishing and Translation Systems. Their characteristics indicate e-Commerce as a primarily target of what we have shown in this use case:

- Very high volume
- Daily content update and rotation
- Immediate time to market (real-time)
- Much lower content management (updates detection and processing)
- Short texts and repetitive descriptions
 - Which are better for MT
 - Which are better for quick post-editing
- Content source independent:
 - HTML from several CMS or ERPs
 - Content from other external applications (banking, pricing, reservations...)

For example, an online shop that offers thousands of products and updates hundreds of them every day.

6.2.3 WEB 2.0 AND WEB 3.0

In the last few years, social media, user content created, and semantic webs are more and more important for the IT society. The amounts of contents, the immediacy and the used information systems make it very difficult to use multilingualism broadly.

RTMPS and online translation system can provide both the technical and the linguistic solution (even using GIST translation) to be optimistic about multilingualweb approaches.

6.2.4 ONLINE PUBLIC SECTOR

The public sector usually produces and publishes enormous amounts of content, and more and more of these contents need to be available to citizens in several languages, in Europe and outside Europe, which rises good business opportunities for ever more urgent needs.

6.2.5 COMPLEX BACK-OFFICE WEB SITES

The evolution of web sites where the origin of content is not a single CMS or information system anymore, but several interoperable information systems connected that deploy html content in real-time needs real-time solutions at the front html page, and not in the back offices (interoperability CMS/TMS, for instance). The use case we have shown treats the localization process from an HTML level and is transparent to all the back-office information systems which produce this HTML.

6.2.6 NEW JOBS FOR YOUNG TRANSLATORS

This use case also involves shifting gears by applying new tasks and methodologies. Thus, young translation and information technology professionals can benefit from it, gaining access to new jobs, where capabilities and profiles require new people.

For example, here are broadly applied new translation tasks:

- Textual analysis and definition of post-editing rules.
- Terminological analysis, management and codification.
- Post-editing of Machine Translation. In this way, Post-editing methodologies (like EDI-TA) also use ITS 2.0 metadata to help people:
 - Providing contextual information (Language information, Domain, Provenance).
 - Containing activation PE rules, Translate, Localization Note, Localization Quality Issue).
- MT error reporting, improvement and customization, using also Localization Quality Issue, for instance.

But also, new localization (technical) tasks:

- Prior programming HTML localization.
- Management of binary files for localization.
- Proactive recurrent testing and fixing.

6.3 Final recommendations and best practices

6.3.1 ORIENTED TO REAL-TIME MULTILINGUAL PUBLICATION SYSTEMS (RTMPS)

- Use the [translate data category](#) and the [language information data category](#) to inform the RTMPS of precisely which sentences or sentence fragments should or should not be translated and which is the source language.
 - *Benefit*: Allows the user to automatically block the machine translation of certain parts of the Web page that are not required to be translated or must not be machine translated because of its difficulty or provenance, i.e. a technical essay or constitutional laws.
 - *Benefit*: Automatically avoids the machine translation of parts of the Web page that are in various languages and must remain that way, i.e. a language selector.
 - *Benefit*: Automatically specifies to the RTMPS the source language of the text and whether it applies to the whole text or not.
- Use the [domain data category](#) to inform the RTMPS, at a paragraph, sentence or word level, of the appropriate training corpora or glossary (depending on the MT System) that should be used on the translation by the MT Systems.
 - *Benefit*: Improves the accuracy and quality of the machine translation.
- The division into sections policy to harnessing the cache system and the validation process, seems to not be very appropriate when dealing with ITS 2.0:
 - The latency reduction can only be perceived in the long-term (The Telefónica cache system solves this problem).
 - There can be a mixture of different languages in the same document which is not usually a best practice.
 - Causes problems when connected to an ITS 2.0 compliant MT System:
 - Mal-formed HTML sent to the MT and thus errors in the DOM generation.
 - The MT should handle the translation-related data categories and leave the rest to the RTMPS.
- Most of the MT Systems do not translate attributes except for those known as translatable attributes which are often hardcoded as such. If we are using a non ITS 2.0 compliant MT System, the only way to translate attributes that the MT does not recognise as translatable, is by mapping those to the ones that are translatable. The restriction here is that this process is carried out using regular expressions and has the limitation that when the element or node has an *id* attribute, the mapping is not possible because the reference is lost, for instance if we want to insert a new *id* as a reference to use with the regular expressions:

WRONG

- `<area id="area" title="this is a text">...</area>` would replace the *id* with the new one assigned losing the reference in the document.

RIGHT

- `<area title="this is a text">...</area>` would turn into `<img id="area_las_translate" title="this is a text">` in order to be translated.

6.3.2 ORIENTED TO ANNOTATORS

- The conversion of the [localization note data categories](#) and the [localization quality issue data categories](#) to SPT can cause problems if the SPT is located in the middle of a sentence (normally when it is inside an inline element), this is due to the languages where word order can be switched with

regard to the source language. For instance, the RBMT System when translating the sentence considers the SPT as a constant (not translatable), so if the position of the SPT is changed during the translation process, afterwards during the reversion to HTML, the reference of the data category and its node owner can be lost, or the data category can end up in the wrong location. The recommendation is not to insert these data category inside inline elements and use block elements referring to the portion of text as much as possible. The SPT when loses the reference can be left in the content without being removed, this problem is solved with regular expressions.

WRONG

- o **English:** <p>On some languages the order of the words can be switched when SPT translating.</p>
- o **Spanish:** <p>En algunos idiomas cuando se SPT traduce el orden de las palabras puede cambiar.</p>

RIGHT

- o **English:** <p its-loc-note="Translating changes position" its-loc-note-type="description">SPT On some languages the order of words can be switched when translating.</p>
- o **Spanish:** <p its-loc-note="Translating changes position" its-loc-note-type="description">SPT En algunos idiomas cuando se traduce el orden de las palabras puede cambiar.</p>
- Related the previous point it should be remarked that in addition when the number of words in the target language are less than in the source language some inline mark-up can be removed from the content, or if there more words the inline mark-up can be misplaced, just like in the above example.
- It is important to highlight that when annotating the [localization note data categories](#) and the [localization quality issue data categories](#) that the content of the attributes *locNote* and *locQualityIssueComment* must not include a segmentation character such as point, colon, question mark, etc...
- Normally a CMS does not allow inserting HTML in certain fields such as in the titles. ITS 2.0 solves this problem by using [global rules](#) with selectors to those elements or attributes, for the rest of cases, the [translate data category](#) can be inserted locally by selecting the text which has to be left as is and use the Tiny editor plug-in of the CMS. The text can vary from a single word to a whole sentence. Nouns, company and other entities' names, brands, some toponyms, products' names, abbreviations, acronyms are usually not translatable, yet always depending on the target language.
- If the language of a certain content of the document is different to the main language of the document, it is recommended to declare it in the correspondent element to make it untranslatable, for instance:

The following code:

```
<html lang="es">...<p lang="en">Winston Churchill - Success is not final, failure is not fatal: it is the courage to continue that counts</p>
```

It would be:

```
<html lang="es">...<p translate="no">Winston Churchill - Success is not final, failure is not fatal: it is the courage to continue that counts</p>
```

6.3.3 ORIENTED TO POST-EDITING

- Use the [localization note data category](#) to provide the editor with the necessary information to review the text in order to help him with the disambiguation and to improve the quality and accuracy of the revision.
 - *Benefit:* Help to improve the accuracy and quality of the reviewed machine translation after the post-edition process.
- The content generated by the RBMT System for being reviewed in this showcase has been revised using a CAT Tool because of lacking an appropriate Web application which would have given not only real-time update of the translation memory, but also a friendly and intuitive interface where the editor can access to the information provided by the data categories and manage new ones with ease.

6.3.4 ORIENTED TO CUSTOMER OR CONSUMER

- Use the [provenance data category](#) to inform the translation consumers of how the HTML content was translated and subsequently revised.
 - *Benefit:* Provides information on the identity of the translation and revision agents whether they are a person, an organization or a piece of software.

6.3.5 ORIENTED TO QA

- Use the [localization quality issue data category](#) to pass feedback on the severity of the errors detected during a language-oriented quality assurance (QA) process, to the translation consumer.
 - *Benefit:* Allows the content author and the consumer of the translations to have a better understanding of the common errors made during the translation and revision processes and to help them in making decisions for the improvement of these quality processes.

6.4 Extensions: Readiness

An important ITS 2.0 extension and part of the showcase implementation is the [readiness data category](#) which is a combination of the previous data categories: processTrigger, legalStatus, processState, proofreadingState and revisionState. This data category is out of scope of the current version of the specification, and its definition and implementation are not complete and stable yet.

6.4.1 DEFINITION

Readiness is an ITS 2.0 extension data category proposal.

Linguaserve and Cocomore are applying Readiness in both use cases involved:

- Applied in CMS-TMS showcase (WP3).
- Applicability in Online Translation system (WP4).
- It indicates the readiness of a document for submission to L10n processes or provides an estimate of when it will be ready for a particular process.
- It can be used in expert systems for automatic processing.

ITS2.0 should be able to indicate the readiness of an element for its submission to different processes or

provide an estimate of when and element will be ready for a particular process.

ITS2.0 should be able to indicate the relative priority elements should be subjected to when submitted to a process.

ITS2.0 should be able to indicate an expectation of when a specific process should be completed for an element.

ITS2.0 should be able to specify if an element previously submitted to a process has subsequently been revised and therefore needs to be re-submitted to that process.

Benefits:

- It provides version control.
- It allows the editor to establish a planning based on priorities and deadlines concerning the different processes that has assigned.

6.4.2 IMPLEMENTATION

According to the last definition the different parameters of the metadata category are:

- A required **ready-to-process** attribute. It contains the value or values of the process requested for the content. One or more comma separated strings, each one identifying a different process. The different values are:
 - *posteditQA*: The content must be postedited.
- An optional **process-ref**. A pointer to an external set of process type definitions used for ready-to-process if the default value set is not used.
- An optional **ready-at** attribute. Defines the time the content is ready for the process.
- An optional **revised** (yes|no). Indicates if this is a different version of content that was previously marked as ready for the declared process.
- An optional **priority** attribute that indicates the priority of the content for the process. Values 1/4|2/4|3/4|4/4. The lower priority would be 1/4 (1 out of 4 that is the total number of priorities), the highest would be 4/4 (urgent)
- An optional **complete-by** attribute that provides a target date-time for completing the process

File 1: HTML5

```
<!DOCTYPE html>
<html lang="en">
  <head>
    <meta charset=utf-8>
    <title>ITS 2.0 - Java Hello World!</title>
    <link href="Rules.xml" rel="its-rules"/>
  </meta>
</head>
<body>
  <section>
    <span id="languageSelector">
      <ul>
        <li><a href="/en/index.html" its-translate="no">English</a></li>
        <li><a href="/es/index.html" its-translate="no">Castellano</a></li>
      </ul>
    </span>
  </section>
```

```

<section>
  <p><span>
    Here it's the code of a "Hello World!" application in Java
  </span></p>
  <p><code its-translate="no">
    class HelloWorldApp {
      public static void main(String[] args) {
        System.out.println("Hello World!"); // Display the string.
      }
    }
  </code></p>
  <p><span>
    The "Hello World!" application consists of two primary components: the
    HelloWorldApp class definition, and the main method. Now compile and run it!
  </span></p>
</section>
</body>
</html>

```

File 2: XML Rules

```

<its:rules version="2.0" xmlns:its="http://www.w3.org/2005/11/its"
xmlns:h="http://www.w3.org/1999/xhtml">
  <itsx:readinessRule ready-to-process="posteditQA" revised="no" priority="2/3"
complete-by="01/10/2012 13:00:00.000 UTC" selector="//h:section/h:p/h:span"
xmlns:itsx="http://www.w3.org/2008/12/its-extensions"/>
</its:rules>

```

This example indicates that the content of the span tags must be postedited after the machine translation, the priority of the process is medium and the deadline is on 01/10/2012 13:00:00.000 UTC, the original text is not yet revised.

6.5 Beyond 2013: future work

We consider the tasks performed and the results achieved as only a starting point for a great deal of work in the future. Real-time Multilingual Publishing and Translation System together with ITS 2.0 and HTML5 open a door for new ways of approaching the multilingual web. Some of the upcoming work to be done can be listed as follows:

6.5.1 SOURCE AND TARGET LANGUAGE TOOLS

- **Pre-editing creation /extension of tools**
 - Full HTML5 compliance and ITS2.0 annotation facilities.
 - Writing tools for content quality, and controlled language for post-editing output adaptation.
- **Post-editing creation /extension of tools**
 - Exploiting specific PE functionalities and ITS 2.0 assistance and viewing functions for post-editors.
 - Specific language-dependent and language-independent post-editing rules and functionalities.
 - ITS 2.0 assistance, and specific interfaces and functions.

6.5.2 LINGUASERVE CHALLENGES

Linguaserve foresees many challenges and future work to be done. All this needs to be prioritized in order to achieve it with the realistic resources and possibilities of the company.

For example, for RTMPS using ITS 2.0 Pre-editing:

- Exploring training and methodology in content creation applying ITS 2.0.
- Advising the clients about ITS 2.0 and how to use it.
- Participating in automatic and semiautomatic annotation tools creation and exploitation.

For example for RTMPS using ITS 2.0 Post-editing

- Exploring best practices using ITS 2.0 data categories.
- ITS 2.0 capabilities, usage, and training kits.
- Improving EDI-TA PE methodology and training.
- Contextual, activation, and identification rules using ITS 2.0.
- Specific language/system/client dependent and independent PE rules and training.
- Requirements from EDI-TA and showcase experience.

On top of this, from the business perspectives ITS 2.0 and RTMPS and online MT awareness is needed among different actors:

- Customers: Qualitative and quantitative advantages.
- CMS developers and integrators: Product adaptations and extensions.
- Content creators: Training, pre-editing tools and best practices.
- LSPs: Compliant tools and best practices.

6.5.3 LUCY CHALLENGES

- **Terminology and Disambiguation:**
 - o PE tools and MT systems should be ready to handle the marked terms and their associated information.
- **MT Confidence:**
 - o Need reliable MT metrics (e.g. unknown terms) and quality measuring systems.
- **External Resource:**
 - o Need CAT and PE tools to handle references to external resources to be translated.
 - o E.g.: audio & video files would need specific pre-and post-processing.

6.5.4 DCU CHALLENGES

- **Quality and performance:**
 - o The biggest challenge for the year 2014 and the ones to come is to improve both the quality and the performance of MaTrEx in terms of response's time, which is not acceptable for a Real-time System, in order to be used not only in a Pre-production but also in an Exploitation environment.
 - o Ability to handle deep nested HTML structures as well as overlapping tags; Current functionality overgeneralizes.
 - o Automate identification of file type: HTML / XML / XLIFF / simple segments.

7. GLOSSARY OF TERMS AND ACRONYMS

Term/Acronym	Definition
ITS	The Internationalization Tag Set (ITS) is a set of attributes and elements designed to provide internationalization and localization support in HTML5 and XML documents
W3C	The World Wide Web Consortium (W3C) is the main international standards organization for the World Wide Web (abbreviated WWW or W3)
RTMPS	Real-time Multilingual Publishing System
MT	Machine Translation
SMT	Statistical Machine Translation
TM	Translation Memory
RBMT	Rule-Based Machine Translation
DCU	Dublin City University
MaTrEx	Machine Translation using Examples, SMT based on Moses that can handle ITS 2.0 metadata
CMS	Content Management System
SPT	Special Plain Text, format used to transform ITS mark-up into plain text understandable by MT Systems
LT	Language Technology
FMT	File Management Tool
LMI/LMO	Input and output files (respectively) generated by Lucy MT System
SSL	Secure Socket Layer is a cryptographic protocol that provides communication security over the Internet

Term/Acronym	Definition
PHP	P (Personal Home Page) H (Hypertext) P (Pre-processor) is a programming Language
DBMS	Data Base Management System

8. REFERENCES

HTML5

Robin Berjon et al. HTML5. W3C Candidate Recommendation 17 December 2012. Available at <http://www.w3.org/TR/2012/CR-html5-20121217/>. The latest version of [HTML5](http://www.w3.org/TR/html5/) is available at <http://www.w3.org/TR/html5/>.

ITS 2.0 standard

Latest ITS 2.0 [Working Draft](#)

Language Technology

Hans Uszkoreit [Language Technology A First Overview](#).

Linguistics

[Linguistic Annotation](#).

MT Systems

[Statistical Machine Translation](#).

Miles Osborne, [MT History and Rule-based systems](#).

Multilingual Web

<http://www.w3.org/International/multilingualweb/lt/>.

<http://www.multilingualweb.eu/>.

Provenance

[Provenance data model](#).

Language Information

[Tags for Identifying Languages](#).

Security

http://cert.inteco.es/extfrontinteco/img/File/intecocert/EstudiosInformes/cert_inf_seguridad_analisis_trafico_wireshark.pdf

Linguaserve

www.linguaserve.com

Dublin City University

<http://nclt.dcu.ie/mt/>

Lucy Software

www.Lucy Software.com

Spanish Tax Office (Agencia Tributaria)

www.agenciatributaria.es