Implementing ITS 2.0 for post-editing purposes

Celia Rico
Universidad Europea
Campus V. de Odón
28670 Madrid
celia.rico@uem.es

Pedro L. Díez Orzas
Linguaserve I.S. S.A.
Seminario de Nobles, 4
28015 Madrid
pedro.diez@linguaserve.com

Felix Sasaki
DFKI / W3C fellow
Alt-Moabit 91
10559 Berlin
fsasaki@w3.org

Abstract

This paper presents part of the work carried out in EDI-TA, in the context of the project MultilingualWeb-LT[1]. The aim is to implement the Internationalization Tag Set 2.0 (ITS 2.0) in an MT context for post-editing purposes. After a brief review of MultilingualWeb-LT's main objectives and a presentation of ITS 2.0 major features, our paper will concentrate on the description of an Online MT showcase. Here ITS 2.0 information, so called "data categories", are tested in a post-editing scenario.

1 Introduction

MultilingualWeb-LT aims at defining the Internationalization Tag Set 2.0 (ITS 2.0), that is: "meta-data for web content (mainly HTML5) and deep Web content that facilitates its interaction with multilingual technologies and localization processes"[2]. The ITS 2.0 specification identifies concepts termed "data categories" (such as "Translate", "Localization note", "Directionality") [3] that are important for internationalization and localization. ITS 2.0 also provides implementations of these data categories among others as a set of markup attributes.

ITS 2.0 applies to the whole process of localization and has a direct impact in the use of MT as "data categories support the different automated backend processes of this service type, thereby adding substantial value to the service results as well as possible subsequent services" (ITS 2.0, 2013). One of such services is MT post-editing (PE). In this context, EDI-TA was designed as a subproject of MultilingualWeb-LT with the aim, among others, of testing the contributions of ITS 2.0 to PE. The broad objectives of EDI-TA are as follows:

- Contribute to defining metadata suitable for post-editing purposes.
- Test the contribution of metadata in order to improve post-editing processes.
- Define a practical methodology for post-editing between distant languages pairs, namely, Spanish into English, French and Basque, and from English into Spanish.
- Suggest improvements in the MT system so as to optimize the output for post-editing specific purposes.
- Show the feasibility and cost reduction of implementing post-editing in a real scenario.
- Identify functions to improve post-editing tools.
- Define a methodology for training post-editors in the following language pairs: ES, EN, FR and EU.

These are certainly ambitious objectives set out with the purpose of comprehensively analysing the different aspects usually involved in a PE project. The present chapter will only concentrate in the description of work carried out towards implementing ITS 2.0 metadata for PE. Other findings have been reported in Rico and Díez Orzas (2013a and 2013b) and are duly referred to when necessary.

[1] MutlingualWeb-LT (LT-Web, European Comission 7FP, Language Technologies, Grant Agreement No. 287815)

[2] http://www.w3.org/International/multilingualweb/lt/

[3] http://www.w3.org/TR/its20/#datacategory-description

2 ITS 2.0 data categories for PE purposes

From the set of the 19 data categories in ITS 2.0, six were identified in this use case for demonstration and for PE purposes. In the following sections we will discuss these in detail.

2.1 Data category: Translate

Definition. The *Translate* data category expresses information about whether the content of an element or attribute should be translated or not. The values of this data category are "yes" (translatable) or "no" (not translatable).

Use for PE purposes. Informing the post-editor of precisely which sentences or sentence fragments should or should not be translated. Viewing not translatable content may help to adjust the possible implications of not automatically translating an element.

2.2 Data category: Localization note

Definition. The *Localization Note* data category is used to communicate notes to localizers about a particular item of content (as a "description" or "alert").

Use for PE purposes. Providing post-editors with the necessary information to review the text in order to help them disambiguate and improve the quality and accuracy of the revision. Some specific notes for post-editing within the *locNote* element are UTS Ratings (Utility, Time and Sentiment) (O'Brien, 2012; Rico, 2012):

- Utility (relative importance of the functionality of the translated content).
- Delivery Time (speed with which the translation is required).
- Sentiment (importance on brand image).
- Expiration level.

2.3 Data category: Language information

Definition. The element *Language Information* is used to express the language of a given piece of content.

Use for PE purposes. The "Language Information" data category allows to point to part of content in a language different from the rest, which could require MT and post-editing for an specific language pair. This way, task assignment can be automatically performed.

2.4 Data category: Domain

Definition. The *Domain* data category is used to identify the topic or subject of a given content.

Use for PE purposes. It enables automatic selection of MT terminology, post-editor selection, and is a key to content disambiguation.

2.5 Data category: Provenance

Definition. It is used to communicate the identity of agents that have been involved in the translation of the content or the revision of the translated content. This data category offers three types of information. First, it allows for the identification of translation agents. Second, it allows for the identification of revision agents. Third, if provenance information is needed that includes temporal or sequence information about translation processes (e.g. multiple revision cycles) or requires agents that support a wider range of activities, the data category offers a mechanism to refer to external provenance information.

Use for PE purposes. It allows post-editors to assess how the performance of these agents may impact the quality of the translation. Translation and translation revision agents can be identified as a person, a piece of software or an organization that has been involved in providing a translation that resulted in the selected content.

2.6 Data category: Localization Quality Issue

Definition. The Localization Quality Issue data category is used to express information related to localization quality assessment tasks.

Use for PE purposes. It allows post-editors to detect possible localization quality issues, as follows:

- Terminology. An incorrect term or a term from the wrong domain was used or terms are used inconsistently.
- Mistranslation. The content of the target mistranslates the content of the source.
- Omission. Necessary text has been omitted from the localization or source.
- Untranslated. Content that should have been translated was left untranslated.
- Addition. The translated text contains inappropriate additions.

- Duplication. Content has been duplicated improperly.
- Inconsistency .The text is inconsistent with itself.
- Grammar. The text contains a grammatical error (including errors of syntax and morphology).
- Legal. The text is legally problematic (e.g., it is specific to the wrong legal system).
- Register. The text is written in the wrong linguistic register, uses slang or other language variants inappropriate to the text.
- Locale specific content. The localization contains content that does not apply to the locale for which it was prepared.
- Locale violation. Text violates norms for the intended locale.
- Style. The text contains stylistic errors.
- Characters. The text contains characters that are garbled or incorrect or that are not used in the language in which the content appears.
- Misspelling. The text contains a misspelling.
- Typographical. The text has typographical errors such as omitted/incorrect punctuation, incorrect capitalization, etc.
- Formatting. The text is formatted incorrectly.
- Inconsistent entities. The source and target text contain different named entities (dates, times, place names, individual names, etc.).
- Numbers. Numbers are inconsistent between source and target.
- Markup. There is an issue related to markup or a mismatch in markup between source and target.
- Pattern problems. The text fails to match a pattern that defines allowable content (or matches one that defines non-allowable content).
- White space. There is a mismatch in whitespace between source and target content.
- Internationalization. There is an issue related to the internationalization of content.
- Length. There is a significant difference in source and target length.
- Uncategorized. The issue has not been categorized or cannot be categorized.
- Other. Any issue that cannot be assigned to any values listed above.

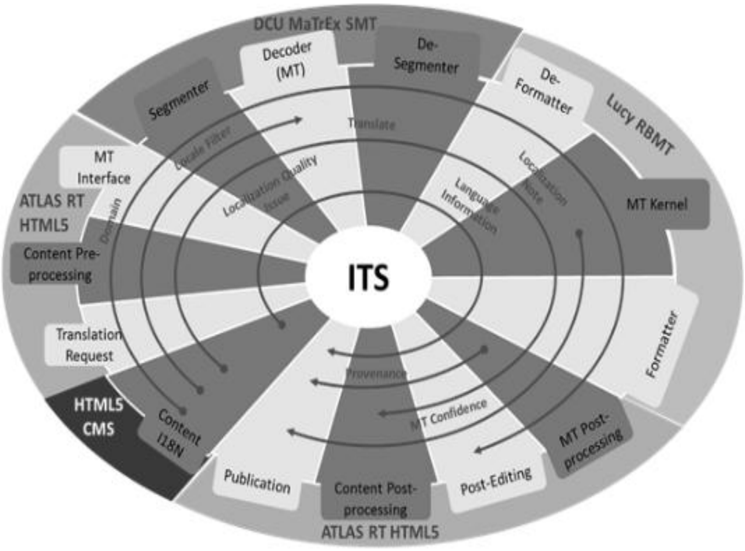
3 Online MT Showcase

The objective of the Online MT System ITS 2.0 showcase is “to define and demonstrate LT-Web metadata in HTML, applying Real Time Multilingual Publishing Systems (RTMPS), using both Rule Base and Statistical Machine Translation”, in industrial showcase with the Spanish Tax Office[4]. What follows here is part of the broader online MT system showcase as designed and conducted by Linguaserve[5] in the context of MultilingualWeb-LT project, using ATLAS Real Time.

3.1 Annotation strategy, processing and output

Figure 1 shows *ITS 2.0 ellipse* with the lifecycle of the different data categories used in the showcase and the different systems involved in their processing. Each data category goes through three stages: a) annotation; b) processing; and c) output. As an illustration we will see how the data category “localization note” is processed. For complete details and a comprehensive description see Nieto et al (2013).

Figure 1. ITS 2.0 Online MT ellipse

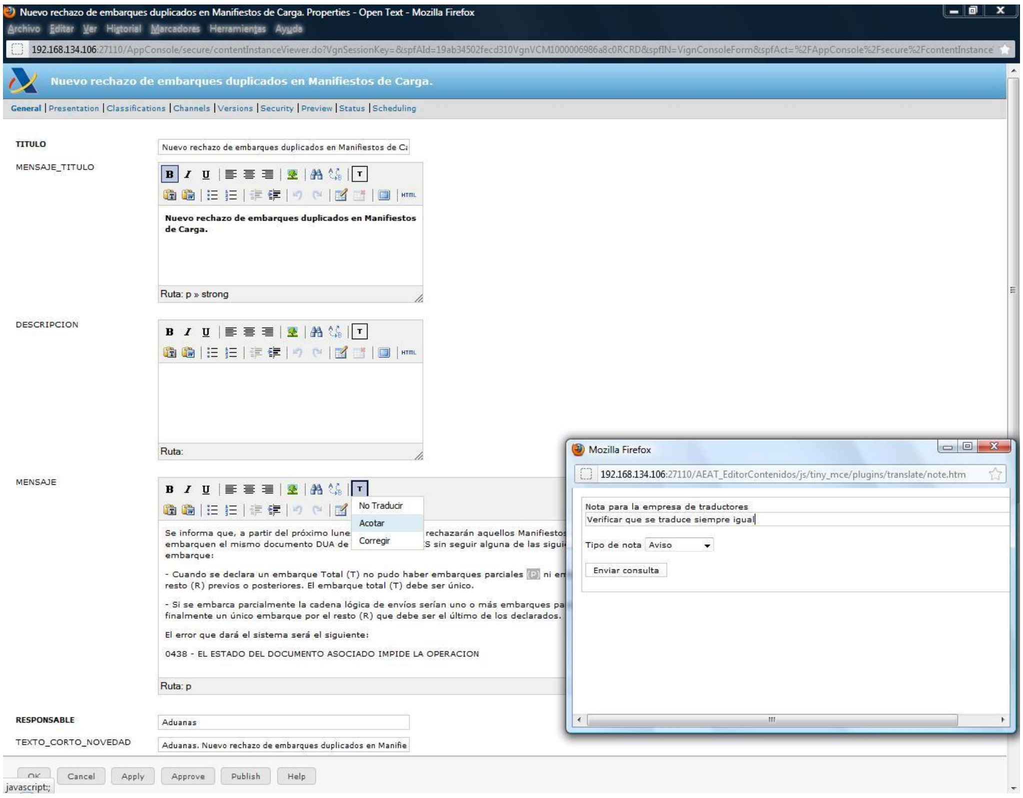


[4] http://www.agenciatributaria.es w3.org/International/multilingualweb/ lt/wiki/WP4
[5] http://www.linguaserve.com/

Data category: Localization note

- Stage 1. Annotation.

ITS 2.0 metadata is embedded into the HTML code. In the example below for the field MENSAJE (message), the editor selects the text where the note applies and then clicks on "Acotar" (annotate) that opens a pop-up window where the editor writes the text of the note and selects the type, and finally clicks on "Enviar consulta" (send query).



Nuevo rechazo de embarques duplicados en Manifiestos de Carga. Properties - Open Text - Mozilla Firefox

192.168.134.106:27110/AppConsole/secure/contentInstanceViewer.do?VgnSessionKey=&spfAId=19ab34502fecd310VgnVCM1000006986a8c0RCRD&spfIN=VignConsoleForm&spfAct=%2FAppConsole%2Fsecure%2FcontentInstance

Nuevo rechazo de embarques duplicados en Manifiestos de Carga.

TITULO: Nuevo rechazo de embarques duplicados en Manifiestos de Ca

MENSAJE_TITULO: **Nuevo rechazo de embarques duplicados en Manifiestos de Carga.**

Ruta: p » strong

DESCRIPCION:

Ruta:

MENSAJE:

Se informa que, a partir del próximo lune rechazarán aquellos Manifiestos embarquen el mismo documento DUA de S sin seguir alguna de las sigui embarque:

- Cuando se declara un embarque Total (T) no pudo haber embarques parciales (P) ni en resto (R) previos o posteriores. El embarque total (T) debe ser único.

- Si se embarca parcialmente la cadena lógica de envíos serían uno o más embarques pa finalmente un único embarque por el resto (R) que debe ser el último de los declarados.

El error que dará el sistema será el siguiente:

0438 - EL ESTADO DEL DOCUMENTO ASOCIADO IMPIDE LA OPERACION

Ruta: p

RESPONSABLE: Aduanas

TEXTO_CORTO_NOVEDAD: Aduanas. Nuevo rechazo de embarques duplicados en Manifie

Mozilla Firefox

192.168.134.106:27110/AEAT_EditorContenidos/js/tiny_mce/plugins/translate/note.htm

Nota para la empresa de traductores: Verificar que se traduce siempre igual

Tipo de nota: Aviso

Archivo Editar Ver Historial Marcadores Herramientas Ayuda
General | Presentation | Classifications | Channels | Versions | Security | Preview | Status | Scheduling
HTML
No Traducir
Acotar
Corregir
Enviar consulta
OK
Cancel
Apply
Approve
Publish
Help
javascript:;

Figure 2. Manual annotation of Localization Note

The code will then look like this[6]:

```
<p>- Cuando se declara un embarque
Total (T) no pudo haber embarques
parciales <span its-loc-
note="Check that translation is
always the same" its-loc-note-
type="alert">(P)</span> ni em-
barques por el resto (R) previos o
posteriores. El embarque total (T)
debe ser único.</p>
```

- Stage 2. Processing/conversions

During processing, the system detects the note and converts it to what is called *Special Plain Text* (SPT)[7], in order to be processed by the MT system[8]:

```
<p>- Cuando se declara un embarque
Total (T) no pudo haber embarques
parciales <span>[@@its:locNote=
Check that translation is always
the same&&locNoteType=alert@@]
(P)</span> ni embarques por el
resto (R) previos o posteriores.
El embarque total (T) debe ser
único.</p>
```

The MT system recognises the SPT pattern and blocks its translation, subsequently, after the pair of plain text files with the list of original text segments and the translated ones are generated, a sub-process parses those input and output files to be processed by the CAT Tool:

[6] All localization notes were originally written in Spanish as this is the source language used in the showcase. Translations into English are provided here for the sake of understanding.

[7] Metadata is transformed from the original markup to a different format called "Special Plain Text" (SPT) and the other way around. The main reason is because the Translation Memories of the MT Systems, as used in the showcase, do not deal with markup, just with plain text.

[8] The difference between `its-loc-note` and `itsLocNote` is only way of naming. The former is used in Web content. Here, case sensitive attribute names are not allowed. The latter is used in XML content which does not have this restriction.

```
- Cuando se declara un embarque
Total (T) no pudo haber embarques
parciales <las_its2 locNote=‶
Check that translation is always
the same″  locNoteType=‶alert″ />
(P) ni embarques por el resto (R)
previos o posteriores. El embarque
total (T) debe ser único.</p>
```

When the revision process ends, the new revised file is generated and parsed to convert again the new mark-up into the original SPT so as to be loaded in the "Translation Memory", where post-editing will then be performed.

- Stage 3. HTML5 Output

After the processing the system will leave the note as is.

3.2 ITS 2.0 PE contextual information

For the purposes of the showcase, only two of the six ITS 2.0 data categories identified as relevant for PE were put to test: Localization Notes and Localization Quality Issues. The PE process was conducted by a team of three post-editors in the following language combinations: Spanish-English; Spanish-French; and Spanish-German. They were experienced translators who were trained in the ability to locate MT errors, in the instrumental competences to understand MT functioning, and the comprehension of ITS 2.0 annotations. They received PE guidelines compiled during the EDI-TA project (Rico and Díez Orzas, 2013a). The whole project consisted in conducting a localisation assignment in real time with ITS 2.0 for the Spanish Tax Agency.

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 adhering to the localization notes. The Localization Notes are an important tool for this purpose, as they are a direct link between the author of the source text and the post-editor. These can be used to give post-editors contextual information in order to make decision making faster and better. A Localization Note can 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. The text that follows is an example of such an annotation:

```
<p its-loc-note=" These paragraphs
are part of the security warning
of the Tax Agency′s website, field
of informatics apart from economic
and judicial " 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 di-
rectivos.</p>
```

Other uses of Localization Note for PE purposes help post-editors to decide on the use of the appropriate terminology in the source language. Still there are some issues which are impossible to prevent and must be part of the post-editing work (morphological, phrase structure, linear order, syntax…). Tables 1 and 2 are examples of how the use of wrong terminology is prevented.

PE without ITS 2.0	PE with ITS 2.0
ES	
Parámetros de liquidación por tipo de entidad.	*Parámetros de <span its-loc-note=" **In this statistic it's referring to the calculation of the tax debt, it's all the operations that quantify the tax amount**" its-loc-note-type="alert">liquidación</span> por tipo de entidad.*
EN	
Settlement *parameters by type of organisation.*	**Assessment** *parameters by type of organisation.*

Table 1. Using the term "assessment" instead of "settlement" in English

PE without ITS 2.0	**PE with ITS 2.0**
ES	
Programa PADRE	*Programa <span its-loc-note=”* ***This is the acronym for ‘Programa de Ayuda a la Declaración de la Renta’, do not translate*** *” its-loc-note-type=”alert”>PADRE</span>*
DE	
VATER *Programm*	**PADRE** *Programm*

Table 2. Using the term “PADRE” instead of “VATER” in German

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, Tables 3 and 4 show stylistic solutions.

PE without ITS 2.0	**PE with ITS 2.0**
ES	
Sede electrónica, todos los trámites online.	*<span its-loc-note=”* ***Translate as if it said ‘oficina de information’ and be consistent*** *” its-loc-note-type=”alert”>Sede electrónica</span>, todos los trámites online.*
FR	
Siège électronique, toutes les demarches en ligne.	*Bureau électronique, toutes les demarches en ligne.*

Table 3.
A note that forces the correct translation of “sede electronica” (ES) into “bureau électronique” (FR)

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 <span its-loc-note=”* ***Translate as ‘declarante registrado’ because it is a taxpayer who has already presented a tax return in previous financial years and is already ‘registered’*** *” its-loc-note-type=”alert”>declarante</span> 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.*

Table 4.
A note that forces the correct translation of “declarante” (ES) into “registered taxpayer” (EN)

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 (Table 5). 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

ES		EN
1	*Congregaciones*	*<las_its: locQualityIssueType=”characters” locQualityIssueComment=”* ***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*** *” locQualityIssueSeverity=”80” />Congregations*
2	*Religiosas*	*Religious*

Table 5. An example of Localization Quality Issue

4. Conclusion

This paper has presented work towards the implementation of ITS 2.0 for PE. It concentrates in two specific categories – Localization Note and Localization Quality Issue – as a way of illustrating their benefit for PE purposes. This is so as far as they provide the post-editor with the necessary information for conducting a successful review: avoiding ambiguity, improving consistency and accuracy, using the correct terminology.

Although these notes are mostly helpful throughout the post-editing process, some drawbacks were still found during the showcase implementation. The use of notes make sentences slightly less understandable and more cryptic as they have tags with the explanations inserted directly between their words. In cases where 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. 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 would hover the mouse over the item to reveal the information. Other possible solution would be having the annotations from a specific segment displayed in an extra window on the post-editing tool or using abbreviated icons.

Additionally, 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.

Acknowledgements

Pablo Nieto Caride, Felix Fernández, Consuelo Aldana, Mauricio del Olmo, Laura Guerrero, Giuseppe Deriard Nolasco, Pablo Badía (Linguaserve), Ankit Srivastava, Declan Groves (Dublin City University), Thomas Ruedesheim (Lucy Software), Román Díez, Alberto Crespo (Spanish Tax Office).

References

ITS 2.0. 2013. *Internationalization Tag Set (ITS) Version 2.0. W3C Last Call Working Draft 21 May 2013*. Accessed June 22, 2013. Available: http://www.w3.org/TR/2013/WD-its20-20130521/

Nieto Caride, Pablo, Giuseppe Deriard Nolasco, Pedro L. Díez Orzas, Felix Fernández, Consuelo Aldana, Pablo Badía, Ankit Srivastava, Declan Groves, Thomas Ruedesheim, Román Díez, and Alberto Crespo. 2013. “Online MT System Linguaserve Showcase”. *Multilingualweb-LT Deliverable 4.2.1.*, public report. Accessed June 22, 2013. Available: http://www.w3.org/International/multilingualweb/lt/wiki/Deliverables

O’Brien, Sharon. 2012 “Towards a Dynamic Quality Evaluation Model for Translation” *The Journal of Specialised Translation*, Issue 17, Jan. 2012

Rico, Celia. 2012. “A Flexible Decision Tool for Implementing Post-editing Guidelines”, *Localisation Focus*, vol. 11, 1: 54-66. Accessed December 14, 2012. Available: http://www.localisation.ie/resources/locfocus/LocalisationFocusVol11_1Web.pdf

Rico, Celia and Pedro Luis Díez Orzas. 2013a. “EDI-TA: Training methodology for Machine Translation Post-editing.” *Multilingualweb-LT Deliverable 4.1.4. Annex II*, public report. Accessed June 22, 2013. Available: http://www.w3.org/International/multilingualweb/lt/wiki/images/d/d4/D4.1.4.Annex_II_EDI-TA_Training.pdf.

Rico, Celia and Pedro Luis Díez Orzas. 2013b. “EDI-TA: Post-editing methodology for Machine Translation”, *Multilingualweb-LT Deliverable 4.1.4. Annex I*, public report. Accessed June 22, Available: http://www.w3.org/International/multilingualweb/lt/wiki/images/1/1f/D4.1.4.Annex_I_EDI-TA_Methology.pdf.