D4.1.4 - ANNEX I

EDI-TA: POST-EDITING METHODOLOGY FOR MACHINE TRANSLATION

Celia Rico Pérez, Pedro L. Díez Orzas, et al.

Distribution: Public

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

FP7-ICT-2011-7

Project no: 287815
### Document Information

<table>
<thead>
<tr>
<th><strong>Deliverable number:</strong></th>
<th>4.1.4 Annex I</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Deliverable title:</strong></td>
<td>EDI-TA: Post-Editing Methodology for Machine Translation</td>
</tr>
<tr>
<td><strong>Dissemination level:</strong></td>
<td>PU</td>
</tr>
<tr>
<td><strong>Contractual date of delivery:</strong></td>
<td>31&lt;sup&gt;st&lt;/sup&gt; December 2012</td>
</tr>
<tr>
<td><strong>Actual date of delivery:</strong></td>
<td>15&lt;sup&gt;th&lt;/sup&gt; February 2013</td>
</tr>
<tr>
<td><strong>Author(s):</strong></td>
<td>Celia Rico Pérez, Pedro L. Díez Orzas, Giuseppe Dériard Nolasco, Lidia Cámara, Igone Regidor, Martin Ariano, Felix Fernández, Johanna Blasco, Pablo Nieto Caride, Roberto Benítez Monje</td>
</tr>
<tr>
<td><strong>Participants:</strong></td>
<td>LINGUASERVE (assisted by Universidad Europea de Madrid)</td>
</tr>
<tr>
<td><strong>Internal Reviewer:</strong></td>
<td>LINGUASERVE</td>
</tr>
<tr>
<td><strong>Workpackage:</strong></td>
<td>WP4</td>
</tr>
<tr>
<td><strong>Task Responsible:</strong></td>
<td>Celia Rico Pérez</td>
</tr>
<tr>
<td><strong>Workpackage Leader:</strong></td>
<td>Pedro L. Díez Orzas</td>
</tr>
</tbody>
</table>

### Revision History

<table>
<thead>
<tr>
<th>Revision</th>
<th>Date</th>
<th>Author</th>
<th>Organization</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18/10/2012</td>
<td>Celia Rico Pérez</td>
<td>Universidad Europea de Madrid (LINGUASERVE)</td>
<td>First final draft</td>
</tr>
<tr>
<td>2</td>
<td>23/10/2012</td>
<td>Pedro L. Díez Orzas</td>
<td>LINGUASERVE</td>
<td>Revision final draft</td>
</tr>
<tr>
<td>3</td>
<td>23/11/2012</td>
<td>Giuseppe Dériard, Lidia Cámara, Igone Regidor, Martin Ariano, Felix Fernández, Johanna Blasco, Pablo Nieto Caride</td>
<td>LINGUASERVE</td>
<td>Revision final draft</td>
</tr>
<tr>
<td>4</td>
<td>20/12/2012</td>
<td>Pedro L. Díez Orzas</td>
<td>LINGUASERVE</td>
<td>Final Version</td>
</tr>
<tr>
<td>5</td>
<td>13/02/2013</td>
<td>Pedro L. Díez Orzas</td>
<td>LINGUASERVE</td>
<td>Revision Final Version</td>
</tr>
</tbody>
</table>
CONTENTS

Document Information ............................................................................................................. 2
Revision History ....................................................................................................................... 2
Contents .................................................................................................................................. 3
1. Executive Summary ............................................................................................................. 5
2. Introduction ......................................................................................................................... 7
2.1. EDI-TA: project description ......................................................................................... 8
   Phase 1. Post-editing pilot project start-up ....................................................................... 8
   Phase 2. MT post-editing experimentation ....................................................................... 11
   Phase 3. Training and methodology ............................................................................... 15
3. EDI-TA: post-editing methodology .................................................................................... 16
   3.1. Preliminary analysis ................................................................................................. 16
       Objectives ..................................................................................................................... 16
       Role & competences ................................................................................................. 16
       Materials and tools needed ...................................................................................... 16
       Workflow ..................................................................................................................... 17
       Post-editing guidelines specification ...................................................................... 17
       Outcomes ................................................................................................................... 23
3.2. Post-editing MT output ................................................................................................. 24
       Objectives ..................................................................................................................... 24
       Competences ............................................................................................................... 24
       Materials and tools needed ...................................................................................... 25
       Tasks & Workflow ....................................................................................................... 25
       Outcomes ................................................................................................................... 26
3.3. Error reporting and quality control ............................................................................. 26
       Objectives ..................................................................................................................... 26
       Role ............................................................................................................................... 26

D4.1.4.Annex I (EDITA)
4. Requirements for a post-editing tool

5. References
1. EXECUTIVE SUMMARY

EDI-TA’s key objectives are as follows:

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

The present document reports work carried out towards achieving objectives 1 to 6 and is complementary to a second report defining training methodology.

In order to achieve these objectives, the main tasks that have been carried out are:

- Designing a methodology for training post-editors.
- Analysing the economic impact of implementing post-editing processes.
- Defining the functionalities for a post-editing tool.

The project’s setting and team are:

- Using the company’s resources and translation workflow.
- MT output was produced by a rule-based system (Lucy Software).
- # words: 50,000 words per language pair.
- Text typology: Administrative and Financial.
- A TM as PE environment: Transit.
- A practical orientation, as a business oriented R&D project.
- Team:
  - 4 junior translators - 1 senior translator
  - 1 technical support - 1 technical manager
  - 1 project coordinator - 1 project director
In conclusion, some significant findings obtained are:

- Deciding on PE guidelines requires a flexible decision tool.
- Terminology management plays a key role.
- MT Output analysis prior to PE.

Some significant findings include:

- A direct relationship between quality and productivity.
- Use of metatags contributes to defining objective PE guidelines.

These findings indicate that for effective PE, a flexible decision tool is crucial. Terminology management is essential in ensuring accurate and consistent translation. Prior to PE analysis helps in evaluating output quality, crucial for accurate evaluation.

Additionally, there is a direct relationship between quality and productivity, which means improving quality leads to increased productivity. Metatags contribute to defining objective PE guidelines, which in turn improves the effectiveness of PE.
2. Introduction

This report is based on work carried out in the framework of EDI-TA, an R&D project conducted by Linguaserve and Universidad Europea de Madrid, as part of the tasks that Linguaserve is developing within The MultilingualWeb-LT (Language Technologies) Working Group, which belongs to the W3C Internationalization Activity and the MultilingualWeb community. The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) Grant Agreement No. 287815.

TAUS/CNGL defines post-Editing as "the correction of machine-generated translation output to ensure it meets a level of quality negotiated in advance between client and post-editor". One of the crucial aspects in a Post-editing (PE) project is to decide on guidelines to be followed by post-editors. Selecting what elements to change and delivering a final text at a sufficient quality level is usually a difficult matter, due to the subjectivity involved in the task and the specification of a desired quality level. Acceptance and use of half- or semi-finished texts determine to what extent MT output should be post-edited, and how much human effort is necessary to improve such imperfect texts.

So specifying PE guidelines involves deciding on text quality acceptance which, in turn, depends on aspects such as client expectations, turn-around time or document life-cycle, among others. Conventional approaches to PE take as a starting point the distinction between full and light PE (Allen 2003, TAUS/CNGL 2011) but this division gets blurred when one tries to implement it in an actual PE project because a human post-editor usually engages in full post-editing (O’Brien, 2011a) and deems this dissociation as irrelevant.

The paradigm shift experienced by Machine Translation in recent years fosters a different approach to PE, using a methodology for guidelines specification which can be formally shared and smoothly incorporated in the translation workflow.

But there are still some questions to be answered:

- How do you go about implementing Post-editing (PE) processes in your company as an LSP?
- How does PE differ from reviewing TM fuzzy matches?
- What is the post-editor’s role and how can it fit in the company’s workflow?
- How is quality to be assessed?
- And what about productivity?
- Is it true that PE contributes to reducing costs?

Lucy LT from Lucy Software is the MT system used in EDI-TA. Also public content from the web sites of the Spanish Tax Agency, City Hall of Barcelona, Vodafone and El Corte Inglés have been used as corpus for the project.
2.1. EDI-TA: project description

EDI-TA had a duration of 6 months, from March 2012 to October 2012, and work was organized into three phases:

1. **Phase 1.** Post-editing pilot project start-up. First stage of the pilot test: test design, implementation, data collection and partial evaluation of results.
2. **Phase 2.** MT post-editing experimentation. Implementation, in two subsequent stages, of a test bed that serves as a showcase.
3. **Phase 3.** Training and methodology

Figure 1 summarizes work on the basis of each of the objectives:

**Figure 1: EDI-TA’s workplan**

**PHASE 1. POST-EDITING PILOT PROJECT START-UP**

This phase focused on setting up a pilot test that would serve as a reference in subsequent phases of the project. More specifically, the test was designed, data was collected and an evaluation was conducted with partial results. In this sense, core tasks included the following:

- **Web content selection.** A first set of web content was selected for this pilot test. Language pairs included EN-ES, EN-EU, EN-FR, ES-EN, and domains referred to online customer information in mobile technology, and information for citizens in the Spanish Internal Revenue Service.
- **Text analysis for post-edit.** This involved the identification of PE problems as well as the registration of MT out errors. The possibility of using Lucy Desktop and WordSmith as analysis tools was also evaluated.
Metadata definition in terms of post-editing exploitation. A first approach to metadata was made in this first part of the project. The extended metadata identification was undertaken in phase 2 (as described below).

RESULTS IN PHASE 1

- A dynamic PE tool was designed, containing the different text characteristics to be taken into account when establishing PE guidelines. The figure below shows in detail these characteristics, some of which would later be used for defining metadata. The full description of how this tool is to be implemented is contained in Section 2 of the present report (see EDI-TA. Post-editing methodology).
### A dynamic model for PE guidelines

*adapted from O’Eifrea 2012, Guarnán 2007 and Torrijón 2002*

<table>
<thead>
<tr>
<th>Communication channel</th>
<th>Content profile</th>
<th>Subject area (indicate)</th>
<th>UTS Ratings (*low, **medium, **<em>high</em>)</th>
<th>MT Output Quality</th>
<th>Post-editing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal</td>
<td>User interface text</td>
<td>Utility (relative importance of the functionality of the translated content). Rate *, **, ***</td>
<td>High</td>
<td>Grammar should be accurate. <strong>Activate rule?</strong></td>
<td></td>
</tr>
<tr>
<td>External: B2C</td>
<td>Marketing material</td>
<td>Delivery Time (speed with which the translation is required). Rate *, **, ***</td>
<td>Medium</td>
<td>Fix syntactic errors as long as they interfere with the message transferred. <strong>Activate rule?</strong></td>
<td></td>
</tr>
<tr>
<td>External: B2B</td>
<td>User documentation</td>
<td>Sentiment (importance on brand image). Rate *, **, ***</td>
<td>Low</td>
<td>Ensure that key terminology is correctly translated. <strong>Activate rule?</strong></td>
<td></td>
</tr>
<tr>
<td>External: C2C</td>
<td>Website content</td>
<td>Experience. Rate *, **, ***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Online help</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Audio/video content</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Social media content</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Training material</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Figure 2: A dynamic model for PE guidelines
Role of terminology management in a PE project. Terminology management plays a key role in smoothing the PE process. In this respect, the identification of metadata related to term management -such as the indication of text domain, the availability of client glossaries, annotating untranslatable entities, among others- is key to a successful completion of a PE project.

MT Output analysis prior to PE. This task is relevant in the sense that error detection contributes to the evaluation of output quality and, thus, PE effort estimation. It is also important for registering errors that later contribute to improving MT performance.

Phase 2. MT Post-editing Experimentation

This phase focused on conducting a PE experiment on the basis of the above mentioned findings. The following tasks were undertaken:

Creating an experimental PE project. This involved selecting a new set of web content. This time the domain referred to information from the Spanish public administration. Language combinations were as follows: EN-ES, EN-EU, EN-FR, ES-EN. This set amounted to a total of 50,000 words per language pair.

Error analysis. MT output was evaluated so as to detect possible errors (lexical, syntactic and terminological). These were identified and handed over to Lucy Software for codification. All errors were annotated according to a pre-defined record card.

Definition of post-editing rules. PE guidelines were specified with the help of the dynamic model as mentioned above. These included explicit reference on what to expect from the MT output in terms of quality and how to proceed in each case. A PE guide was also prepared containing all details.

Analysis of metadata contribution to improving post-editing processes. An exhaustive analysis of metadata was carried out, following directions from The MultilingualWeb-LT (Language Technologies) Working Group. Each of them was evaluated towards defining its possible effect in a PE project, whether it would contribute to obtain better quality in the PE output.

Results in Phase 2

PE guide. A guide containing practical information on how to approach a PE project has been designed and successfully tested.

Productivity tests. First tests on productivity account for an improvement in time and quality, reaching an average PE productivity of 893 words per hour.

Metadata identification. The list of metadata identified as relevant for PE purposes is as follows, as used in the Online MT System ITS 2.0 demonstration:

---

1 Identification of metadata relevant for PE purposes is based on [http://www.w3.org/TR/2012/WD-its20-20120829/#datacategory-description](http://www.w3.org/TR/2012/WD-its20-20120829/#datacategory-description)
### DATA CATEGORY: TRANSLATE

<table>
<thead>
<tr>
<th>Definition</th>
<th>Use for PE purposes</th>
</tr>
</thead>
</table>
| 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). | - Informing the RTTS of precisely which sentences or sentence fragments should or should not be translated.  
  - Input applies:  
    - Proxy system for a dynamic treatment of page view  
    - MT text treatment.  
    - PE text treatment -> Viewing not translatable content to adjust the possible implications of not automatically translating an element. |

### DATA CATEGORY: LOCALIZATION NOTE

<table>
<thead>
<tr>
<th>Definition</th>
<th>Use for PE purposes</th>
</tr>
</thead>
</table>
| The Localization Note data category is used to communicate notes to localizers about a particular item of content. | - Providing post-editors with the necessary information to review the text in order to help them disambiguation and to improve the quality and accuracy of the revision.  
  - Automatic processing:  
    - locNoteType and certain text values agreed with the client: description for posteditor/Alert send a notification to ProjectManager.  
    - A locNotePointer/locNoteRef/locNoteRefPointer to make available to post-editor and Project Manager the notes in the post-editing package.  
  - Manual processing:  
    - Using localization note to perform the post-edition task.  
  - Some specific notes for post-editing within the locNote element are UTS Ratings (utility, time and sentiment):  
    - 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.  
  - Editing in CAT Tool:  
    - Input applies:  
      - Post-editing workflow (e.g. sending a notification to PM with tooltip visualization).  
      - Import with specific filtering and tagging rules in the PE tool to block non-translatable content. |
## DATA CATEGORY: LANGUAGE INFORMATION

<table>
<thead>
<tr>
<th>Definition</th>
<th>Use for PE purposes</th>
</tr>
</thead>
</table>
| The element **Language Information** is used to express the language of a given piece of content. | • Informing the RTTS about the source language.  
• Allows the user to block automatically the machine translation of certain parts of the Web page that are not required to be translated or that must not be machine translated because of its difficulty or provenance, i.e. a technical essay or constitutional laws.  
• Avoids automatically the machine translation of parts of the Web page that are in various languages and must remain that way, i.e. a language selector.  
• Additionally, part of a content in another language from the rest which could require MT and post-editing for this specific language pair.  
• Automatic processing and manual use.  
  - Automatic processing:  
    o Automatic task assignment.  
    o Semiautomatic assignment of post-editors.  
  - Manual use:  
    o Include/exclude post-editing.  
    o Contextual information for post-editors.  
  - Input applies:  
    o Import with specific filtering and tagging rules in the CAT tool to block non-translate content. |

## DATA CATEGORY: DOMAIN

<table>
<thead>
<tr>
<th>Definition</th>
<th>Use for PE purposes</th>
</tr>
</thead>
</table>
| The **Domain** data category is used to identify the domain of content. | • Automatic processing and manual use.  
- Automatic processing:  
  o Automatic selection of MT terminology.  
  o Semiautomatic assignment of post-editors.  
- Manual use:  
  o Context for post-editing, for example:  
    ▪ Post-editor selection.  
    ▪ Content disambiguation.  
    ▪ New terminology and neologisms.  
- Input applies:  
  o Post-editing workflow.  
  o Import with specific filtering and tagging rules in the CAT tool to block non-translatable content. |
### DATA CATEGORY: PROVENANCE

<table>
<thead>
<tr>
<th>Definition</th>
<th>Use for PE purposes</th>
</tr>
</thead>
</table>
| **Translation Agent/Revision Agent.** | • Translation Agent: Automatic processing.  
  - Automatic processing:  
    - Metadata coding with the information of translator (MT system).  
  - Input applies: none.  
  • Revision Agent: Automatic processing.  
  - Automatic processing:  
    - Metadata coding with the information of reviewer (ID of human reviewer).  
  - Input applies: none.  
  - Output applies: identification of post-editor. |

### DATA CATEGORY: LOCALIZATIONQUALITYISSUE

<table>
<thead>
<tr>
<th>Definition</th>
<th>Use for PE purposes</th>
</tr>
</thead>
</table>
| **The Localization Quality Issue data category is used to express information related to localization quality assessment tasks.** | Automatic processing:  
  - Allow the post-editing system to detect possible localization quality issues.  
  - Categories relevant for post-editing purposes:  
    - Severity. A decimal value representing the severity of the issue, as defined by the model generating the metadata.  
    - Profile reference. A reference to a document describing the quality assessment model used for the issue.  
    - 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). |
PHASE 3. TRAINING AND METHODOLOGY

Work concentrated on the following tasks:

- Training method for language independent post-editing aspects.
- Training method for language dependent post-editing aspects.
- Evaluation of training method and use of metadata in post-editing training contexts.

EDI-TA’s training methodology is reported comprehensively in a separate (and complementary) report.
3. EDI-TA: POST-EDITING METHODOLOGY

According to EDI-TA findings, the methodology for post-editing (PE) is designed into three steps:

1) preliminary analysis;
2) post-editing MT output,
3) error reporting and quality control.

3.1. Preliminary analysis

OBJECTIVES

The main objective of this preliminary step is to analyze MT output quality with a view to:

• Establishing PE patterns for each language combination in the project.
• Report on recurrent MT errors that can be fixed prior to starting the PE process.
• Report new terms to be included in project glossaries.

ROLE & COMPETENCES

This role would typically be assigned to a PE team coordinator who supervises the correct implementation of PE rules. Competences for this role can be grouped as follows:

• Core competences. These are in line with the attitudinal or psycho-physiological competence that allows the post-editor to cope with subjectivity issues involved in defining, and adequately handling client’s expectations in terms of text quality acceptance, and overcoming uncertainty in the PE task. To these, we add the strategic competence that helps post-editors reach informed decisions when choosing between different PE alternatives, based on the preliminary analysis of the sample MT output.
• Linguistic skills. These can be seen as related to skills usually demanded of a translator, and refer to excellent knowledge of source and target language, familiarity with post-editing directions/guidelines, communicative and textual competence in the project languages, and subject area competence.
• Instrumental competences. These are related to technical skills that allow an understanding of what MT involves, and developing a positive attitude/tolerance towards the machine. Instrumental competences refer, then, to knowledge of MT systems and their capabilities (either rule-based, example-based, statistical engines or hybrid systems), term management skills, MT dictionary maintenance (for rule-based systems), and corpus quality assessment skills (for example-based and statistical engines).

MATERIALS AND TOOLS NEEDED

In order to carry out an efficient preliminary analysis, the following materials and tools are needed:

• Access to a representative sample of MT output in the project languages so that all necessary tests can be adequately conducted.
• Access to the MT engine with complete functionalities.
• Access to a client’s glossary for term consistency control.
• PE guidelines specification, including explicit reference on what to expect from the MT output in terms of quality and how to proceed in each case.

Rico and Torrejón, forthcoming
**WORKFLOW**

This preliminary step involves the following sequence of operations:

![Workflow diagram]

**POST-EDITING GUIDELINES SPECIFICATION**

The specification of PE guidelines involves gathering in a single source all the aspects that influence the post-editor’s decision so that PE guidelines can be easily drawn, adequately supported with actual examples and, what is more important, shared and replicated along different PE projects.

The main elements to be considered are listed in Figure 3 below and refer to the following:

- Data set 01: project information.
- Data set 02: text profile.
- Activation rules: text related guidelines and language specific rules.
- Example card: registers typical PE samples for each language pair.
The **data sets** provide practical information on the PE project as well as a formalization of other aspects which, subsequently, contribute to specifying PE guidelines, and serve the PE team coordinator both to keep track of its most practical aspects and to gather broad knowledge on the task at hand. The list of features to be considered is defined as follows:

**Data set 01: project information**

- Client identification (this refers to the internal project identification code).
- Client description (this is a short description of who the client is together with any particularities the project coordinator might deem necessary).
- Text identification (this would typically be an internal project code).
- Text description (a short description of the particularities of the text not covered in any of the other categories).
- Glossary availability (indicating whether there are any available glossaries –from the client or internal to the company- which the post-editor might need access to).
- Domain (this refers to the specification of content subject area).
- MT engine (this is a reference to the MT system used, with indication of any specific internal rules which have been activated, glossaries used, training data, and interaction with translation memories, if any).
• MT output quality (this refers to a grading of the output text quality).

Data set 02: text profile

• Communication channel. This refers to the description of the communicative purposes of the document, which can be used either for internal purposes or for external communication, as previously described. This latter category is further divided into three subcategories: Business to Customer, Business to Business and Customer to Customer.
• Content profile. The information gathered in this category relates to text type and complements from data set 01 regarding “text description” and “domain”.
• Utility, Time and Sentiment. These subcategories refer to the importance of the functionality of the translated content (Utility), the speed at which the PE output is to be handed (Time), and the importance of impact on brand image. Each of these is rated according to three metrics: low, medium and high.

The activation rules aim at offering clear indications to post-editors on how to carry on so that they can stick to them with no hesitation, and are divided into a) text related rules, and b) language specific rules.

Text related rules

<table>
<thead>
<tr>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fix wrong terminology.</td>
</tr>
<tr>
<td>Spend time in terminology research.</td>
</tr>
<tr>
<td>Fix syntactic errors (wrong part of speech, incorrect phrase structure, wrong linear order).</td>
</tr>
<tr>
<td>Fix morphological errors (number, gender, case, tense, voice).</td>
</tr>
<tr>
<td>Fix misspelling errors.</td>
</tr>
<tr>
<td>Fix punctuation errors.</td>
</tr>
<tr>
<td>Fix any omissions as long as they interfere with the message transferred.</td>
</tr>
<tr>
<td>Edit any offensive, inappropriate or culturally unacceptable information.</td>
</tr>
<tr>
<td>Fix any problem related to textual standards (cohesion, coherence).</td>
</tr>
</tbody>
</table>

The way to proceed is to review each of the rules and decide whether to activate them or not, depending on the information previously gathered in the two data sets above.

Language specific rules

Together with the general PE guidelines activated in the data set above, there might be some language specific guidelines that need to be taken into consideration, when they are not covered in text related guidelines.

Language specific rules are, for example, the use of a particular language locale, lexical collocations or specific sentence structures, how product names should be dealt with (whether there is an equivalent available or the source language name should be used). In the language combination ES-EN, rules would typically include instructions on how to deal with the translation of sentences using the infinitive tense, how to PE third person singular, or an indication of when to delete unnecessary uses of “the”, among others.

---

3 These categories are inspired on the work of O’Brien (2012).
Finally, the **example card** is key to providing post-editors with a set of representative examples for each of the rules so they know what to look for, how to deal with the different rules and what PE implies. Each PE project should compile its own example card taking into account the particularities of the text, language combination as well as all other parameters as we have seen above.

Examples below show PE rules for the language combinations in EDI-TA: ES-EN, EN-ES, ES-FR, ES-EU.

**Examples of rules for post-editing Spanish into English**

<table>
<thead>
<tr>
<th><strong>PE rule</strong></th>
<th><strong>MT input</strong></th>
<th><strong>MT output</strong></th>
<th><strong>PE output</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Fix wrong terminology.</td>
<td>Sin derecho a deducción</td>
<td>Without law to deduction</td>
<td>Without deduction right</td>
</tr>
<tr>
<td>Fix morphological errors (number, gender, case, tense, voice).</td>
<td>El espacio más alto será la sala de ensayo de los castellers, que se sitúa en la planta del sótano,</td>
<td>The highest space will be the rehearsing room of the castellers, *&lt;A&gt;[which</td>
<td>who]&gt;* are situated in the plant of the basement,</td>
</tr>
<tr>
<td>Fix syntactic errors (wrong part of speech, incorrect phrase structure, wrong linear order).</td>
<td>Para aprender más cosas sobre la ciencia</td>
<td>To learn more things on the science</td>
<td>Plans of forecast guaranteed</td>
</tr>
<tr>
<td>Fix any omissions as long as they interfere with the message transferred.</td>
<td>Yo creo que un poco más crudo y más abierto, pero me parece que bastante fiel al espíritu del disco, a los cruces instrumentales y a las atmósferas vocales.</td>
<td>I believe that a little more raw and more open, but seems me that quite faithful to the spirit of the &lt;A&gt;[disk</td>
<td>record]&gt;, to the instrumental &lt;A&gt;[crosses</td>
</tr>
<tr>
<td>Fix stylistic problems only when needed.</td>
<td>...para ofrecer un escaparat de la <em>industria española del videojuego</em></td>
<td>to offer a shop window of the Spanish <em>industry of the video game</em></td>
<td>to offer a showcase of the Spanish <em>industry of the video game</em></td>
</tr>
</tbody>
</table>

---

4 A full set of detailed examples can be found in the document *EDI-TA Training Methodology*, which is complementary to this report.
<table>
<thead>
<tr>
<th>PE rule</th>
<th>MT input</th>
<th>MT output</th>
<th>PE output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fix wrong terminology.</td>
<td>The show is a fantasy-filled parade created by the imagination of a clown, the leading character</td>
<td>El espectáculo es un desfile lleno de fantasía creado por la imaginación de un payaso, el carácter principal.</td>
<td>El espectáculo es un desfile lleno de fantasía creado por la imaginación de un payaso, el personaje principal.</td>
</tr>
<tr>
<td>Fix grammatical mistakes.</td>
<td>And to enjoy the story from even closer up, you can see it in 3D, a technology that is even included in the film’s title.</td>
<td>Y disfrutar de la historia de incluso más cercano hacia arriba de, lo puedes ver en 3D, una tecnología que se incluye incluso en el título de la película.</td>
<td>Y para disfrutar de la historia incluso desde más cerca, la puedes ver en 3D, una tecnología que se incluye incluso en el título de la película.</td>
</tr>
<tr>
<td>Fix morphological errors (number, gender, case, tense, voice).</td>
<td>This musical comedy now on at the Teatre Poliorama, is directed, adapted and produced by the Tricicle company, who found the work fascinating when they saw it in Oslo and decided to export it.</td>
<td>Esta comedia musical ahora en el Teatre Poliorama, es dirigido, adaptado y producido por la compañía de Tricicle, que encontró el trabajo fascinante cuando lo vieron en Oslo y decidieron exportarlo.</td>
<td>Esta comedia musical ahora en el Teatre Poliorama, es dirigida, adaptada y producida por la compañía de Tricicle, que encontró el trabajo fascinante cuando lo vieron en Oslo y decidieron exportarlo.</td>
</tr>
<tr>
<td>Fix syntactic errors (wrong part of speech, incorrect phrase structure, wrong linear order).</td>
<td>Alba Sarraute returns to the SAT! with this clown and cabaret show in which she travels to Amazonia to reveal a world we all have deep inside us.</td>
<td>Alba Sarraute vuelve a SAT! con este el payaso y cabaret enseñan en cual viaja a Amazonia para revelar un mundo que todos nosotros tenemos profundo dentro de nosotros.</td>
<td>Alba Sarraute vuelve a SAT! con este espectáculo de payasos y cabaret en el cual viaja a la Amazonia para revelar un mundo que todos nosotros tenemos muy dentro de nosotros.</td>
</tr>
<tr>
<td>Fix any omissions as long as they interfere with the message transferred.</td>
<td>Cirque du Soleil is coming back to Barcelona next January with their show Corteo.</td>
<td>El Cirque du Soleil está volviendo a Barcelona (x) próximo enero con su espectáculo Corteo.</td>
<td>El Cirque du Soleil está volviendo a Barcelona el próximo enero con su espectáculo Corteo.</td>
</tr>
<tr>
<td>Fix stylistic problems only when needed.</td>
<td>The Municipal government launches a plan to promote employment for young people without jobs.</td>
<td>El Gobierno Municipal lanza un plan para promover empleo para gente joven sin trabajos.</td>
<td>No cambiar “sin trabajos” por “desempleada”.</td>
</tr>
</tbody>
</table>
Examples of rules for post-editing Spanish into French

<table>
<thead>
<tr>
<th>PE rule</th>
<th>MT input</th>
<th>MT output</th>
<th>PE output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fix wrong terminology.</td>
<td>Novedades para ampliar público</td>
<td>Nouveautés pour agrandir du public</td>
<td>Nouveautés pour élargir le public</td>
</tr>
<tr>
<td>Fix grammatical mistakes.</td>
<td>Además, en este comunicado el Ayuntamiento también manifiesta que “la defensa de unas ideas legítimas no puede ser, en ningún caso, la excusa para bloquear el acceso al Parlamento de Cataluña de los diputados y diputadas que han sido democráticamente escogidos”.</td>
<td>De plus, en ce la Mairie communiquée aussi manifeste que “la défense d’idées légitimes ne peut être, en aucun cas, l’excuse pour bloquer l’accès au Parlement de la Catalogne des députés et de députées qui ont été démocratiquement choisis”.</td>
<td>De plus, dans ce communiqué la Mairie manifeste aussi que la « défense d’idées légitimes ne peuvent être, en aucun cas, une excuse pour bloquer l’accès au Parlement de la Catalogne aux députés qui ont été démocratiquement choisis ».</td>
</tr>
<tr>
<td>Fix morphological errors (number, gender, case, tense, voice).</td>
<td>El estado de la arena, del mar, las banderas, las temperaturas y la insolación se pueden consultar en el web de playas</td>
<td>L’état du sable, de la mer, les drapeaux, les températures et l’insolation peuvent être consultés en le web de plages</td>
<td>L’état du sable, de la mer, les drapeaux, les températures et l’insolation peuvent être consultés sur le site web des plages.</td>
</tr>
<tr>
<td>Fix any omissions as long as they interfere with the message transferred.</td>
<td>Los chiringuitos de Ciutat Vella tienen autorización para poner la televisión con el volumen limitado.</td>
<td>Les buvettes de &lt;U[Ciutat Vella]&gt; ont autorisation pour mettre la télévision avec le volume limité.</td>
<td>Les buvettes de Ciutat Vella ont l’autorisation pour mettre la télévision avec le volume limité.</td>
</tr>
<tr>
<td>Fix stylistic problems only when needed.</td>
<td>Sube arriba. / Entra dentro de la casa</td>
<td>Il monte en haut. / Il entre dans la maison</td>
<td>Il monte en haut / Il entre dans la maison (the sentence remains as it is).</td>
</tr>
</tbody>
</table>
### Examples of rules for post-editing Spanish into Euskera

<table>
<thead>
<tr>
<th>PE rule</th>
<th>MT input</th>
<th>MT output</th>
<th>PE output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fix wrong terminology.</td>
<td>tipo de archivo</td>
<td>@A[artxibo</td>
<td>artxibatze]@-</td>
</tr>
<tr>
<td>Fix morphological errors (number, gender, case, tense, voice).</td>
<td>La muestra, además, se acompaña de audiovisuales, grabados por el mismo autor, con el testimonio directo de familiares afectados.</td>
<td>Gainera, laguntzen dio erakusketari ikus-entzuneko <em>etatik</em>, egile berak grabatuta, eragindako senide <em>etako</em> zuzeneko lekukotasunarekin.</td>
<td>Gainera, erakusketari ikus-entzuneko <em>etatik</em>, egile berak grabatutakoak, eragindako senide zuzeneko lekukotasunarekin.</td>
</tr>
<tr>
<td>Fix any omissions as long as they interfere with the message transferred.</td>
<td>Por defecto, cualquier PC con Windows 7 lo está.</td>
<td>Lehenetsi, Windows 7 duen edozen ordenagailu dago.</td>
<td>Lehenetsi, Windows 7 duen edozen ordenagailu <em>prestatuta</em> dago</td>
</tr>
<tr>
<td>Fix stylistic problems only when needed.</td>
<td>Es necesario que introduzcas tus datos correctamente al registrarte en el servicio.</td>
<td>Beharrezkoa da zu erregistratzerakoan zuk zure datuak zuzen zerbitzuan sartzea.</td>
<td>Beharrezkoa da zerbitzuan erregistratzerakoan zure datuak zuzen sartzea.</td>
</tr>
</tbody>
</table>

### OUTCOMES

This preliminary analysis generates the following outcomes:

- **Data analysis for post-editing guidelines specification.** This includes data for defining language independent PE rules, language dependent PE rules, system specific PE rules and client specific rules.
- **PE guide.** A guide containing practical information on how to approach the PE project. These include explicit reference on what to expect from the MT output in terms of quality and how to proceed in each case.
3.2. Post-editing MT output

OBJECTIVES

Three main objectives are identified in this step:

- Post-editing MT output according to the specifications and guidelines established in the previous phase, and registered in the PE guide.
- Registering MT errors to be reported later in the process.
- Responding to output quality as required by the client.

COMPETENCES

The skills and competences to look for in a post-editor share some similarities to those described for the PE coordinator. These are grouped, again, into three main categories: core competences, linguistic skills and instrumental competences, as summarized in Figure 4 below.

![Figure 4: PE skills and competences](Source: Rico and Torrejón, forthcoming)

- **Core competences.** These are in line with the attitudinal or psycho-physiological competence that allows the post-editor to cope with subjectivity issues involved in defining and applying PE specifications, adequately handling client’s expectations in terms of text quality acceptance, and overcoming uncertainty. To these, we add the strategic competence that helps post-editors reach informed decisions when choosing between different PE alternatives, following the directions as mechanically as possible and showing no stylistic concerns, if required, even when faced with low quality output. As Guzmán (Guzmán, 2007) puts it, this means «specifying the scope of manual MT post-editing and sticking to it stoically».
• **Linguistic skills.** These can be seen as related to skills usually demanded of a translator, and refer to excellent knowledge of source and target language, familiarity with post-editing directions/guidelines, communicative and textual competence in at least two languages and cultures, cultural and intercultural competence and subject area competence.

• **Instrumental competences.** These are related to technical skills that help the post-editor understand what is behind the MT output and develop a positive attitude/tolerance towards the machine. Instrumental competences refer, then, to knowledge of MT systems and their capabilities (either rule-based, example-based, statistical engines or hybrid systems), term management skills, MT dictionary maintenance (for rule-based systems), corpus quality assessment skills (for example-based and statistical engines), and, finally some programming skills (for creating macros for automated correction).

---

**MATERIALS AND TOOLS NEEDED**

The *post-editor's kit* should contain the following tools and materials:

• PE guidelines, as described in the previous step and including explicit reference on what to expect from the MT output in terms of quality and how to proceed in each case.

• *Access to a client’s glossary for term consistency control.*

• *Fehler! Verweisquelle konnte nicht gefunden werden.* MT output with relevant meta data for post-editing purposes, as described previously on page *Fehler! Textmarke nicht definiert.*.

---

**TASKS & WORKFLOW**

The following tasks are identified in the PE process:

• **Source text related processes.** These refer essentially to reading the source text, either entirely or in segments, looking forward to recognize some pattern for reformulation in the target text (morphological, syntactic or semantic) or to decide on textual coherence. In the translation industry, the question of whether the post-editor should get access to the source text is still under consideration as in some contexts it is deemed as a barrier to reaching optimal productivity. In our experience, using the source text as a reference is key when dealing with low quality MT output for there are no other means of understanding (and thus correcting) the target text.

• **Machine translation related processes.** This is one of the most important process categories, together with those related to the target text (either in production or in evaluation). The category refers to reading the MT output either entirely or in segments, directing attention to elements which need further confirmation in the source text and evaluating whether a reformulation is necessary.

• **Target text production.** By far the largest proportion of the process falls into this category and it is concerned with producing a new text either from old elements already present in the text or adding new ones. Tasks involved in this process are, then, related to language correction according to PE guidelines (spelling, morphology, agreement, syntax order, lexical choice), dealing with terminology issues, guaranteeing style/terminology consistency/coherence. In this respect, we should remember that defining PE guidelines is a complex task since it involves considering quality acceptance from the client’s point of view, turn-around time and final use of the text, among others.

• **Target text evaluation.** Tasks involved in this process are related to making positive or negative evaluations of the MT output and comparing it with the source text, which, in turn, is related to defining quality in terms of client expectations.

---

5 This sections follows the work of Krings and Koby, 2001
• **Reference work related processes.** This category is concerned with choosing dictionaries to be used (if any), collocations to be found, parallel texts to be consulted and/or informants to be asked for. Similarly, this involves tasks related to dictionary/terminology maintenance in MT and TM systems in order to improve lexical accuracy.

• **Physical writing processes.** These refer to the action of writing linearly, deleting or inserting elements, leaving a gap, marking specific elements, overwriting and rewriting.

• **Global task-related processes.** These are processes that serve to control the procedure of dealing with the PE task, i.e. are related to task management: the post-editor «must not only determine how the task should be best divided, but also in what order the subtasks should then be processed» (Krings and Koby, 2001: 510).

## OUTCOMES

This second phase produces the following resulting elements:

- Post-edited output.
- List of terms to be updated in project’s glossaries.
- List of observations toward the PE guidelines, should any of these be updated.

### 3.3. Error reporting and quality control

## OBJECTIVES

The objectives of this last phase in the PE process are:

- Reporting feedback to allow for MT improvement and/or source content optimization, which could help solve repetitive mistakes of MT output.
- Conducting a quality control following on demand client specifications and expectations.
- Additionally, this phase involves collecting samples of different post-editing issues in order to facilitate training of other fellow post-editors in the team, and keeping up-to-date with the latest advances in the field of MT and pre/post-editing tools.

## ROLE

During this phase, post-editors work in close collaboration with the PE team coordinator. They provide project feedback which would be used for improving MT performance, updating project glossaries and revising PE guidelines (see “Tasks & Workflow” below).

The PE team coordinator is ultimately responsible for the adequate completion and delivery of the PE project.

## TOOLS

The following tools are needed:

- **MT error reporting template**

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input*</td>
<td>Original segment which generates the error</td>
</tr>
<tr>
<td>MT output*</td>
<td>Machine Translated output</td>
</tr>
<tr>
<td>Correct Output*</td>
<td>Post-edited output</td>
</tr>
</tbody>
</table>
### Field Description

- **Context**: Complete context where the error is found
- **Comments**: Any comments or questions that need to be taken into account
- **Priority**: Priority for the post-editor in terms of error frequency, importance or effort needed. (Highest priority: 1, lowest: 3)
- **Date**: Reporting date
- **Post-editor**: Reporting person’s name
- **Lexicon version**: Number of the version of lexicon used in MT
- **Status**: To be completed by the MT system team
- **Type**: To be completed by the MT system team
- **Comments**: To be completed by the MT system team

* = compulsory information

In this kind of reporting, post-editors often try to help MT linguists and developers suggesting the solution of problems, without having a real knowledge of the possible causes. It would be like an ornithologist telling an aeronautic engineer how to improve an airplane, the ideas would probably be valid, but not the analysis and the implementation approach. In that respect the field “Priority” is especially relevant. This field refers to the importance of the problem, specifically for the post-editing task (time consuming, complexity, etc.), so MT linguists and developers can order the large amount of problems to solve them by using this prioritization.

**Quality control template**

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MT Source text</td>
<td>Source language text, MT input</td>
</tr>
<tr>
<td>MT Target text</td>
<td>Output without post-editation</td>
</tr>
<tr>
<td>Post-edited target</td>
<td>Post-edited output</td>
</tr>
<tr>
<td>Category/error type</td>
<td>Classification of error categories and error types to identify the problem.</td>
</tr>
<tr>
<td>Relevance of error</td>
<td>Serious or minor</td>
</tr>
<tr>
<td>Resolution approach</td>
<td>Where and how can be fixed</td>
</tr>
<tr>
<td>MT Segment Quality</td>
<td>Quality of the segment for post-editing</td>
</tr>
<tr>
<td>Reviewer remarks</td>
<td>Comments of the reviewer</td>
</tr>
</tbody>
</table>

An example of error classification is the following:

<table>
<thead>
<tr>
<th>Error Category</th>
<th>Error type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Format</td>
<td>Incorrect capitalization, Malformed, Incorrect format conversion</td>
</tr>
<tr>
<td>Lexicon</td>
<td>Term choice (terminology), Problems with phraseological units / placements, Translation of names or words in other languages, Misspellings, typos and punctuation / use of nonstandard abbreviations</td>
</tr>
<tr>
<td>Morphology</td>
<td>Concordance: subject, gender, number, Wrong Case: Genitive, Wrong case: other cases, Problem with gender, Misuse of article, Lack of preposition in source text</td>
</tr>
<tr>
<td>Syntax</td>
<td>Auxiliary verb / modal verb choice, Wrong tense, Negative clause, Ambiguity, Genitive (English), Incorrect coordination, Word order in the entire sentence / word order in a phrase, Clause connectors, Missing words, Overuse or misuse of prepositions</td>
</tr>
</tbody>
</table>
* PE guidelines comments template.

<table>
<thead>
<tr>
<th>PE rule template. Instructions</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PE rule #</strong> [indicate rule number]: [write the name of the rule]</td>
<td><strong>PE rule 01: fix any wrong term</strong></td>
</tr>
<tr>
<td><strong>MT input:</strong> EN</td>
<td><strong>MT input:</strong> EN</td>
</tr>
<tr>
<td>[Write here an example of the MT input which illustrates the rule]</td>
<td><strong>MT output:</strong> ES</td>
</tr>
<tr>
<td><strong>MT output:</strong> ES</td>
<td><strong>PE output:</strong> ES</td>
</tr>
<tr>
<td>[Include here the MT output which illustrates the rule]</td>
<td><strong>Comments</strong></td>
</tr>
<tr>
<td><strong>PE output:</strong> ES</td>
<td><strong>PE output:</strong> ES</td>
</tr>
<tr>
<td>This rule reads as follows: [write here the complete rule]</td>
<td><strong>Comments</strong></td>
</tr>
<tr>
<td>[Explain the examples used above]</td>
<td></td>
</tr>
</tbody>
</table>

* Term registration template

<table>
<thead>
<tr>
<th>Term number identification: [indicate identification number]</th>
<th>Term number identification:001</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source term</strong></td>
<td><strong>Source term</strong></td>
</tr>
<tr>
<td>[Write here the source term ]</td>
<td><strong>Source term in context</strong></td>
</tr>
<tr>
<td><strong>Source term in context</strong></td>
<td><strong>Source term in context</strong></td>
</tr>
<tr>
<td>[Write here an example of the term in context in the source text]</td>
<td><strong>Machine translated term</strong></td>
</tr>
<tr>
<td><strong>Machine translated term</strong></td>
<td><strong>Machine translated term</strong></td>
</tr>
<tr>
<td><strong>Target term</strong></td>
<td><strong>Target term</strong></td>
</tr>
<tr>
<td>[Write here the correct term in the target language]</td>
<td><strong>Target term in context</strong></td>
</tr>
<tr>
<td><strong>Target term in context</strong></td>
<td><strong>Target term in context</strong></td>
</tr>
<tr>
<td><strong>Comments</strong></td>
<td><strong>Comments</strong></td>
</tr>
</tbody>
</table>

In this template, linguistic information (such as part of speech, morphological paradigm, etc.) to code the term in the MT System is not included. This information is developed by MT expert terminologists.
Tasks are organized along the following workflow:
4. Requirements for a Post-editing Tool

After the analysis of a series of post-editing tools available on the market\(^6\), EDI-TA advanced the following list of requirements if such a tool is to be smoothly implemented in the Machine Translation workflow:

- Delivering an open tool, which can be integrated into different MT systems and solution through and open API.
- Using standard web formats HTML5, XML, XLIFF, RDF as well as ITS 2.0 metadata.
- Full integration with Translation Memory data (searching for alternative translations).
- Full integration with project glossaries for terminology look-up and update.
- Propagating post-edited segments along the project.
- Access to online reference materials.
- Source and target text visualization.
- Blocking and unblocking meta data and non-translatable elements.
- Dividing and joining segments.
- Copy and paste features.
- Inserting comments.
- Filtering options (so that segments can be grouped along different post-editing criteria).
- Configuring number of allowed characters.
- Search and replace options.
- Controlling and tracking changes.
- Measuring post-editing effort and productivity (words per hour).
- Spell checker.
- Terminology checker.
- Tag and format checker.
- Visualization options (horizontal/vertical distribution of windows).
- Report generation.

Integration of the tool in the translation workflow is represented in Figure 5, which describes what EDI-TA calls a Controlled Translation Suite. This is a different approach to MT performance enhancement, by developing an open-source tool that allows for automatic pre-editing the input text before it is processed by the engine, and then automatically post-editing the output before it is finally published. The implementation of such a tool facilitates the creation of a Controlled Translation Scenario where MT output is enhanced by controlling both input and output text.

Integration of the tool in the translation workflow is represented in Figure 5, which describes what EDI-TA calls a Controlled Translation Suite. This is a different approach to MT performance enhancement, by developing an open-source tool that allows for automatic pre-editing the input text before it is processed by the engine, and then automatically post-editing the output before it is finally published. The implementation of such a tool facilitates the creation of a Controlled Translation Scenario where MT output is enhanced by controlling both input and output text.

\(^6\) The tools analyzed were Boltran, PET, Globalsight, OmegaT and CAITRA

---

Figure 5: Post-editing in a Controlled Translation
5. REFERENCES


Rico, C. (forthcoming): “A flexible decision tool for implementing Post-editing guidelines” in Localisation Focus

Rico, C. and E. Torrejón (forthcoming): “Skills and Profile of the New Role of the Translator as MT Post-editor” in Tradumática
