MultilingualWeb-LT

D4.1.4 - ANNEX I
EDI-TA: POST-EDITING METHODOLOGY FOR MACHINE TRANSLATION

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

Distribution: Public

Document Information

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

Revision History

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

CONTENTS

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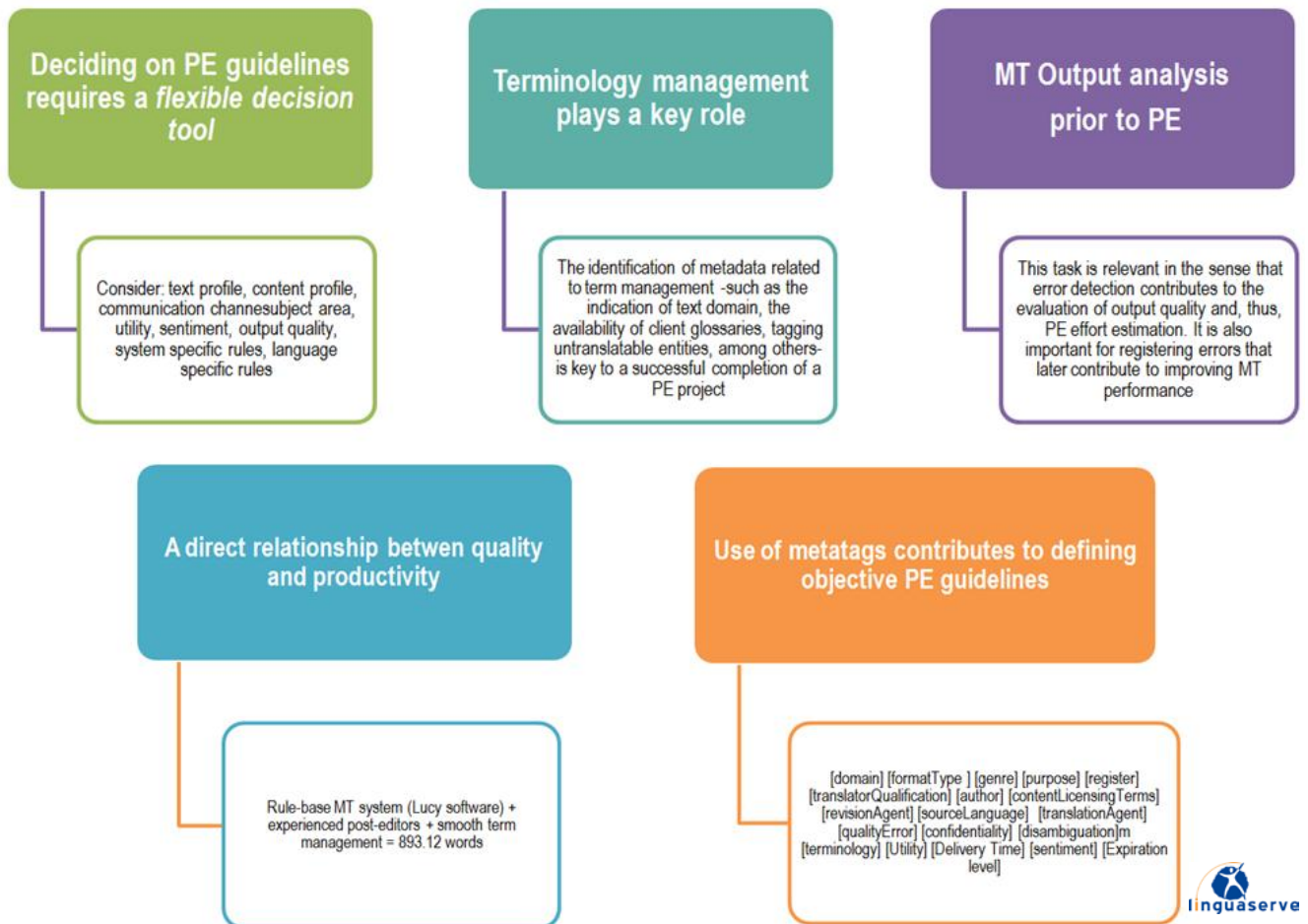
In order to achieve these objectives, the main tasks that have been carried out are:

- 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).
- Language pairs: EN-ES, ES-EN, ES-CAT, ES-EU.
- # words: 50,000 words per language pair.
- Text typology: Administrative and Financial.
- A TM as PE environment: Transit.
- March – July 2012.
- 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*

Consider: text profile, content profile, communication channesubject area, utility, sentiment, output quality, system specific rules, language specific rules

Terminology management plays a key role

The identification of metadata related to term management -such as the indication of text domain, the availability of client glossaries, tagging 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

A direct relationship betwen quality and productivity

Rule-base MT system (Lucy software) + experienced post-editors + smooth term management = 893.12 words

Use of metatags contributes to defining objective PE guidelines

[domain] [formatType] [genre] [purpose] [register] [translatorQualification] [author] [contentLicensingTerms] [revisionAgent] [sourceLanguage] [translationAgent] [qualityError] [confidentiality] [disambiguation]m [terminology] [Utility] [Delivery Time] [sentiment] [Expiration level]

linguaserve

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

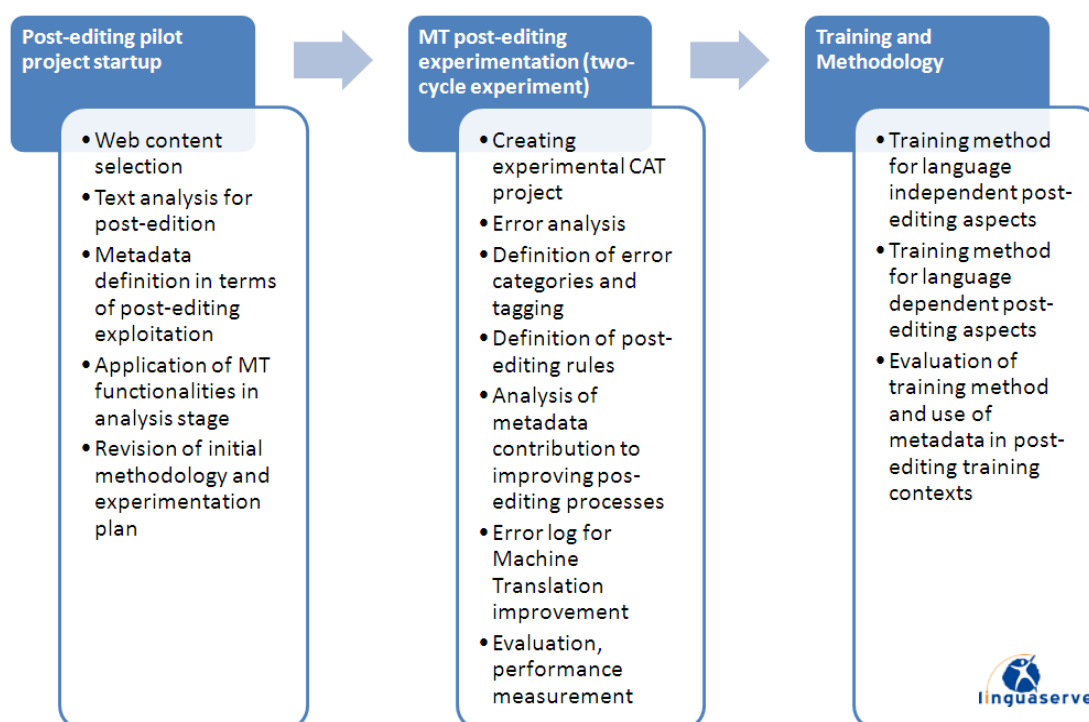Figure 1 summarizes work on the basis of each of the objectives:

Post-editing pilot project startup
- Web content selection
- Text analysis for post-edition
- Metadata definition in terms of post-editing exploitation
- Application of MT functionalities in analysis stage
- Revision of initial methodology and experimentation plan

MT post-editing experimentation (two-cycle experiment)
- Creating experimental CAT project
- Error analysis
- Definition of error categories and tagging
- Definition of post-editing rules
- Analysis of metadata contribution to improving pos-editing processes
- Error log for Machine Translation improvement
- Evaluation, performance measurement

Training and Methodology
- 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

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-edition**. 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 *

Client ID:	Client description:
Text ID:	Text description:
Glossary (Yes/No):	

Communication channel	Content profile	Subject area (*indicate*)	UTS Ratings (*low, **medium, ***high*)	MT Output Quality	Post-editing	
Internal	User interface text		*Utility* (relative importance of the functionality of the translated content). Rate *, **, ***	High	• Grammar should be accurate. Activate rule?	
External: B2C	Marketing material		*Delivery Time* (**speed** with which the translation is required). Rate *, **, ***	Medium	• Fix syntactic errors as long as they interfere with the message transferred. Activate rule?	
External: B2B	User documentation		*Sentiment* (importance on brand image). Rate *, **, ***	Low	• Ensure that key terminology is correctly translated. Activate rule?	
External: C2C	Website content		*Expiration level. Rate *, **, ***		• Spend time researching terms. Activate rule?	
	Online help				• Fix any omissions as long as they interfere with the message transferred. Activate rule?	
	Audio/video content				• Fix morphological errors. Activate rule?	
	Social media content				• Fix misspelling errors. Activate rule?	
	Training material				• Fix punctuation errors. Activate rule?	
					• Edit any offensive, inappropriate or culturally unacceptable information. Activate rule?	
					• All basic rules regarding spelling still apply. Activate rule?	
					• Textual standards (cohesion, coherence, standard word order etc.) are not so important. Activate rule?	
					• Ignore stylistic problems: infinitives vs. nouns in titles, personal vs. Impersonal, active vs. passive, prepositions and articles, word order. Activate	
					• For tagged formats, ensure all tags are present and in the correct positions. Activate rule?	

linguaserve

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 meta data identified as relevant for PE purposes is as follows, as used in the Online MT System ITS 2.0 demonstration[1]:

---

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

DATA CATEGORY: TRANSLATE	
Definition	*Use for PE purposes*
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	
Definition	*Use for PE purposes*

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 and manual use:
	- Automatic processing:
	   o **locNoteType** and certain text values agreed with the client: description for posteditor/Alert send a notification to ProjectManager.
	   o A **locNotePointer**/**locNoteRef**/A **locNoteRefPointer** to make available to post-editor and Project Manager the notes in the post-editing package.
	- Manual processing:
	   o 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):
	   o Utility (relative importance of the functionality of the translated content).
	   o Delivery Time (speed with which the translation is required).
	   o Sentiment (importance on brand image).
	   o Expiration level.
	• Editing in CAT Tool:
	- Input applies:
	   o Post-editing workflow (e.g. sending a notification to PM with tooltip visualization).
	   o Import with specific filtering and tagging rules in the PE tool to block non-translatable content.

DATA CATEGORY: LANGUAGE INFORMATION	
Definition	*Use for PE purposes*

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 assigment 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	
Definition	*Use for PE purposes*

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	
Definition	*Use for PE purposes*

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	
Definition	*Use for PE purposes*
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).*

	- 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. **Manual processing:** - Allow the post-editor to provide quality issue information. - Categories relevant for post-editing purposes: - Comment. A human-readable description of the quality issue. - Profile reference. A reference to a document describing the quality assessment model used for the issue. - 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 (NB: not for use with terminology inconsistency). - 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 of 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. - Internationalization. There is an issue related to the internationalization of content. Uncategorized. The issue has not been categorized.

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[2]:

- *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.

---

[2] Rico and Torrejón, forthcoming

WORKFLOW

This preliminary step involves the following sequence of operations:
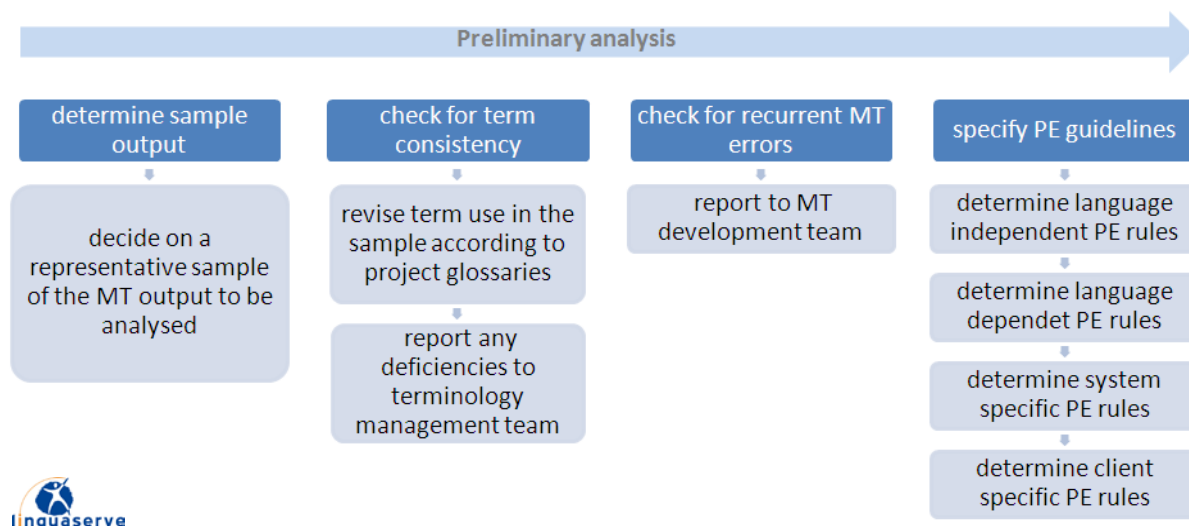
POST-EDITING GUIDELINES SPECIFICATION

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.
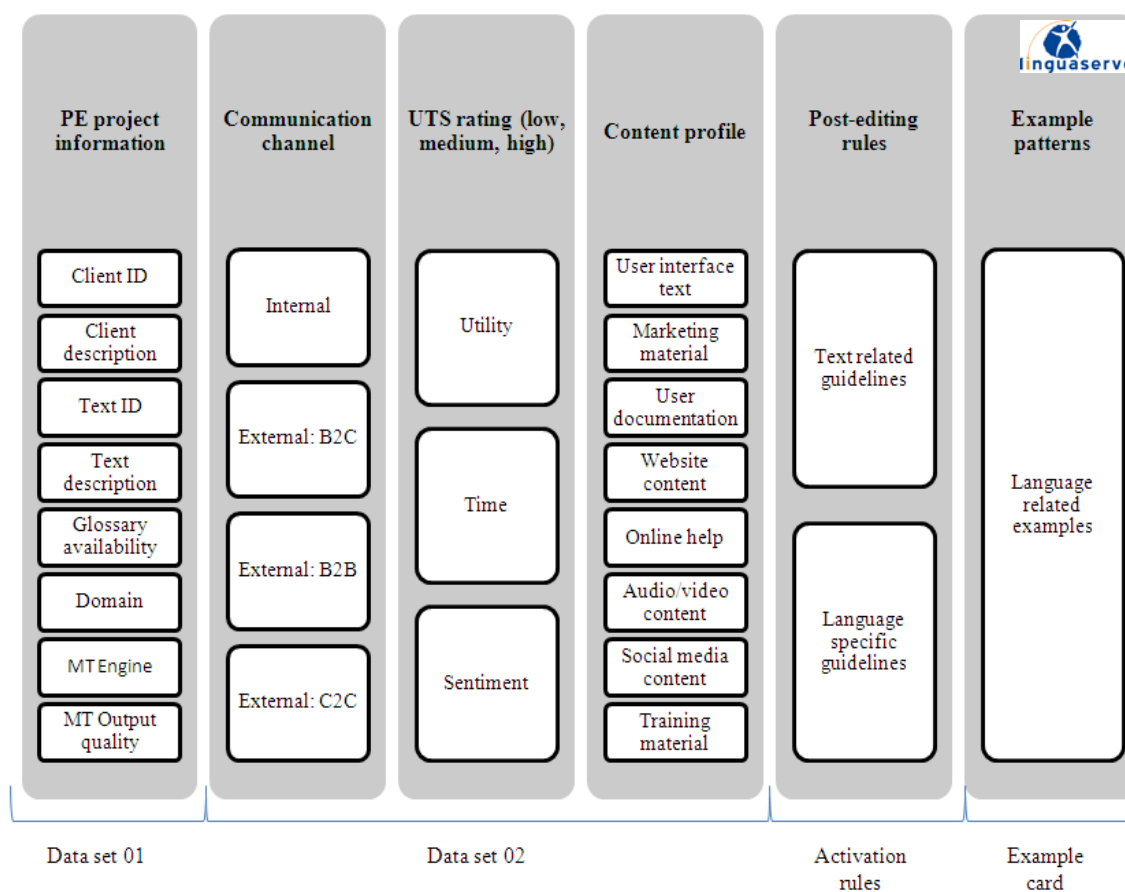- Example card: registers typical PE samples for each language pair.

Figure 3: elements to be considered in designing PE guidelines

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[3]

- *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

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

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 [4]

<u>Examples of rules for post-editing Spanish into English</u>

PE rule	*MT input*	*MT ouput*	*PE output*
Fix wrong terminology.	Sin derecho a deducción	Without **law** to deduction	Without deduction **right**
Fix morphological errors (number, gender, case, tense, voice).	El espacio más alto será la sala de ensayo de los castellers, que se sitúa en la planta del sótano,	The highest space will be the rehearsing **room** of the castellers, <A[which|who]> **are situated** in the plant of the basement,	The highest space will be the rehearsing **room** of the castellers, which **is situated** in the basement
Fix syntactic errors (wrong part of speech, incorrect phrase structure, wrong linear order).	Para aprender más cosas sobre la ciencia Planes de previsión asegurados	To **learn** more things **on** the science Plans of **forecast guaranteed**	To learn more things **about** science Plans of **guaranteed forecasts**
Fix any omissions as long as they interfere with the message transferred.	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.	I believe that a little more raw and more open, but seems me that quite faithful to the spirit of the <A[disk|record]>, to the instrumental <A[crosses|crossovers]> and to the vocal atmospheres.	I believe that a little rawer and more open, but **it** seems **to** me that quite true to the spirit of the record, the instrumental crossings and the vocal atmospheres
Fix stylistic problems only when needed.	…para ofrecer un escaparate de la **industria** española **del videojuego**	to offer a shop window of the Spanish **industry of the video game**	to offer a showcase of the Spanish **industry of the video game**

---

[4] A full set of detailed examples can be found in the document *EDI-TA Training Methodology*, which is complementary to this report.

Examples of rules for post-editing English into Spanish

PE rule	MT input	MT ouput	PE output
Fix wrong terminology.	The show is a fantasy-filled parade created by the imagination of a clown, the leading character	El espectáculo es un desfile lleno de fantasía creado por la imaginación de un payaso, el **carácter** principal.	El espectáculo es un desfile lleno de fantasía creado por la imaginación de un payaso, el **personaje** principal.
Fix grammatical mistakes.	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.	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.	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.
Fix morphological errors (number, gender, case, tense, voice).	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.	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.	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.
Fix syntactic errors (wrong part of speech, incorrect phrase structure, wrong linear order).	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.	Alba Sarraute vuelve a SAT! **con esto el payaso y cabaret enseñan** en cual viaja a Amazonia para revelar un mundo que todos nosotros tenemos profundo dentro de nosotros.	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.
Fix any omissions as long as they interfere with the message transferred.	Cirque du Soleil is coming back to Barcelona next January with their show Corteo.	El Cirque du Soleil está volviendo a Barcelona **(x)** próximo enero con su espectáculo Corteo.	El Cirque du Soleil está volviendo a Barcelona **el** próximo enero con su espectáculo Corteo.
Fix stylistic problems only when needed.	The Municipal government launches a plan to promote employment for young people without **jobs**.	El Gobierno Municipal lanza un plan para promover empleo para gente joven sin trabajos.	No cambiar "sin trabajos" por "desempleada".

Examples of rules for post-editing Spanish into French

PE rule	*MT input*	*MT ouput*	*PE output*
Fix wrong terminology.	Novedades para **ampliar público**	Nouveautés pour **agrandir du public**	Nouveautés pour **élargir le public**
Fix grammatical mistakes.	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".	De plus, en ce la Mairie **communiquée** aussi <A[il|elle]> 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".	**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 ».
Fix morphological errors (number, gender, case, tense, voice).	El estado de la arena, del mar, las banderas, las temperaturas y la insolación se pueden consultar en **el web de playas**	L'état du sable, de la mer, les drapeaux, les températures et l'insolation peuvent être consultés **en le web de plages**	L'état du sable, de la mer, les drapeaux, les températures et l'insolation peuvent être consultés sur **le site web des plages**.
Fix syntactic errors (wrong part of speech, incorrect phrase structure, wrong linear order).	Barcelona empieza la temporada de verano con 210 nuevos agentes desplegados **por las zonas** con más afluencia y un modelo coordinado con los Mossos d'Esquadra.	Barcelone commence la <M[saison d'été]> avec 210 nouveaux agents déployés <A[par|pour]> **les zones a**vec plus d'affluence et un modèle coordonné avec les <U[Mossos]> <U[d'Esquadra]>.	Barcelone commence la saison d'été avec 210 nouveaux agents déployés **dans les zones** ayant plus d'affluence et un modèle coordonné avec les Mossos d'Esquadra.
Fix any omissions as long as they interfere with the message transferred.	Los chiringuitos de Ciutat Vella **tienen autorización** para poner la televisión con el volumen limitado.	Les buvettes de <U[Ciutat Vella]> **ont autorisation** pour mettre la télévision avec le volume limité.	Les buvettes de Ciutat Vella **ont l'autorisation** pour mettre la télévision avec le volume limité.
Fix stylistic problems only when needed.	**Sube arriba. / Entra dentro** de la casa	**Il monte en haut. / Il entre dans** la maison	**Il monte en haut / Il entre dans la maison** (*the sentence remains as it is).*

Examples of rules for post-editing Spanish into Euskera

PE rule	MT input	MT ouput	PE output
Fix wrong terminology.	tipo de archivo	@A[artxibo|artxibatze]@-@A[tipo|gorpuzkera]@	fitxategi-mota
Fix morphological errors (number, gender, case, tense, voice).	La muestra, además, se acompaña de audiovisuales, grabados por el mismo autor, con el testimonio directo de familiares afectados.	Gainera, laguntzen dio erakusketari ikus-entzunezko**etatik**, egil**e** berak grabatuta, eragindako senide**etako** zuzen**eko** lekukotasunarekin.	Gainera, erakusketari ikus-entzunezko**ek** laguntzen di**ote**, egile**ak** berak grabatutakoak, eragindako senide zuzen**en** lekukotasunarekin.
Fix syntactic errors (wrong part of speech, incorrect phrase structure, wrong linear order).	En total, más de doscientas actuaciones repartidas por siete espacios de la ciudad, entre los cuales están la Pedrera o el Palau de la Música.	Guztira, **hiriaren zazpi espaziok**, zeinen artean Pedrera edo Palau de la Música daudenek, **banatutako berrehun emanaldi baino gehiago.**	Guztira, **berrehun emanaldi baino gehiago, hiriaren zazpi espaziotan banatuta,** horien artean Pedrera edo Palau de la Música.
Fix any omissions as long as they interfere with the message transferred.	Por defecto, cualquier PC con Windows 7 lo está.	Lehenetsi, Windows 7 duen edozein ordenagailu dago.	Lehenetsi, Windows 7 duen edozein ordenagailu **prestatuta** dago
Fix stylistic problems only when needed.	Es necesario que introduzcas tus datos correctamente al registrarte en el servicio.	Beharrezkoa da **zu** erregistratzerakoan **zuk zure** datuak zuzen zerbitzuan sartzea.	Beharrezkoa da zerbitzuan erregistratzerakoan zure datuak zuzen sartzea.

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

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.
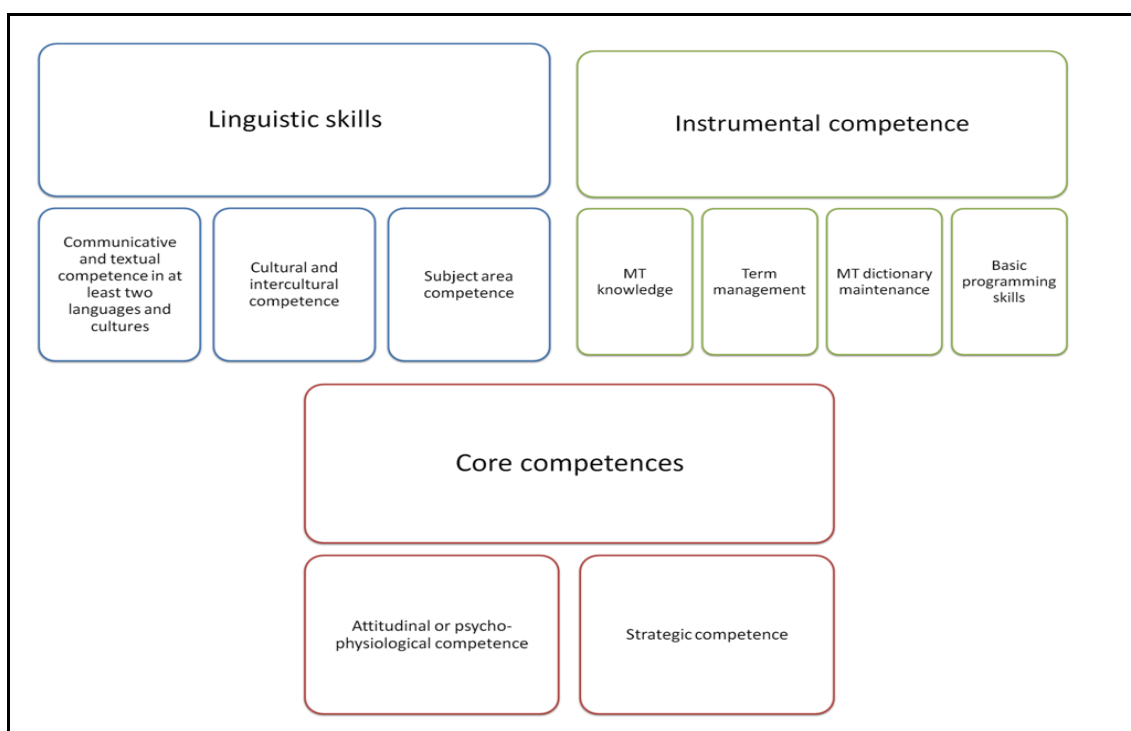

Figure 4: PE skills and competences
(Source: Rico and Torrejón, forthcoming)

- *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.
- MT output with relevant meta data for post-editing purposes, as described previously on page **¡Error! Marcador no definido.**.

TASKS & WORKFLOW

The following tasks are identified in the PE process[5]:

- *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**

Field	Description
Input*	Original segment which generates the error
MT output*	Machine Translated output
Correct Output*	Post-edited output

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
Subject area*	Domains 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**

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

An example of error classification is the following:

Error Category	Error type
Format	Incorrect capitalization
	Malformed
	Incorrect format conversion
Lexicon	Term choice (terminology)
	Problems with phraseological units / placements
	Translation of names or words in other languages
	Misspellings, typos and punctuation / use of nonstandard abbreviations
Morphology	Concordance: subject, gender, number
	Wrong Case: Genitive
	Wrong case: other cases
	Problem with gender
	Misuse of article
	Lack of preposition in source text
Syntax	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

Error Category	Error type
Translation	Untranslated words

- **PE guidelines comments template.**

PE guidelines template. Instructions

PE rule # [indicate rule number]: [write the name of the rule]	
MT input: EN	[Write here an example of the MT input which illustrates the rule]
MT output: ES	[Include here the MT output which illustrates the rule]
PE output: ES	[Write here the post-edited output]
Comments	This rule reads as follows: [write here the complete rule] [Explain the examples used above]

Example

| PE rule 01: fix any wrong term	
MT input: EN	The show is a fantasy-filled parade created by the imagination of a clown, the leading *character*
MT output: ES	El espectáculo es un desfile lleno de fantasía creado por la imaginación de un payaso, el *carácter* principal.
PE output: ES	El espectáculo es un desfile lleno de fantasía creado por la imaginación de un payaso, el *personaje* principal.
Comments	This rule reads as follows: "*Fix any wrong term in the text, either technical or non-technical term. Also, correct any inconsistent use of the same term*". In the example, the word *character* (EN) has been translated as *carácter* (ES). This is not a technical term but it needs to be changed to *personaje* (ES) as, otherwise, the translated sentence introduces an ambiguity not present in the source text. Additionally, this word needs post-editing because the final text is bound for publication.

- **Term registration template**

Term number identification: [indicate identification number]	
Source term	[Write here the source term]
Source term in context	[Write here an example of the term in context in the source text]
Machine translated term	[Write here the term as machine translated]
Target term	[Write here the correct term in the target language]
Target term in context	[Write here an example of the target term in context]
Comments	[Use this field to introduce any necessary comments]

Term number identification:001	
Source term	mobile sector
Source term in context	This was the first of a series of initiatives aimed at promoting and creating new work lines in the **mobile sector.**
Machine translated term	Éste fue la primera de una serie de iniciativas de las cuales se esperaba que promovieran y crearan nuevas líneas de trabajo en el **sector ambulante|cambiable**.
Target term	sector de telefonía móvil
Target term in context	[…] crearan nuevas líneas de trabajo en el **sector de la telefonía móvil.**
Comments	

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 & WORKFLOW

Tasks are organized along the following workflow:
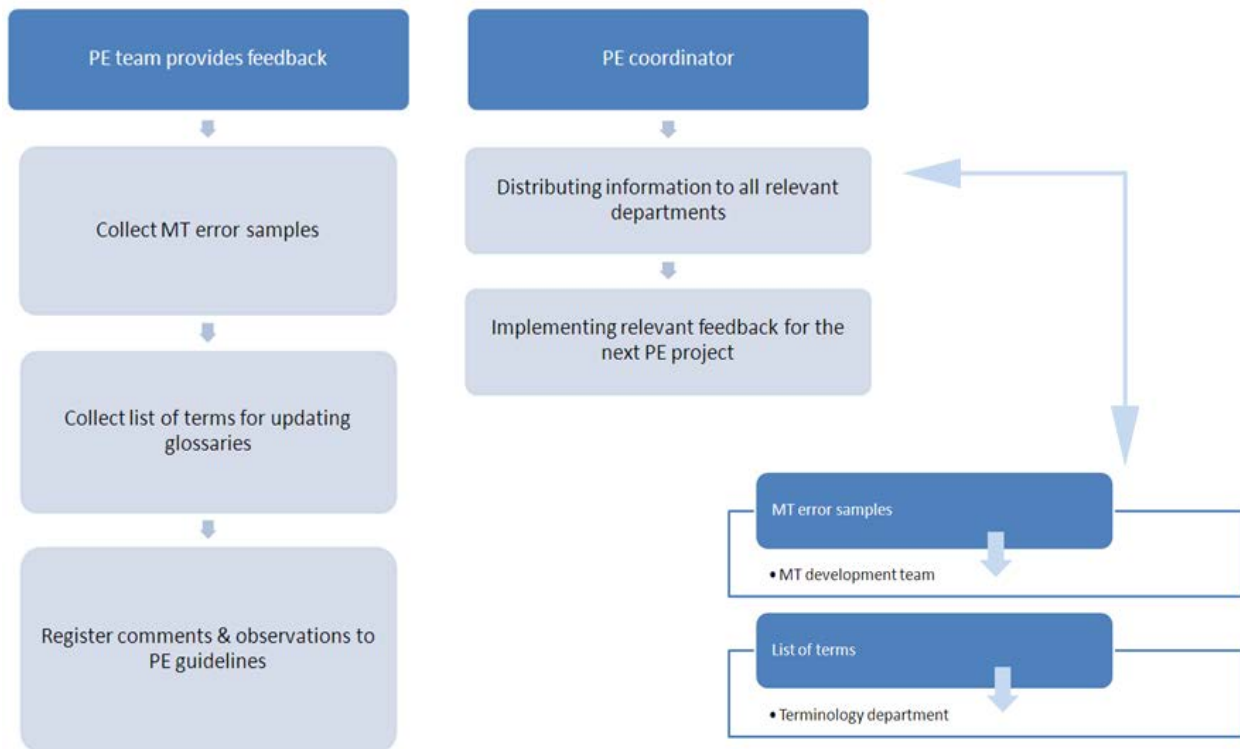
4. REQUIREMENTS FOR A POST-EDITING TOOL

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 publised. 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.
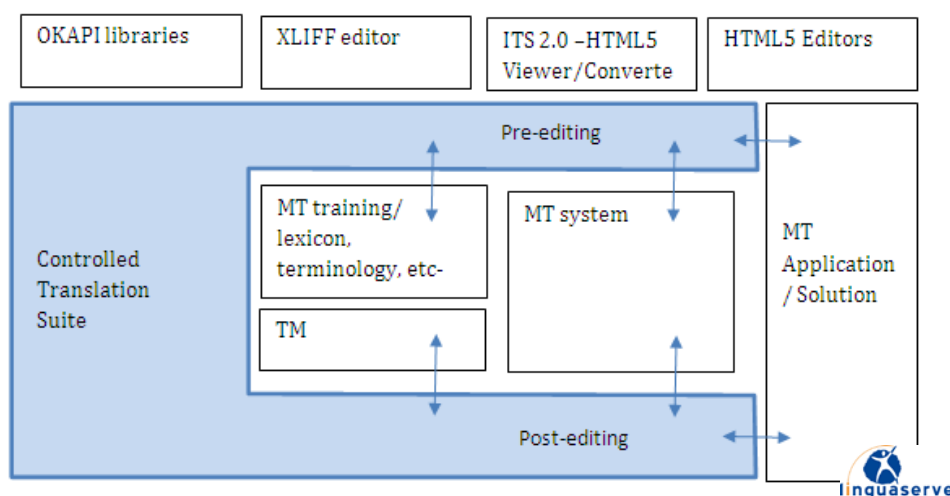

Figure 5: Post-editing in a Controlled Translation

---

[6] The tools analyzed were Boltran, PET, Globalsight, OmegaT and CAITRA

5. REFERENCES

Allen, J. (2003) "Post-editing" in Somers, H. ed., *Computers and Translation. A translator's guide*. Amsterdam/Philadelphia: John Benjamins. 297-317.

Guzmán, R. (2007) "Manual MT Post-editing: if it's not broken, don't fix it" *Translation Journal*, vol. 11, 4. October 2007.

Krings, H.P (2001) *Repairing Texts. Empirical Investigations of Machine Translation Pos-Editing Processes*. Kent: The Kent State University Press.

O'Brien, S. (2002): "Teaching post-editing: a proposal for course content". *6th EAMT Workshop, Teaching Machine Translation, November 14-15, 2002. European Association for Machine Translation*. 99-106. [online] Available: http://www.mt-archive.info/EAMT-2002-OBrien.pdf [accessed 13 September 2012]

O'Brien, S. (2012) "Towards a Dynamic Quality Evaluation Model for Translation" *The Journal of Specialised Translation*, Issue 17, Jan. 2012

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*

Torrejón Díaz, E. y C. Rico Pérez (2002) "Controlled Translation: A new teaching scenario tailor-made for the translation industry" *6th EAMT Workshop, Teaching Machine Translation*, November 14-15, 2002. European Association for Machine Translation, pp. 107-116

Vasconcellos, M. (1987): "A comparison of MT postediting and traditional revision". In *Proceedings of the 28th Annual Conference of the American Translators Association*. Ed. K. Kummer, 409-415. Medford, NJ: Learned Information.