Post-editing practices and the multilingual web: sealing gaps in best practices and standards

Dra. Celia Rico
W3C Workshop: New Horizons for the Multilingual Web
7-8 May 2014, Madrid
Definition of PE
“the correction of machine-generated translation output to ensure it meets a level of quality negotiated in advance between client and vendor”
(TAUS/CNGL)
EDI-TA’s fact sheet

Objectives

a) define the functionalities for a post-editing tool
b) design a methodology for training post-editors
c) analyze the economic impact of implementing post-editing processes

Project’s setting

• 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 project coordinator

Definition of PE

“the correction of machine-generated translation output to ensure it meets a level of quality negotiated in advance between client and post-editor” (TAUS/CNGL)
Is there a real benefit in using standards for post-editing purposes in daily practice?
Do annotation tags make sentences slightly less understandable and more cryptic for post-editors?
In cases where there is more than one annotation per phrase, the post-editor may miss the visual continuity of the sentence, spend too much time rereading it or even leave syntax mistakes from the MT uncorrected. How should this information be presented (if at all)?
Should post-editors be allowed to insert annotations?
The case for Post-editing as multilingual Web enabler
Growth

- Common Sense Advisory calculates that the market for outsourced language services is worth US$34.778 billion in 2013 (1,022 companies surveyed)
- As of 2013, Common Sense Advisory calculates that the language services market is growing at an annual rate of 5.13%.
“150-200% more words Using Machine Translation and Post-Editing instead of human translation when working with a translation agency could mean translating 150-200% more words for the same money”
Volume + Turnaroud + Prices = Let’s automate
At some point you need a person checking MT output!

Post-production?

– **No post-editing**: internal documentation, browsing, gisting, tightly controlled languages, KBs with customised MT

– **Rapid post-editing**: perishable information and urgent texts (only serious errors are fixed)

– **Partial post-editing**: minimum changes

– **Full post-editing**: complete revision (external publication)
Post-editing and ITS 2.0
ITS 2.0

- Facilitating automated creation and processing of web content
- Defining metadata for language technology in the Web (MT, Localization)
- Metadata needed for web content (HTML5), deep Web (XML), Localization formats (XLIFF)

Adding value to content
ITS 2.0 metatags were reviewed in terms of PE needs.
Post-editing information
<table>
<thead>
<tr>
<th>Text related guidelines</th>
<th>&lt;indicate whether this rule should be activated&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fix wrong terminology</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Spend time in terminology research</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Fix syntactic errors (wrong part of speech, incorrect phrase structure, wrong linear order)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Fix morphological errors (number, gender, case, tense, voice)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Fix misspelling errors</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Fix punctuation errors</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Fix any omissions as long as they interfere with the message transferred</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Edit any offensive, inappropriate or culturally unacceptable information</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Fix any problem related to textual standards (cohesion, coherence)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Fix stylistic problems</strong></td>
<td></td>
</tr>
</tbody>
</table>
Language dependent rules

- *LS EN-ES PE Rule 01.* Replace upper-case letters for low-case letters, when applicable.
- *LS EN-ES PE Rule 02.* Time format.
- *LS EN-ES PE Rule 03.* Date format.
- *LS EN-ES PE Rule 04.* Change order of figures when used as adjectives.
- *LS EN-ES PE Rule 05.* Correct –ING adjectives by translating them as adjectives or relative clauses.
- *LS EN-ES PE Rule 06.* Translate –ING forms as infinitive forms, when used as subject.
- *LS EN-ES PE Rule 07.* Translate the infinitive phrase ‘to be + infinitive’ with a future tense.
- *LS EN-ES PE Rule 08.* Translate the present continuous with a future tense, when used to refer to a future event with a future tense.
- *LS EN-ES PE Rule 10.* Replace the “de” preposition if appearing excessively in the text.
- *LS EN-ES PE Rule 11.* Insert articles when necessary to convey the meaning.
- *LS EN-ES PE Rule 12.* Translate ‘for’ as *para/por* as the case may be.
Text related guidelines

- 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)
- Fix stylistic problems

Language dependent rules

- LS EN-ES PE Rule 01. Replace upper-case letters for low-case letters, when applicable.
- LS EN-ES PE Rule 02. Time format.
- LS EN-ES PE Rule 03. Date format.
- LS EN-ES PE Rule 04. Change order of figures when used as adjectives.
- LS EN-ES PE Rule 05. Correct -ING adjectives by translating them as adjectives or relative clauses.
- LS EN-ES PE Rule 06. Translate -ING forms as infinitive forms, when used as subject.
- LS EN-ES PE Rule 07. Translate the infinitive phrase 'to be + infinitive' with a future tense.
- LS EN-ES PE Rule 08. Translate the present continuous with a future tense, when used to refer to a future event with a future tense.
- LS EN-ES PE Rule 09. Correct translation for verbs 'estar/ser'.
- LS EN-ES PE Rule 10. Replace the "de" preposition if appearing excessively in the text.
- LS EN-ES PE Rule 11. Insert articles when necessary to convey the meaning.
- LS EN-ES PE Rule 12. Translate 'for' as para/por as the case may be.
# Mapping tags and rules

<table>
<thead>
<tr>
<th>Data category</th>
<th>PE purposes</th>
<th>PE rule activation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translate</td>
<td>Informing the post-editor of sentences or sentence fragments should or should not be translated</td>
<td>Block text when NO post-editing is to be done</td>
</tr>
<tr>
<td>Localization note</td>
<td>Providing post-editors with the necessary information to review the text in order to help them disambiguate and improve the quality and accuracy of the revision. <strong>Utility</strong> (relative importance of the functionality of the translated content). <strong>Delivery Time</strong> (speed with which the translation is required). <strong>Sentiment</strong> (importance on brand image).</td>
<td>Trigger PE rules (from zero to full PE) according to text functionality, delivery time and importance of brand image (O’Bien, 2012; Rico, 2012)</td>
</tr>
<tr>
<td>Language information</td>
<td>Points to part of content in a language different from the rest, which could require MT and post-editing for a specific language pair.</td>
<td>Block text when NO post-editing is to be done</td>
</tr>
<tr>
<td>Domain</td>
<td>It enables automatic selection of MT terminology, post-editor selection, and is a key to content disambiguation.</td>
<td>Check domain &amp; disambiguate when necessary</td>
</tr>
</tbody>
</table>
# Mapping tags and rules

<table>
<thead>
<tr>
<th>Data category</th>
<th>PE purposes</th>
<th>PE rule activation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provenance</td>
<td>Assessing how translation agents may impact the quality of the translation. Translation and translation revision agents can be identified as a person, a piece of software or an organization that has been involved in providing a translation that resulted in the selected content.</td>
<td>Confirm provenance</td>
</tr>
<tr>
<td>Localization</td>
<td>Detecting possible localization issues such as: Terminology, Mistranslation, Omission, Untranslated, Addition, Duplication, Grammar, Legal, Register, Locale specific content, Locale violation, Style, Characters, Misspelling...</td>
<td>Trigger PE rules accordingly</td>
</tr>
<tr>
<td>MT Confidence</td>
<td>Confidence score for each translated segment. Those above a certain threshold will be blocked for no post-editing</td>
<td>Prevent text modification above a certain threshold</td>
</tr>
</tbody>
</table>
Maximising the post-editor’s interface

- Postediting is usually carried out in a Translation Environment
- Lack of support may lead to cognitive friction (Moorkens and O’Brien, 2013)
- Simplicity and customizability
- Provenance of MT or TM suggestions kept separate
- Meta-data showing the origin of match suggestions is important to translators and post-editors
- UI clean and uncluttered
- Show only those tags with relevant information for taking a decision
- Use a color code but not on tags (which might distract attention) but on PE rules
- Show tags only when post-editor deems it necessary
Conclusion

There is a case for Post-editing as a multilingual web enabler

At some point you need a person checking MT output

ITS 2.0 facilitates automation

Keep it clean and simple
References


