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

Universidad Europea

LAUREATE INTERNATIONAL UNIVERSITIES

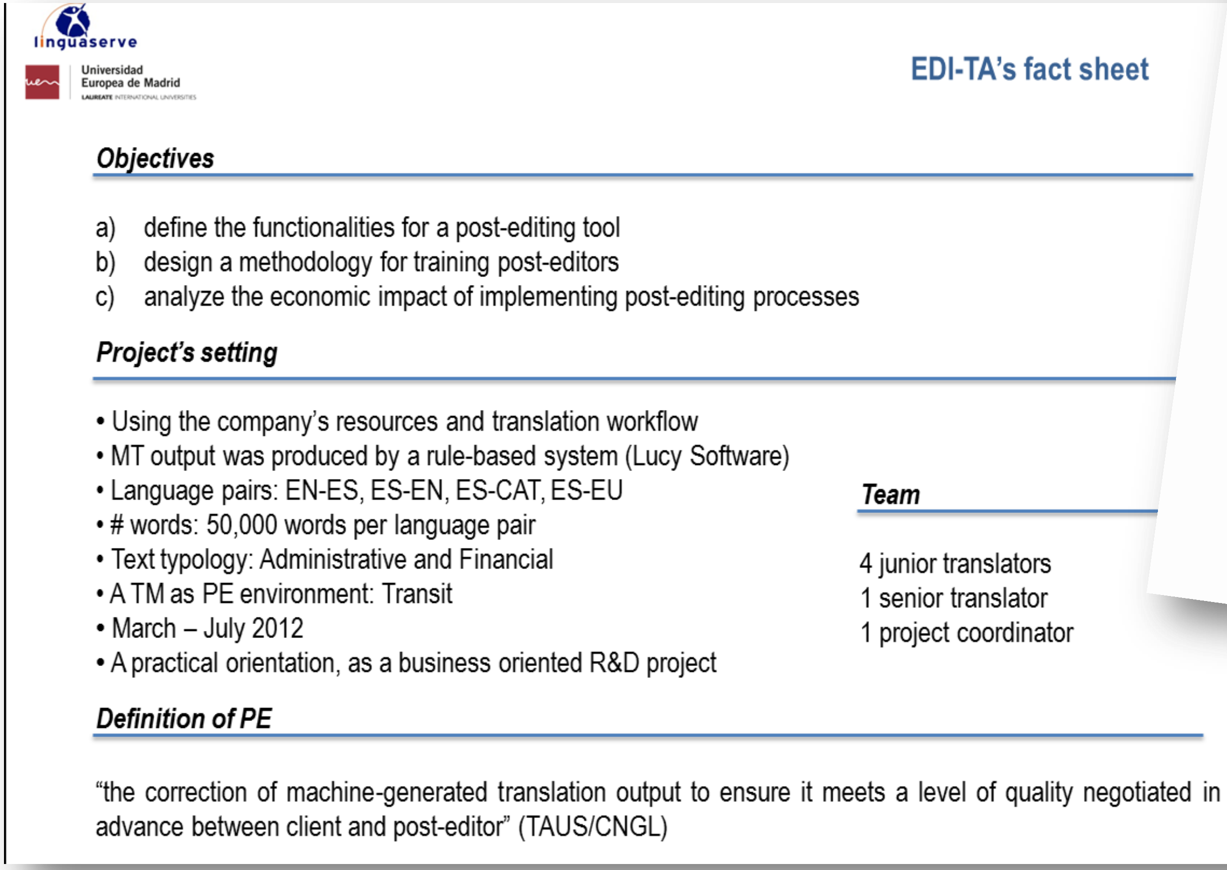
Dra. Celia Rico

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) http://www.w3.org/International/multilingualweb/lt/wiki/Deliverables



linguaserve

Universidad Europea de Madrid
LAUREATE INTERNATIONAL UNIVERSITIES

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)

MultilingualWeb-LT

D4.1.4 - ANNEX II
TRAINING METHODOLOGY FOR MACHINE TRANSLATION POST-EDITING

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

Distribution: Public

MultilingualWeb-LT

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

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

Distribution: Public

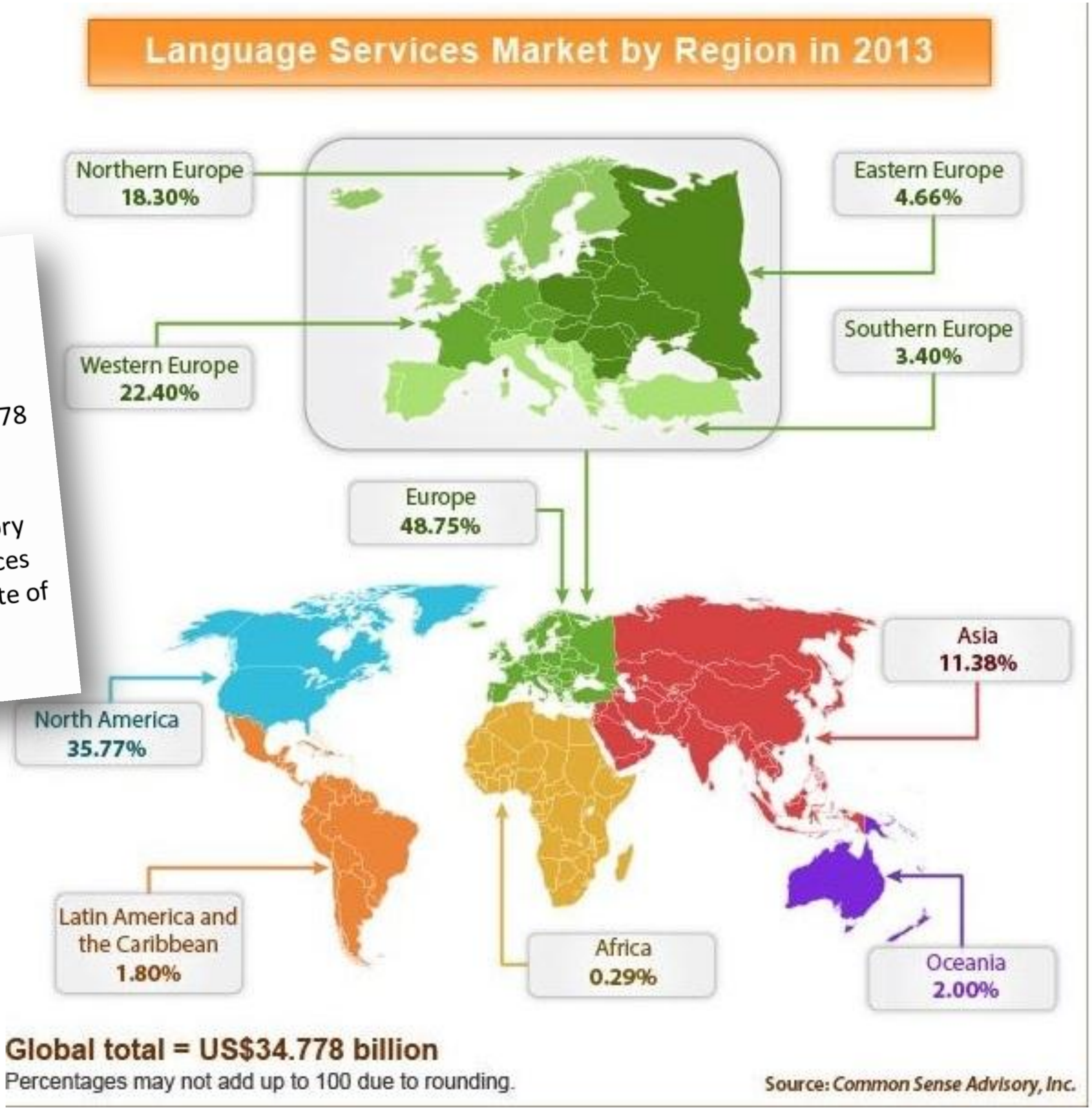
MultilingualWeb-LT (LT-Web)
Language Technology in the Web
FP7-ICT-2011-7

- □ 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

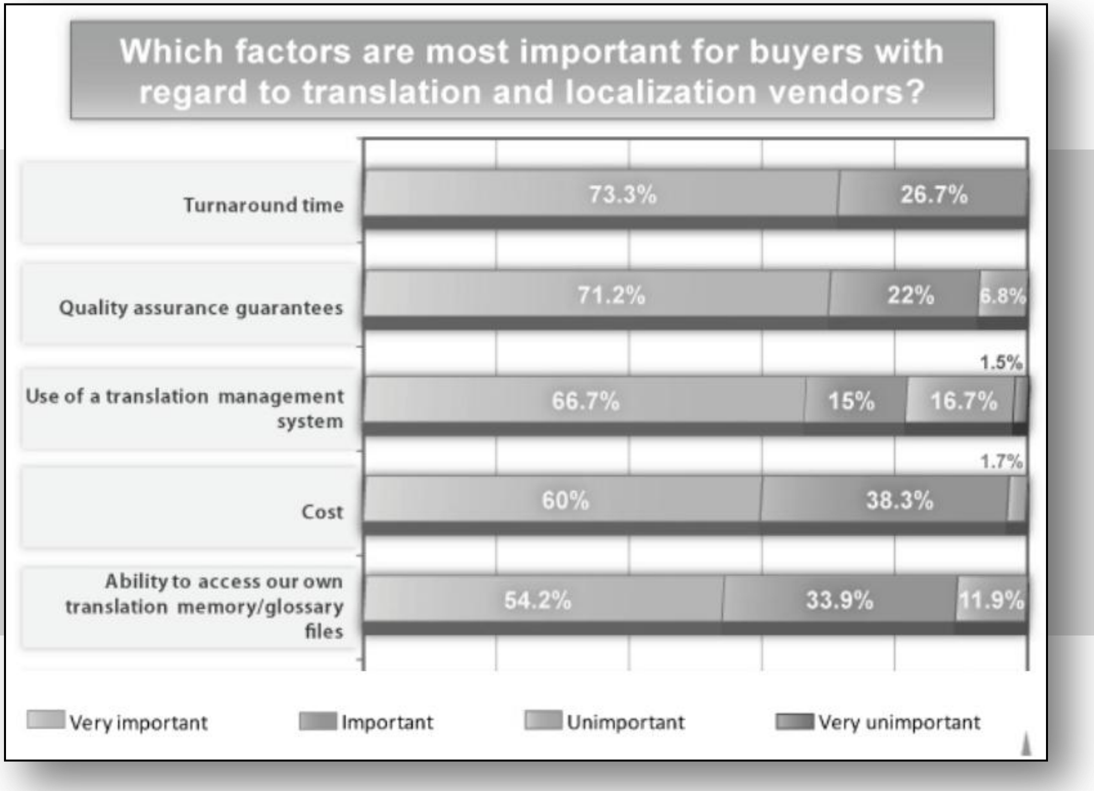
- Common Sense Advisory calculates that the market for outsourced language services is worth US$34.778 billion in 2013 (1,022 companies surveyed)
- As of 2013, Common Sense Advisory calculates that the language services market is growing at an annual rate of 5.13%.


Language Services Market by Region in 2013

Global total = US$34.778 billion

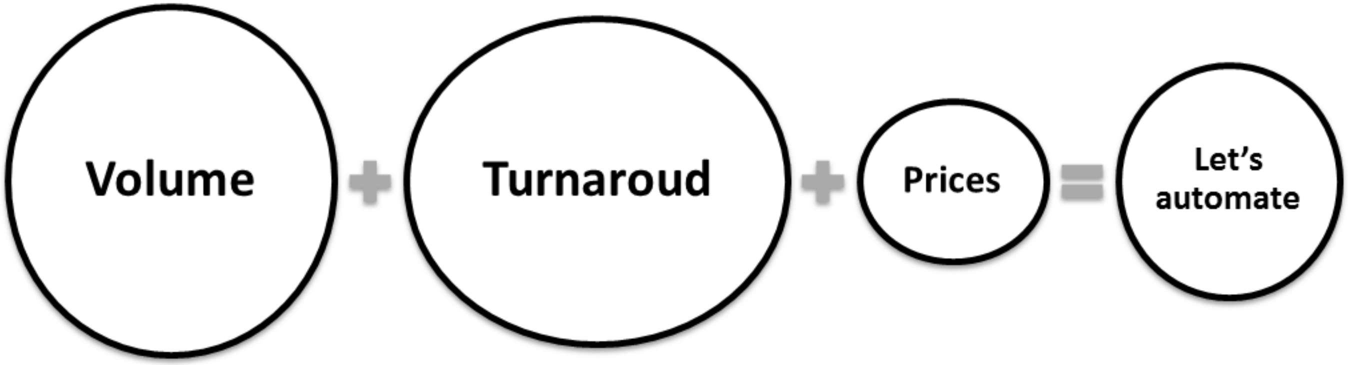
Percentages may not add up to 100 due to rounding.

Source: *Common Sense Advisory, Inc.*



Source: *Trends in Machine Translation*, Common Sense Advisory, 2011

“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

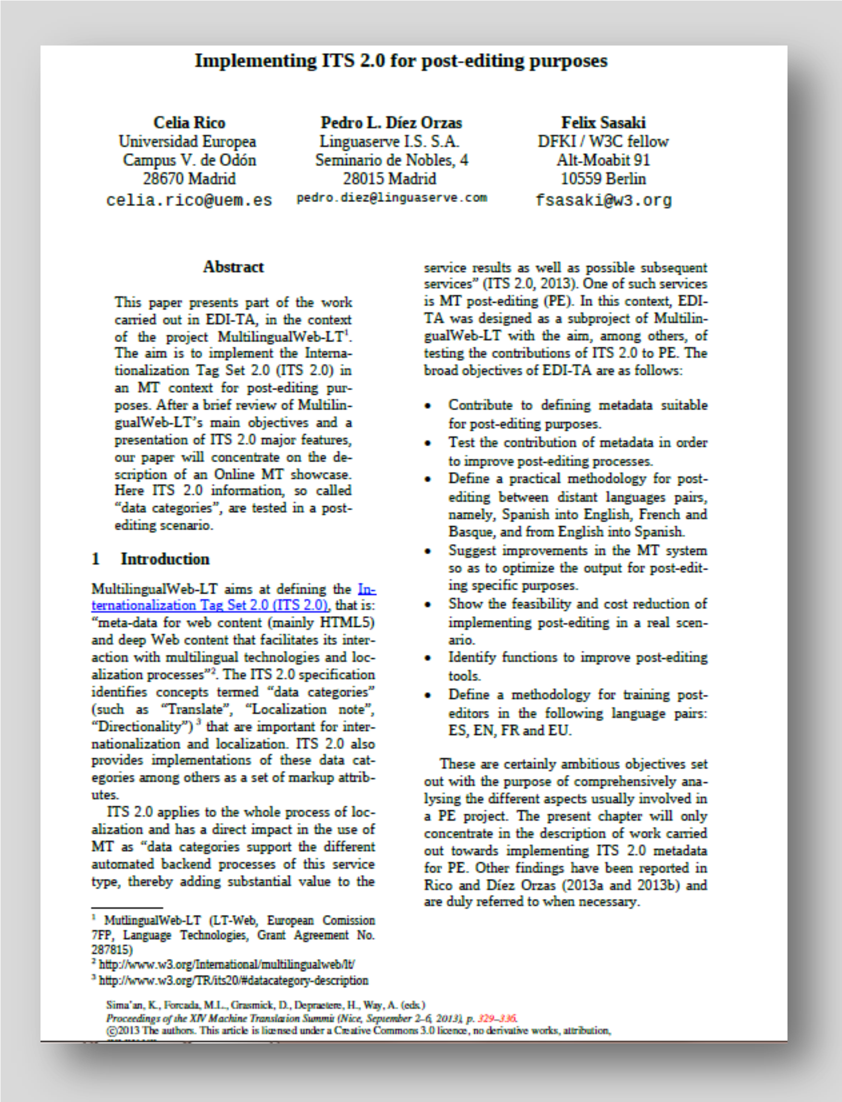
ITS 2.0

- 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

What info?

http://www.mtsummit2013.info/files/proceedings/main/mt-summit-2013-rico-et-al.pdf



Implementing ITS 2.0 for post-editing purposes

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

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

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

Abstract

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

1 Introduction

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

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

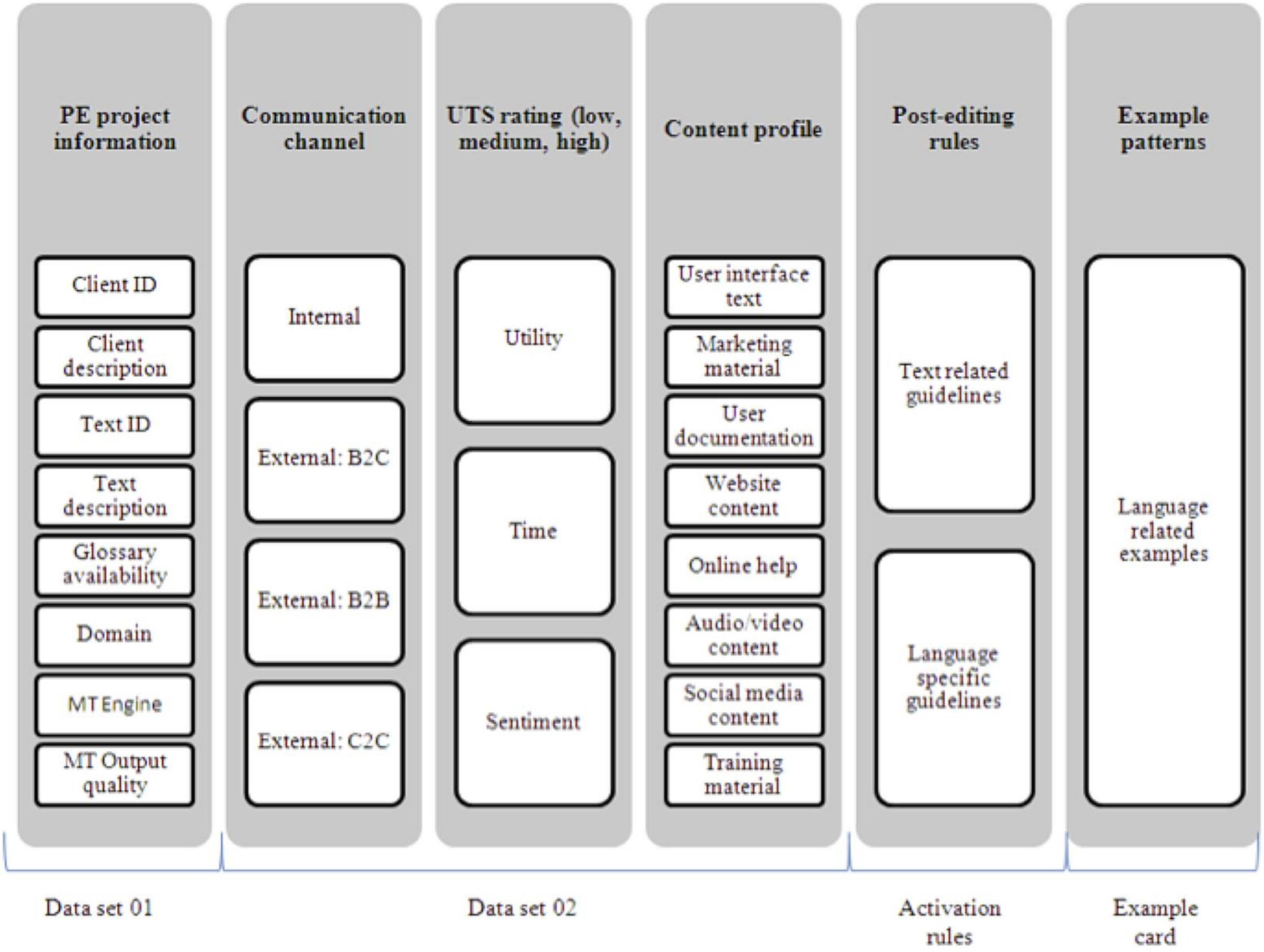
- Contribute to defining metadata suitable for post-editing purposes.
- Test the contribution of metadata in order to improve post-editing processes.
- Define a practical methodology for post-editing between distant languages pairs, namely, Spanish into English, French and Basque, and from English into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

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

[1] MutlingualWeb-LT (LT-Web, European Comission 7FP, Language Technologies, Grant Agreement No. 287815)
[2] http://www.w3.org/International/multilingualweb/lt/
[3] http://www.w3.org/TR/its20/#datacategory-description

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 329–336.

ITS 2.0 metatags were reviewed in terms of PE needs



PE project information

- Client ID
- Client description
- Text ID
- Text description
- Glossary availability
- Domain
- MT Engine
- MT Output quality

Communication channel

- Internal
- External: B2C
- External: B2B
- External: C2C

UTS rating (low, medium, high)

- Utility
- Time
- Sentiment

Content profile

- User interface text
- Marketing material
- User documentation
- Website content
- Online help
- Audio/video content
- Social media content
- Training material

Post-editing rules

- Text related guidelines
- Language specific guidelines

Example patterns

- Language related examples

Data set 01

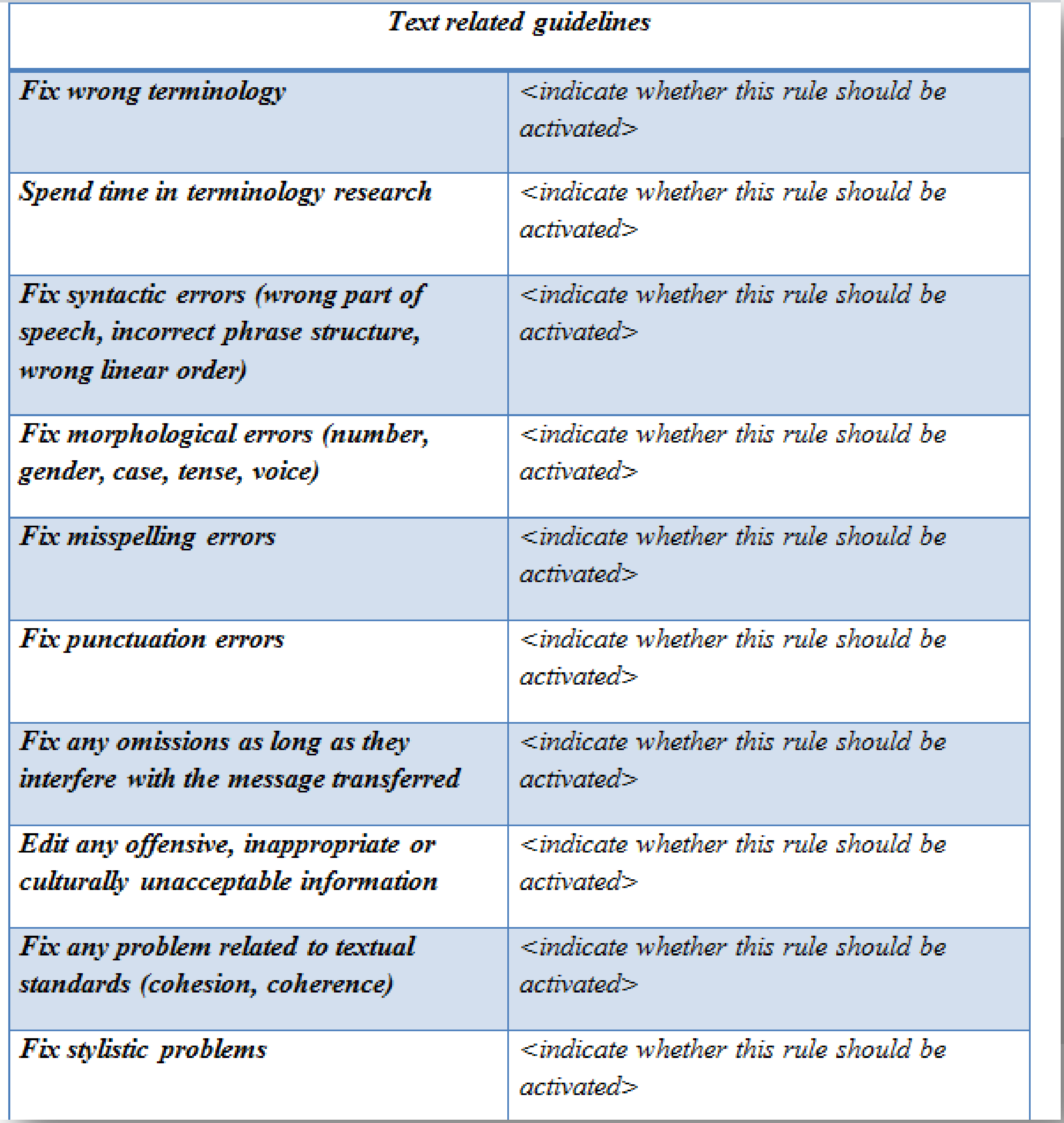
Data set 02

Activation rules

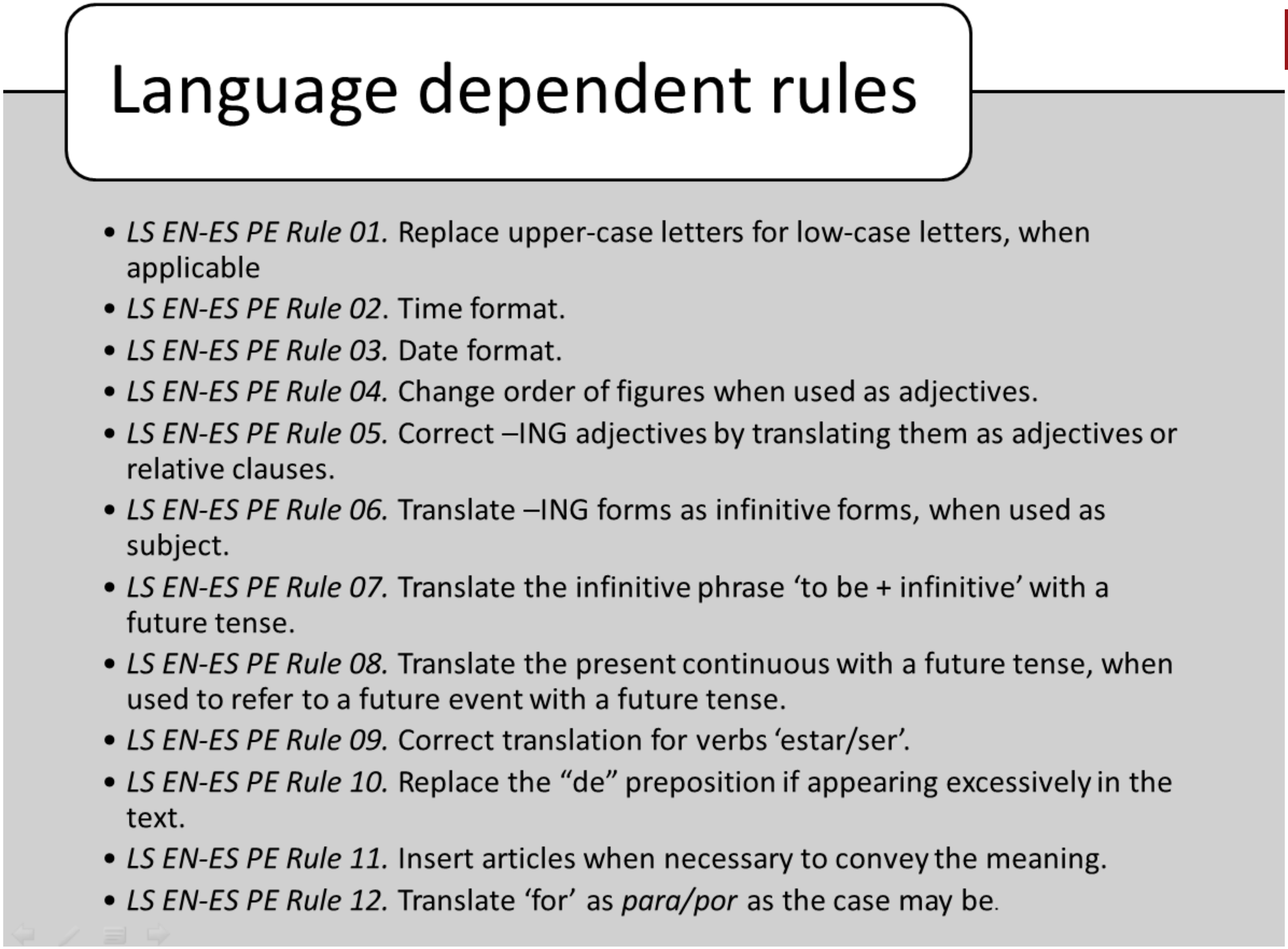
Example card

Post-editing information

http://www.localisation.ie/resources/locfocus/LocalisationFocusVol11_1Web.pdf

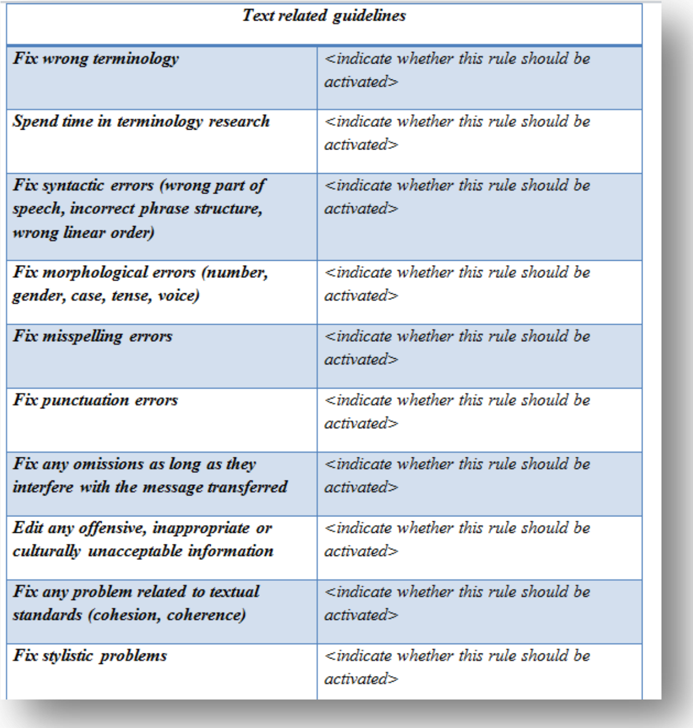


Text related guidelines	
Fix wrong terminology	*<indicate whether this rule should be activated>*
Spend time in terminology research	*<indicate whether this rule should be activated>*
Fix syntactic errors (wrong part of speech, incorrect phrase structure, wrong linear order)	*<indicate whether this rule should be activated>*
Fix morphological errors (number, gender, case, tense, voice)	*<indicate whether this rule should be activated>*
Fix misspelling errors	*<indicate whether this rule should be activated>*
Fix punctuation errors	*<indicate whether this rule should be activated>*
Fix any omissions as long as they interfere with the message transferred	*<indicate whether this rule should be activated>*
Edit any offensive, inappropriate or culturally unacceptable information	*<indicate whether this rule should be activated>*
Fix any problem related to textual standards (cohesion, coherence)	*<indicate whether this rule should be activated>*
Fix stylistic problems	*<indicate whether this rule should be activated>*

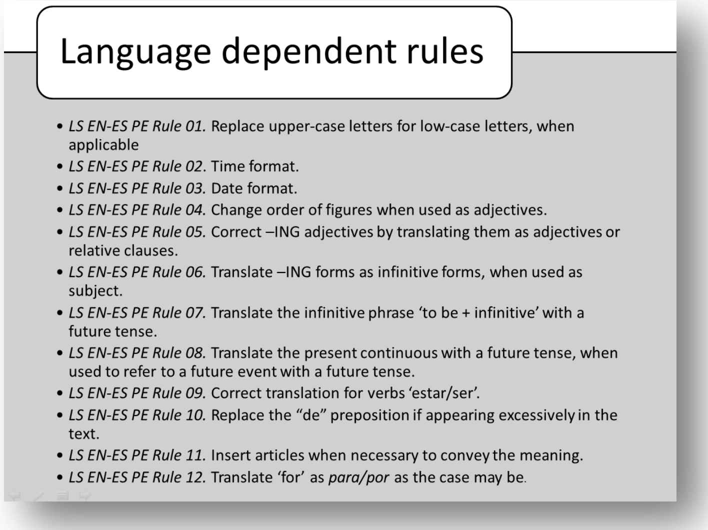


Language dependent rules

- *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.



Text related guidelines	
Fix wrong terminology	*<indicate whether this rule should be activated>*
Spend time in terminology research	*<indicate whether this rule should be activated>*
Fix syntactic errors (wrong part of speech, incorrect phrase structure, wrong linear order)	*<indicate whether this rule should be activated>*
Fix morphological errors (number, gender, case, tense, voice)	*<indicate whether this rule should be activated>*
Fix misspelling errors	*<indicate whether this rule should be activated>*
Fix punctuation errors	*<indicate whether this rule should be activated>*
Fix any omissions as long as they interfere with the message transferred	*<indicate whether this rule should be activated>*
Edit any offensive, inappropriate or culturally unacceptable information	*<indicate whether this rule should be activated>*
Fix any problem related to textual standards (cohesion, coherence)	*<indicate whether this rule should be activated>*
Fix stylistic problems	*<indicate whether this rule should be activated>*



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.

ITS 2.0 data categories

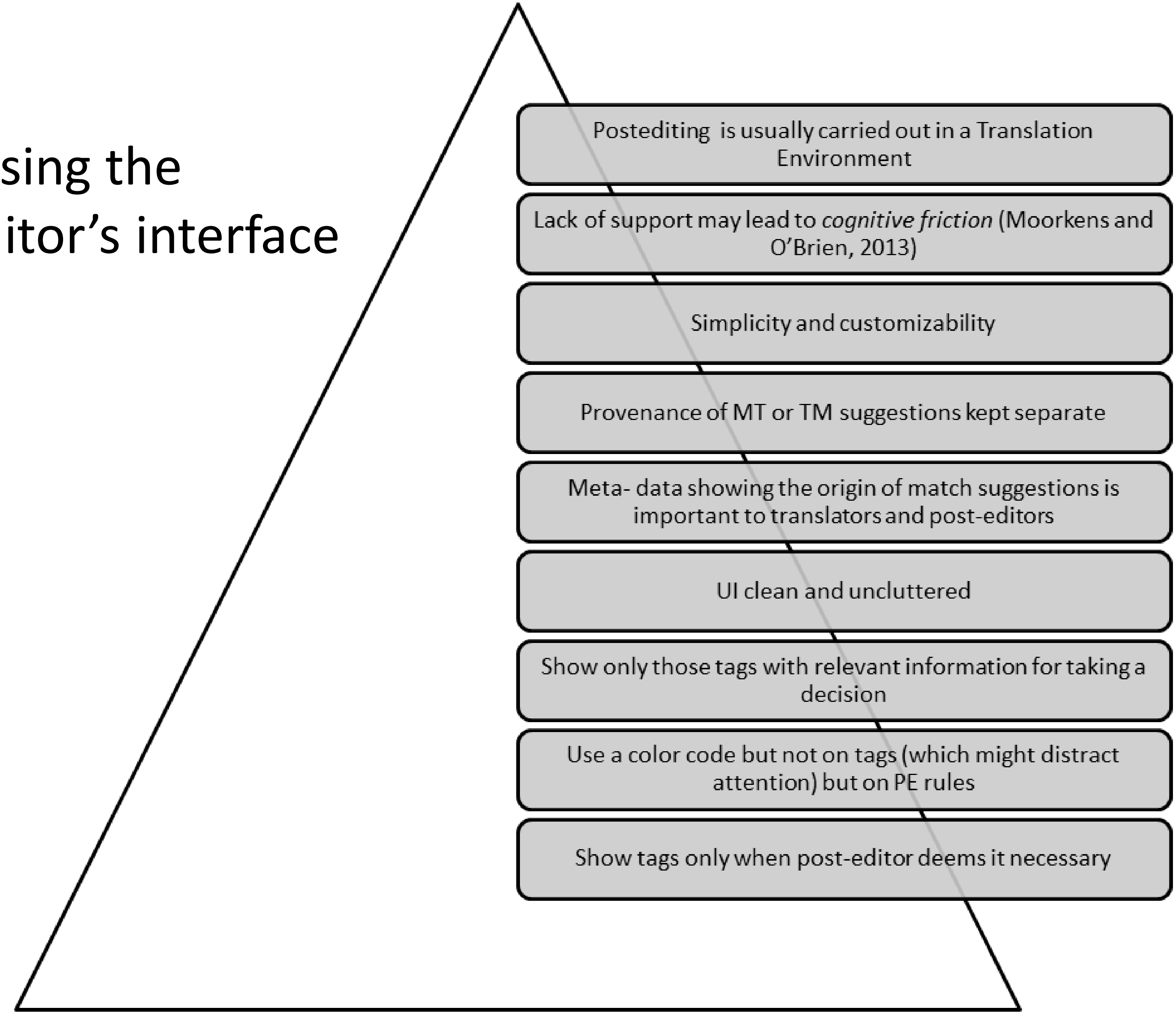
Mapping tags and rules

Data category	PE purposes	PE rule activation
Translate	Informing the post-editor of sentences or sentence fragments should or should not be translated	Block text when NO post-editing is to be done
Localization note	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. Utility (relative importance of the functionality of the translated content). Delivery Time (speed with which the translation is required). Sentiment (importance on brand image).	Trigger PE rules (from zero to full PE) according to text functionality, delivery time and importance of brand image (O'Brien, 2012; Rico, 2012)
Language information	Points to part of content in a language different from the rest, which could require MT and post-editing for a specific language pair.	Block text when NO post-editing is to be done
Domain	It enables automatic selection of MT terminology, post-editor selection, and is a key to content disambiguation.	Check domain & disambiguate when necessary

Mapping tags and rules

Data category	PE purposes	PE rule activation
Provenance	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.	Confirm provenance
Localization quality issue	Detecting possible localization issues such as: Terminology, Mistranslation, Omission, Untranslated, Addition, Duplication, Grammar, Legal, Register, Locale specific content, Locale violation, Style, Characters, Misspelling…	Trigger PE rules accordingly
MT Confidence	Confidence score for each translated segment. Those above a certain thresold will be blocked for no post-editing	Prevent text modification above a certain thresold

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

- Bikmatov, R., N. Glenn , S. Gladkoff, A. Melby (2013): “Visualization of ITS 2.0 Metadata for Localization Process”, *Localization Focus*, vol. 12, 1: 74-77
- Moorkens, J, and s. O’Brien, 2013: **“**User Attitudes to the Post-Editing Interface” Sharon O’Brien, Michel Simard and Lucia Specia (eds*.): Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*, Nice, September 2, 2013, p. 19–25.
- O’Brien, S. (2012) “Towards a Dynamic Quality Evaluation Model for Translation” *The Journal of Specialised Translation*, Issue 17, Jan. 2012
- Rico, C. and Díez Orzas, P.L. (2013): “EDI-TA: Training methodology for Machine Translation Post-editing”, *Multilingualweb-LT Deliverable 4.1.4. Annex II*, public report, Available: http://www.w3.org/International/multilingualweb/lt/wiki/images/d/d4/D4.1.4.Annex_II_EDI-TA_Training.pdf [05/05/2014]
- Rico, C. adn Díez Orzas, P.L. (2013): “EDI-TA: Post-editing methodology for Machine Translation”, *Multilingualweb-LT Deliverable 4.1.4. Annex I*, public report, Available: http://www.w3.org/International/multilingualweb/lt/wiki/images/1/1f/D4.1.4.Annex_I_EDI-TA_Methology.pdf [05/05/2014]
- Rico, C., P.L. Díez Orzas and F. Sasaki (2013): “Implementing ITS 2.0 for post-editing purposes”, *Proceedings of the MT Summit 2013,* 2-6- Sept. Nice, France. http://www.mtsummit2013.info/files/proceedings/main/mt-summit-2013-rico-et-al.pdf [05/05/2014]
- Rico, C. (2012): “A Flexible Decision Tool for Implementing Post-editing Guidelines”, *Localisation Focus*, vol. 11, 1: 54-66 http://www.localisation.ie/resources/locfocus/LocalisationFocusVol11_1Web.pdf [05/05/2014]

Universidad Europea

LAUREATE INTERNATIONAL UNIVERSITIES

Madrid Valencia Canarias