Qualitative Analysis of Post-Editing for High Quality Machine Translation

Frédéric Blain* and **Jean Senellart**
Systran SA
Paris, France
{blain,senellart}@systran.fr

Holger Schwenk
*LIUM, University of Le Mans
Le Mans, France
schwenk@lium.univ-lemans.fr

Mirko Plitt
Autodesk, Development Sàrl
Neuchâtel, Switzerland
mirko.plitt@autodesk.com

Johann Roturier
Symantec Limited
Blanchardstown, Dublin 15, Ireland
johann_roturier@symantec.com

Abstract

In the context of massive adoption of Machine Translation (MT) by human localization services in Post-Editing (PE) workflows, we analyze the activity of post-editing high quality translations through a novel PE analysis methodology. We define and introduce a new unit for evaluating post-editing effort based on *Post-Editing Action* (PEA) - for which we provide human evaluation guidelines and propose a process to automatically evaluate these PEAs. We applied this methodology on data sets from two technologically different MT systems. In that context, we could show that more than 35% of the remaining effort can be saved by introducing of global PEA and edit propagation.

1 Introduction

During the last five years, Machine Translation (MT) providers - boosted by corpus-driven approaches - have renewed their offers and started presenting "highly customized" translation solutions for specific domains/usages. Typically for technical documentation and online technical assistance material. Multiple reports attest the reality of this activity and market analysis shows a trend for language service providers (42% in 2010 according to DePalma and Hegde (2010)) to offer post-edited MT to their customers. Large corporations like Symantec, Autodesk, Cisco are also turning to Post-Editing (PE) as a way to reduce cost and "time-to-market" (Roturier, 2009). Machine PE is a new domain and differs from traditional translation reviewing by the nature of the errors to correct.

This paper strives to analyze formally PE activity based on actual data built from several types of translation engines along the following questions:

1. What's left for post-editing analysis?
2. How can we measure post-editing effort?
3. Can we reduce the effort?

In the following sections, we describe existing work with some answer elements for these different questions.

1.1 What's left for post-editing analysis?

Several studies consider PE through MT error classification in order to better rationalize PE effort: *"familiarity with the pattern of errors produced by a particular MT system is an important factor in reducing post-editing time"* (Martinez, 2003). However describing the errors does not provide us with a methodology for fixing them and always leads to system-dependent remediation approaches. In our approach, we are less interested in understanding the errors than defining the correct action to obtain a good translation.

1.2 How can we measure post-editing effort?

The measure of PE effort is important from a business perspective since it sets up the productivity of post-editors and subsequently the potential for additional cost-saving. The most criteria are the measure of PE time (Specia, 2011) or the comparison with human translation (Plitt and Masselot, 2010). Other approaches take into account "user activity data" covering keystrokes (Barrett et al., 2001) or eye movement detection (Doherty et al., 2010).

Implicitly, estimating PE effort is the driver for establishing better quality evaluation metrics. For instance HTER (Snover et al., 2006) calculating translation edit rate towards *targeted reference translation* provides a reproducible metric, well correlated with human judgment on translation quality and close by definition to "translation post-editing".

In WMT09, Callison-Burch et al. (2009) introduced a new task: editing to evaluate translation where the edited translation is not used as a reference nor the reviewer asked to perform the least number of edits, but to make the translation fluent without access to reference translation. The edited translation is then evaluated in a second phase of the evaluation task. However, the result of this task is not conclusive due to the variability between post-editors, and no strong correlation is observed with sentence quality judgment.

With METEOR, Lavie and Agarwal (2007) introduced the possibility of evaluating quality based on intuitive "human assimilation": matches on lemmatized forms, and synonymy seek to address deficiencies of simpler word-based metrics.

In our context, post-editors are professional translators with very strict guidelines to perform "light" PE (which is possible on technical documentation for already highly customized translation). This creates natural "human targeted reference" and is therefore naturally suited for HTER evaluation, however in our approach, translation edit rate based on "mechanical edits" count is just an intermediate analysis to expose "logical edits" taking into account part of speech, lemmatization, and constituent structure of the sentences.

1.3 Can we reduce the effort?

Beyond analysis, the general problem is how PE effort can be reduced. Multiple approaches can be quoted for that purpose: Guzmán (2007) describes a set-up where MT output passes through a set of PE rules designed to smooth out translation output for a highly customized system.

Dugast et al. (2007) and Simard et al. (2007) describe a set-up where an SMT system is trained on a bilingual corpus constituted with both MT output and human reference, and show how the system learn how to "correct the translation output". Schwenk et al. (2009) reproduce this with a Statistical Post-Editing (SPE) system trained on very large corpus making the initial translation as a mere preprocessing. In both cases, the SMT system benefits from higher similarity between pretranslated text and reference compared to source and reference; however, if the final quality is higher, the system does not learn post-editing.

Through the introduction of PEA, our study shows that a large part of the PE effort can be classified and automatically learn. The rest of this paper is organized as follows: in section 2, we introduce and define the notion of PEA. We also present the experimental data and the PE typology we used. In section 3, we describe the process used to automatically analyze our experimental data. Then we present and discuss our results in section 4.

2 Our Approach

We argue that PE activity can be modeled by a set of rules, resulting of the decomposition and qualitative analysis of the PE results. Our approach consists in automatically extracting a set of *minimal* and logical edits, called "Post-Editing Actions" (PEA). These logical edits are opposed to mechanical edits: typically insertion, deletion, substitution and move used by edit distance.

A PEA is *minimal* in sense that we cannot find a smaller independent edit. A PEA is said "logical" meaning that the transformation it describes linguistically makes sense. For instance, the French sentence "le bord est affiché", post-edited by "la bordure est affichée" can be seen as a mechanical edit operation of 3 words substituted by 3 words, or as a single logical edit of the word "bord" by "bordure" (both being valid meanings of the word "border"). This word edit comes with a propagation of the gender of the noun headword to its modifiers: here, the determiner and the adjective are in a predicate position of the subject. *In that PE, the intent of the post-editor was to correct only one single word, and the introduction of PEA is to reflect that intent.*

This approach is possible to the extent that modifications during post-editing make sense, *i.e.* the number of edits is limited and it is possible to identify minimal logical changes. Some edits won't fit into that category, either because the source text was not making any sense to analyze (what we will call "word salad"), or the post-editor introduced a mistake, or he decided to radically change the structure of the sentence making the decomposition into PEAs impossible. Our preliminary human analysis (set as our reference) shows that this condition is reasonable for PE of customized translation for technical documentation material.

Note that key differences between logical edits compared with mechanical edits are: 1. more intuitive for the post-editor; 2. more difficult to extract in an automatic way; 3. generally multiple word changes will be involved in a single PEA which may group several classical edit operations (insertion, deletion, substitution or shift).

2.1 PEA Typology

For our purpose of defining minimal logical edits relevant to PE, we considered existing classifications presented in (Font-Llitjós et al., 2005; Vilar et al., 2006; Dugast et al., 2007).

Based on these error classifications, we defined the following PEA typology (with examples where SRC is the source sentence, TGT is its MT, and PE is the human post-editing):

Noun-Phrase (NP) — related to lexical changes.

- Determiner choice — change in determiner
  SRC: *enable a drawing preview of the DWG overlay*
  TGT: *activer l'aperçu du dessin de la superposition DWG*
  PE : *activer un aperçu du dessin de la superposition DWG*

- Noun meaning choice — a noun is replaced by another noun **changing its meaning**
  SRC: *the border displays as stripes*
  TGT: *la bordure s'affiche sous forme de rayures.*
  PE : *la bordure s'affiche sous forme de bandes.*

- Noun stylistic change — a noun is replaced by a synonym (no meaning change)
  SRC: *[...]that placing[...]*
  TGT: *[...]que le placement[...]*
  PE : *[...]que le positionnement[...]*

- Noun number change
  SRC: *[...]their proper locations*
  TGT: *[...]leur_ emplacement_ approprié_*
  PE : *[...]leurs emplacements appropriés*

- Case change

- Adjective choice — change in adjective choice for better fit with modified noun
  SRC: *[...]regardless of the active project.*
  TGT: *[...]quel que soit le projet en cours.*
  PE : *[...]quel que soit le projet actif.*

- Multi-word change — multiword expression change (meaning change)
  SRC: *credit card*
  TGT: *carte bancaire*
  PE : *carte de crédit*

- NP structure change — structure change of NP but the sense is preserved
  SRC: *preview color*
  TGT: *couleur de l'aperçu*
  PE : *couleur d'aperçu*

Verbal-Phrase (VP) — related to grammatical changes

- Verb agreement — correction of agreement in verb
  SRC: *connectors can be used*
  TGT: *les connecteurs peut être utilisé_[...]*
  PE : *les connecteurs peuvent être utilisés[...]*

- Verb phrase structure change
  SRC: *the options are displayed[...]*
  TGT: *les options s'affichent[...]*
  PE : *les options sont affichées[...]*

- Verb meaning choice — a verb is replaced by another verb **changing its meaning**
  SRC: *the actual distance decreases[...]*
  TGT: *la distance réelle réduit[...]*
  PE : *la distance réelle diminue[...]*

- Verb stylistic change — a verb is replaced by a synonym
  SRC: *activate this check box*
  TGT: *activez cette case*
  PE : *sélectionnez cette case*

Preposition change
SRC: *snapping to sketches*
TGT: *accrochage des esquisses*
PE : *accrochage aux esquisses*

Co-reference change — generally through introduction/removal of a pronoun, or change of a definite to possessive determiner
SRC: *the distance increases*
TGT: *la distance augmente*
PE : *elle augmente*

Reordering — repositioning of a constituent at a better location (adjective, adverb)
SRC: *then, click*
TGT: *ensuite, cliquez*
PE : *cliquez ensuite*

PE Error — Post-editor made a mistake in his review
SRC: *the dialog box lets you choose to display this option*
TGT: *la boîte de dialogue vous permet de choisir d'afficher cette option*
PE : *la boîte de dialogue __ permet de choisir d'afficher cette option*

Misc style — unnecessary stylistic change
SRC: *can be toggled on or off as required*
TGT: *[...]peut être activée ou désactivée si nécessaire*
PE : *[...]peut être activée ou désactivée selon vos besoins*

Misc — all PEAs that we cannot classify

3 Automation of PEA Analysis

In order to automate our process, we developed a framework to achieve an accurate analysis of post-edited data. The global overview of this framework is represented in figure 1. Our system works as a classifier based on linguistic rules which takes in input a set of sentence pairs made of MT outputs and its post-edited version, and produces PE report based on PEA (following the typology defined above). Each sentence pair is analyzed through a "logical" point of view by applying a three steps protocol. Results are finally aggregated at the document level in particular to produce a PE report and, in a next step of our work, to deduce new translation rules.

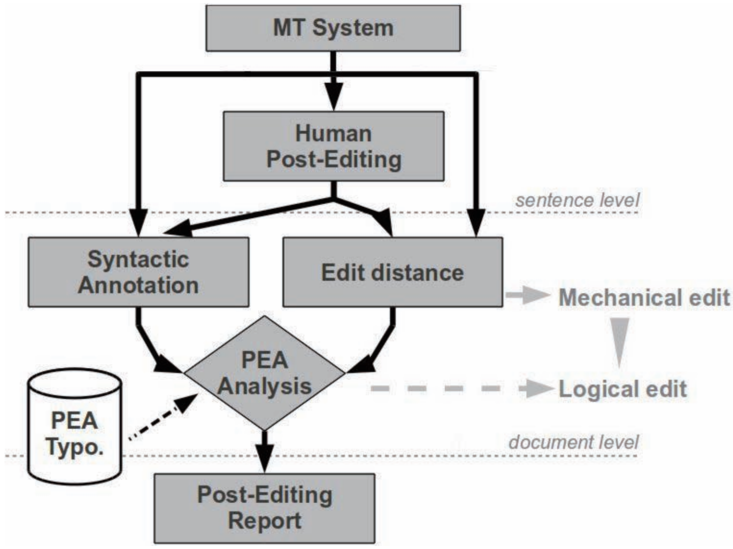

Figure 1: Overview of our PE analysis system.

3.1 Protocol

First, linguistic annotations (part-of-speech (POS), lemma and constituent structure tags such as NP or VP) are generated for initial translation and its post-edited version. This annotation is performed by SYSTRAN syntactic analyzer.

Then, both sentences are aligned in order to identify all changes made during the PE process, and the edit distance is computed. The standard edit operations are considered (insertion, deletion, substitution and move), and a new operation, called "near", corresponding to a substitution of a word by a cognate is introduced. This new operation is useful to localize potential morphology differences on the determiner, noun, verb or adjective. We use for calculating this edit distance an adapted version of TER (Snover et al., 2006).

Finally, PEAs are identified through pattern matching with predefined rules (see next section) by using the first level edits (insertion, deletion, substitution and shifts) and the linguistic annotations. A PEA is produced and add as an annotation in the document if a match is found.

```xml
<annotations>
  &lt;segment id="1"&gt;
    <source>
      Create factory layouts using the default system units.
    </source>
    <target>
      <pea type="verbStyle" id="1">Créez</pea>
      des présentations
      <edit type="agreement">de</edit>
      <pea type="nounMeaningChoice" id="2">famille</pea>
      avec
      <pea type="determinerChoice" id="3">des</pea>
      unités système par défaut.
    </target>
    <pstedt>
      <pea type="verbStyle" id="1">Créer</pea>
      des présentations
      <edit type="agreement">d'</edit>
      <pea type="nounMeaningChoice" id="2">usine</pea>
      avec
      <pea type="determinerChoice" id="3">les</pea>
      unités système par défaut.
    </pstedt>
  &lt;/segment&gt;
</annotations>
```

Figure 2: Sample of PE annotations using XML. The PEA are represented by <pea> nodes simultaneously on the post-editing source (raw MT output: the <target> node) and the post-editing output (<pstedit> node).

3.2 Linguistic Rules

A set of linguistic rules has been defined for our PEA typology. For each class (except for the "misc" class which contains all unclassified cases), patterns are defined based on the linguistic characteristics of the corresponding PEA.

In the current implementation, all classes of the PEA typology are not yet implemented. We chose to focus on the following PEA classes (which are corresponding to the most frequent and simplest patterns):

- Changes in NP: *determiner choice, noun meaning choice, noun number, case change, adjective choice*
- Changes in VP: *verb agreement, verb meaning choice*
- Preposition change
- Co-reference change

4 Experimental Data

The support of our work is based on data from actual PE workflows and provided by Autodesk and Symantec.

The corpus referenced in this paper is a software technical documentation material from a real PE workflow: translated first from English to French, it was post-edited by four different professional translators, who were French native speakers (Plitt and Masselot, 2010). The post-editors were provided with simple PE guidelines to produce publishable quality at the lowest effort, avoiding changes due to stylistic or personal preferences. The post-editors are presented once sentence at a time, in the same order in which they appear in the original source document, without any further functionality supporting the PE activity (*e.g.* no terminology lookup). Some of the PE tasks used MT outputs generated with a Moses engine, an SMT system trained on in-domain data (Koehn et al., 2007), others with the SYSTRAN system. Note that the post-editors were not informed which MT system was used. Although our aim was not to compare RBMT versus SMT, it was interesting to note that our approach applies equally on both system outputs.

4.1 Human Baseline

A subset of 100 sentences (the baseline) was tagged manually using XML format as shown in figure 2. The aim is to compare our automatic results to this reference analysis.

Table 1 describes the human analysis of 100 sentences: in these sentences each PEA has been classified according to the previous typology. The left part corresponds to the RBMT system outputs and the right part to the SMT system output. Counts and proportions of each considered PEA are provided together with their word coverage.

We can observe that the main part of the PE effort involves NPs (about 90%). This fact is interesting since NPs - and in particular in technical documentation - constitute a relatively easy subset on which specific and simple approaches can be applied.

Also, we compared the numbers for the RBMT system with typology of PE performed within the statistical PE process: in (Dugast et al., 2007), the authors analyze the type of modifications performed by a SPE system. The distribution reported is very similar to our analysis of PEA. This shows that the SPE layer is preparing the work of the human post-editor, but has a limited capacity.

Another interesting outcome of this analysis is to

Class / Sub-class	RBMT system #PEA	RBMT system %PEA	SMT system #PEA	SMT system %PEA
Noun-Phrase (NP)	**74**	**90%**	**125**	**92%**
Determiner choice	1	1.2%	3	2.2%
Noun meaning choice	49	59%	84	62%
Noun number	3	3.6%	0	0%
Case change	19	23%	37	27%
Adjective choice	2	2.4%	1	0.7%
Verbal-Phrase (VP)	**6**	**7.2%**	**4**	**3%**
Verb agreement change	3	3.6%	2	1.5%
Verb meaning choice	3	3.6%	2	1.5%
Preposition change	1	1.2%	0	0%
Co-reference change	2	2.4%	7	5%
TOTAL	**83**	**100%**	**136**	**100%**

Table 1: Human analysis of a post-edited subset (100 sentences). This table shows the number of PEA for the implemented classes. We can see that the principal category concerns NP changes with 90% of the total amount for both systems. Terminological changes are the principal source of PEA with 59% and 62%, respectively.

see the PEA repetitions: if we count how many times each PEA is used (a PEA will be uniquely identified by the modification which is obtained independently from the context) - we can extract the most frequent PEA - see table 2. In that context, a first significant reduction of the PE effort (by 23% to 37%) will come by adding four simple rules/phrases in the MT systems. This shows that the training corpus had obviously some terminology gap - which is somewhat expected for this customized MT for technical documentation: training is always related to a previous version of the documentation but, more important, this also gives an idea of the potential for learning even from the current data.

RBMT System	PE	**#**	**%**
famille	usine	96	20%
sol	atelier	65	13%
plancher	sol	11	2%
archive	actif	9	2%
TOTAL (top 4)		**181**	**37%**
TOTAL (all)		**488**	**100%**
SMT System	PE	**#**	**%**
archive	actif	60	11%
superposition	calque	39	7%
archive	ressource	19	3%
sol	atelier	13	2%
TOTAL (top 4)		**131**	**23%**
TOTAL (all)		**558**	**100%**

Table 2: Top four of the most frequent PEAs. They represent simple noun meaning change and cover respectively 37% and 23% of the total PEAs.

4.2 Results from the Automatic Analysis

The results obtained with the automatic system are presented in Table 3. The difference between human and automatic results can be explained, apart from potential analysis errors, by the fact that the human annotation is performed on one or more tokens at the same time, while our automatic process considers the token, one after the other, following the edit path. As a result, some decisions are taken too early, especially when propagations happen after the current modification (which is the case for *e.g.* determiner where a low precision is observed).

Having a closer look at the "Noun meaning choice" class, we can see that a significant amount of terminological changes are detected. This will be particularly useful to adapt the MT system and thus avoid that some mistakes appear again.

Table 4 shows PEA and propagation coverages on our global corpus. On both RBMT and SMT systems outputs, we currently have a 35% coverage for the current set of PEA patterns and agreement propagation detection. With the current Precision and

Class			RBMT system			SMT system		
Sub-class	#PEA	#Match	%Prec.	%Rec.	#PEA	#Match	%Prec.	%Rec.
Noun-Phrase (NP)	**125**	**48**	—	—	**145**	**95**	—	—
Determiner choice	15	1	7%	100%	16	1	6%	33%
Noun meaning choice	89	35	40%	71%	97	69	71%	82%
Noun number	3	0	0	0	4	0	0	0
Case change	18	12	67%	63%	27	25	93%	68%
Adjective choice	0	0	0	0	1	0	0	0
Verbal-Phrase (VP)	**9**	**2**	—	—	**8**	**2**	—	—
Verb agreement change	1	0	0	0	2	0	0	0
Verb meaning choice	8	2	25%	67%	6	2	33%	100%
Preposition change	34	0	0	0	53	0	0	0
Co-reference change	11	1	9%	50%	7	1	14%	14%

Table 3: Automatic PEA analysis of our post-edited data subset. Column #PEA contains the number of PEAs identified, column #Match is the number of correctly recognized PEA, and the last two columns show the Precision and Recall.

	RBMT		SMT	
	#	%	#	%
# Edit	3231	100%	3947	100%
# PEA	1133	35%	1340	34%
# Agr Prop	169	5,2%	255	6,5%
# Det	40	1,2%	99	2,5%
# Prep	102	3,2%	97	2,5%
# Verb	27	0,8%	59	1,5%

Table 4: PEA and agreement propagation coverage for our both RBMT and SMT system outputs. The first column shows the edit number while the second shows the cover rate.

Recall levels of our automatic analysis seen before, the final potential seems interesting and it can probably be further improved.

5 Conclusion

We have defined the notion of Post-Editing Actions (PEA) as "logical edits" of PE by opposition to a "mechanical edits" on which current metrics are defined like BLEU (n-gram precision), WER (insertion, deletion, substitution) and TER (WER + move). We have introduced a categorization of PEAs observed in real data, and we have manually annotated the post-edited outputs of two different MT systems. This gave us interesting insights in the error patterns of both RBMT and SMT systems.

In a second step, we have proposed a procedure to automatically detect these PEAs starting from a TER. The results obtained so far show a good potential for automatically retrieving actual PEAs. We are currently working on several improvements, in particular the refinement of the patterns used for detecting the PEAs and the ability to deal with multiword modifications.

5.1 The human factor

Our approach applies to workflows where the initial MT quality is high and where post-editors are requested to perform light editing. This situation does not apply for general purpose translation tasks. As an example, Martinez (2003) gives the following advices for post-editors for marketing brochures: "*to look for synonyms [in order to] avoid the repetitive style caused by MT consistency, to simulate the performance of a human translator...*". These instructions would definitely harm our automatic extraction process since they would reduce learnability.

Note that even in this context of "light editing" on high quality MT, informal feedback from post-editors show that learning from their comments is a key element to keep them motivated.

5.2 Perspectives

Our next goal is to use the PEA analysis to increase translation quality by taking into account recurrent PEAs. This will reduce time (*i.e.* cost) but also repetitiveness of this task, which are significant points in a business context.

The incremental modification of a RBMT system will be achieved by including an additional dynamic dictionary for the new terminology and single rules. SMT systems on the other hand are more difficult to adapt on the fly since their models are usually trained on large amounts of data implying a time-consuming for retraining process. A possible approach could be based on suffix arrays and incremental EM as introduced by Levenberg et al. (2010).

Acknowledgments

Research partially funded by both the DGA and the ANRT under CIFRE-Defense 7/2009. This work is also supported by the ANR project COSMAT under ANR-09-CORD-004.

References

Robert Carl Barrett, Paul Philip Maglio, and George Milton Underwood. 2001. User-centered push methods and system, May 23. US Patent App. 20,020/054,057.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 workshop on statistical machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28. Citeseer.

Donald A. DePalma and Vijayalaxmi Hegde. 2010. The market for MT post-editing. page 4, November.

Stephen Doherty, Sharon O'Brien, and Michael Carl. 2010. Eye tracking as an MT evaluation technique. *Machine translation*, pages 1–13.

Loïc Dugast, Jean Senellart, and Philipp Koehn. 2007. Statistical post-editing on systran's rule-based translation system. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 220–223. Association for Computational Linguistics.

Ariadna Font-Llitjós, Jaime G. Carbonell, and Alon Lavie. 2005. A framework for interactive and automatic refinement of transfer-based machine translation. In *European Association of Machine Translation (EAMT) 10th Annual Conference. Budapest, Hungary*. Citeseer.

Rafael Guzmán. 2007. Automating MT post-editing using regular expressions.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231.

Abby Levenberg, Chris Callison-Burch, and Miles Osborne. 2010. Stream-based translation models for statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 394–402. Association for Computational Linguistics.

Lorena Guerra Martinez. 2003. *Human Translation versus Machine Translation and Full Post-Editing of Raw Machine Translation Output*. Ph.D. thesis, Citeseer.

Mirko Plitt and Francois Masselot. 2010. A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague Bulletin of Mathematical Linguistics*, 93(-1):7–16.

Johann Roturier. 2009. Deploying novel MT technology to raise the bar for quality: A review of key advantages and challenges. *The twelfth Machine Translation Summit, Ottawa, Canada, August. International Association for Machine Translation*.

Holger Schwenk, Sadaf Abdul-Rauf, Loïc Barrault, and Jean Senellart. 2009. SMT and SPE machine translation systems for wmt'09. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 130–134. Citeseer.

Michel Simard, Nicola Ueffing, Pierre Isabelle, and Roland Kuhn. 2007. Rule-based translation with statistical phrase-based post-editing. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 203–206. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Lucia Specia. 2011. Exploiting objective annotations for measuring translation post-editing effort. In *15th Annual Conference of the European Association for Machine Translation, EAMT*, volume 11.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error analysis of statistical machine translation output. In *Proc. of the Fifth Int. Conf. on Language Resources and Evaluation (LREC)*. Citeseer.