# Design and Analysis of a Large Corpus of Post-Edited Translations: Quality Estimation, Failure Analysis and the Variability of Post-Edition

**Guillaume Wisniewski**
Université Paris Sud and LIMSI
Orsay, France
wisniews@limsi.fr

**Anil Kumar Singh**
LIMSI
Orsay, France
anil@limsi.fr

**Natalia Segal**
Reverso–Softissimo
Neuilly, France
nsegal@softissimo.com

**François Yvon**
Université Paris Sud and LIMSI
Orsay, France
yvon@limsi.fr

## Abstract

Machine Translation (MT) is now often used to produce approximate translations that are then corrected by trained professional post-editors. As a result, more and more datasets of post-edited translations are being collected. These datasets are very useful for training, adapting or testing existing MT systems. In this work, we present the design and content of one such corpus of post-edited translations, and consider less studied possible uses of these data, notably the development of an automatic Quality Estimation (QE) system and the detection of frequent errors in automatic translations. Both applications require a careful assessment of the variability in post-editions, that we study here.

## 1 Introduction

Post-editing, the process of editing the outputs of a Machine Translation (MT) system in order to correct the translations in terms of fluency and adequacy, is becoming more and more popular both to produce human-quality translations at a reduced cost (Garcia, 2011) or to evaluate the quality of MT systems. Indeed, the hTER score (Snover et al., 2006), which depends on the number of editions required to transform a MT hypothesis into a correct (post-edited) translation has proved to be a good indicator of the quality of a MT system.

With the development of post-edition, more and more datasets of post-edited translations are being collected and distributed (Potet et al., 2012; Callison-Burch et al., 2012). These corpora have been accumulated in the context of MT evaluation campaigns and have mainly been used to estimate translation quality. They can also serve several other purposes: our first contribution is to show how they can be used to identify and analyze the limits of a MT system and to train a quality estimation (QE) system. For these tasks we present results achieved on the TRACE corpus,[1] a new, large corpus of French to English and English to French post-editions, which has been recently assembled using data collected from a public web portal and from datasets used in MT evaluation campaigns.

The second contribution of this work is a study of the variability of post-edition, a question that the growing role of the TER score, both in MT evaluation and as a measure of the post-edition effort,[2] makes more and more important. Since it has long been recognized that MT evaluation (especially at the sentence level) is plagued with a low inter-rater agreement (Koehn and Monz, 2006), it seems appropriate to raise the same issues in relationship to the QE task. Our analysis relies on a subpart of the TRACE corpus containing automatic translations that have been post-edited independently by two translators. To the best of our knowledge, this is the first time that several post-editions of the same sentences have been collected, allowing us to perform both a qualitative comparison of the differences between the post-editions of two translators as well as a quantitative analysis of the inter-rater agreement for the hTER score.

The rest of the paper is organized as follows. We first describe in Section 2 a large corpus of post-editions that has been collected for this work. We

[1] The corpus is downloadable from anrtrace.limsi.fr

[2] For instance, the quality estimation task organized for the 2013 edition of the ACL Workshop on Machine Translation (WMT) is cast as the problem of predicting the hTER score.

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. 117–124.*

then present several experiments that have been made with this corpus to identify and analyze some limitations of existing MT systems (Section 3) and to develop a QE system (Section 4). We finally present, in Section 5, a study on the variability of post-editions and on the inter-rater agreement of hTER score.

## 2 Corpus Description

The TRACE corpus of post-edited translations contains $6,693$ French sentences ($109,689$ words), accompanied by two automatic translations in English and the post-edition of one of these translations by a professional translator. An analogous corpus contains $5,929$ sentences ($120,378$ words) for the English to French direction. For the two directions, $1,000$ additional sentences that have been post-edited independently by two translators have also been prepared. These corpora can be freely downloaded from the TRACE website.

Half of the source sentences have been collected through a public web portal which serves each month several millions of translation requests between French and English. These requests cover a wide variety of genres and domains. The other half of the corpus is made of parts of the datasets provided by MT evaluation campaigns (WMT[3] (Callison-Burch et al., 2012) and IWSLT (Cettolo et al., 2012)) and by Word Sense Disambiguation campaigns (Lefever and Hoste, 2010). Examples from this part of the corpus are accompanied by additional information provided by the campaigns organizers such as reference translations or semantic annotations.

These sentences have been translated by two MT systems: the first one, denoted by SYSRULE, is a commercial rule-based system; the second, denoted SYSSTAT, a state-of-the-art phrase-based statistical MT system developed for the WMT'12 evaluation campaign (Le et al., 2012).

Precise guidelines were given to the translators to ensure that the corrections of the automatic translations were *minimal*: they were asked to produce correct translations (with respect to both adequacy and fluency), while remaining as close as possible to the original translations. To guarantee the quality of the post-editions, samples of the post-editions were further reviewed and corrected when appropriate. As a sanity check, post-edited translations were then used to compute standard MT metrics on the automatic output. As reflected in Table 1 for the English to French direction, the metric values are much higher than what is usually observed in MT evaluation campaigns. This shows that the post-edited references are indeed much closer to the translations than the references used in these campaigns. For instance, when SYSSTAT is evaluated against the references of the WMT campaign, its TER score is $56.27$, nearly twice as worse as when evaluated using post-edited translations as reference. It should also be noted that, as mentioned in many past studies (Callison-Burch et al., 2006), rule-based systems are highly disfavored by automatic metrics.

|  | SYSSTAT | SYSRULE |
|---|---|---|
| BLEU↑ | $56.98$ | $47.62$ |
| hTER↓ | $29.08$ | $36.83$ |
| Meteor↑ | $40.64$ | $33.76$ |

Table 1: Automatic evaluation for the English to French direction on the TRACE corpus using post-edited hypotheses as references. The higher (resp. the lower) scores followed by a ↑ (resp. ↓) are, the better the system performance.

## 3 Failure Analysis of MT systems

We show, in this section, how comparing translation hypotheses with their post-editions can help identify and analyze failures of MT systems. For space reasons, only results for the English to French direction are presented.

### 3.1 Error Patterns

By computing the edit distance at the word-level between translation hypotheses and their post-edition, it is possible to automatically detect the modifications required to make MT output both fluent and adequate. The careful analysis of the most frequent corrections is then likely to give some hints about failures of existing MT systems. All the edit distances used in this section have been computed using TERCom (Snover et al., 2006), an implementation of the Levenstein distance that considers word and block movements; with such extended set of operations, distance computation

[3] Only sentences in their original language have been selected: French source sentences are 'pure' French, and not translations from some other language into French.

has to resort to an approximate search algorithm.

The results reported in Table 2 show that most of edit operations correspond to substitutions. A significant proportion of these substitutions (almost 9%) are, in fact, a modification of the word ending and can be attributed to a morphological error, such as the choice of the wrong gender or number, or of a wrong tense/mood: two typical errors are , "penserai" that is changed to "penserais" (from future tense to conditional mood) and "spéciales" to "spécial" (from feminine plural to masculine singular). This observation is quite surprising as it could be expected that, at least for SYSSTAT, the language model would resolve such difficulties. It is however difficult to distinguish which of these editions have been made to correct an error in the MT output from the ones that result from another correction of the sentence (e.g. when the ending of an adjective is modified because of the substitution of the word it is qualifying). Another striking fact is the very high number of deletions for SYSRULE, which has a clear tendency to produce translations that are too long.

| Edition | SYSRULE | SYSSTAT |
|---|---|---|
| movement | 3 473 | 2 861 |
| substitution | 10 991 | 10 065 |
| deletion | 7 371 | 3 572 |
| insertion | 2 263 | 2 502 |

Table 2: Number of editions required to correct MT output.

Extracting error patterns is difficult as almost 70% of the editions are unique; most of the frequent corrections involve frequent words, typically function words (Table 3). However, once these have been filtered out, it is possible to identify some patterns. For instance, among the $5,929$ translations in the corpus, the (automatic) translation of the English word "order" into the French word "ordre" has been corrected 23 times into "commande" and the translation of "home" into "maison" has been corrected 12 times into "chez...". Both errors suggest a domain mismatch between the expectation of the translation engine and the actual input sentences. Almost 100 of such error patterns have been found, even if all of them are not as easily interpretable.

| Substitution | | Insertion | | Deletion | |
|---|---|---|---|---|---|
| 148 | les → des | 380 | de | 799 | de |
| 93 | des → les | 233 | la | 335 | à |
| 60 | la → le | 204 | le | 329 | la |
| 57 | du → le | 204 | a | 278 | le |
| 55 | des → de | 184 | à | 277 | que |
| 53 | du → de | 141 | dans | 256 | les |
| 51 | de → des | 131 | que | 242 | en |
| 46 | de → pour | 99 | en | 215 | et |
| 43 | cela → il | 97 | un | 212 | des |
| 42 | une → un | 96 | des | 167 | pour |

Table 3: Most frequent editions.

### 3.2 Differences between Automatic Translations and their Post-Edition

To characterize the differences between automatic translations and their post-edition, we propose to learn a classifier that could distinguish between these two kinds of translations. We hope that finding which features are relevant for making this distinction will provide us some insight about the limits of MT systems. This approach is directly inspired by earlier work in QE like (Kulesza and Shieber, 2004), where the authors try to learn the difference between a good and a bad translation.

In the experiments described in this section, each translation is represented by 336 numerical features, most of which are inspired by works in QE for MT (Callison-Burch et al., 2012).[4] These features can be classified into four categories:

- **Association Features**: Measures of the quality of the 'association' between the source and the target sentences like, for instance, features derived from the IBM 1 model scores;
- **Fluency Features**: Measures of the 'fluency' or the 'grammaticality' of the target and source sentences such as features based on language model scores;
- **Surface Features** extracted mainly from the source sentence such as the number of words, the number of out-of-vocabulary words or words that are not aligned;
- **Syntactic Features**: some simple syntactic features like the number of nouns, modifiers, verbs, function words, WH-words, etc.

[4] Features are distributed with the corpus and are described in (Wisniewski et al., 2013).

In our experiments, we used a random forest classifier (Breiman, 2001). Random forest is an ensemble method that learns many classification trees and predicts an aggregation of their results. Random forests have proven to be very good 'out-of-the-box' learners and have achieved state-of-the-art performance in many tasks. They also provide a quantification of the *importance* of a feature with respect to the predictability of the target variable. This importance is derived from the position of a feature in a decision tree: features used in the top nodes of the trees, which contribute to the final prediction decision of a larger fraction of the input samples, play a more important role than features used near the leaf of the tree.

In our experiments, we use the implementation provided by scikit-learn (Pedregosa et al., 2011). Parameters of the forest are estimated on 2/3 of the data; the last third is used for evaluating prediction performance. Hyper-parameters of the random forest (the number of trees and the stopping criterion) were chosen by 10-fold cross-validation.

The task consisting in distinguishing automatic translations from post-edited translations is intuitively hard because it requires to automatically characterize good translations and because the two kind of sentences are very similar on many aspects. That may explain the rather poor performance of the classifier: the precision on the train set is 63%, and only 59% on the test set.

The 8 most discriminative characteristics and their importance are displayed in Figure 1. Only a very small number of characteristics is useful and most of them are derived from language model scores. Continuous space language models (Le et al., 2011) (features having SOUL in their name) are playing a key role: the importance of most relevant feature, a raw sentence probability estimated by a continuous space language model, is four time larger than the importance of the second most relevant feature. Other features derived from LM scores are POSLMLOGPROB that stands for the log-probability of POS sequence and BIGRAMSFREQQUARTILE1 that describes the percentage of bigrams in the first frequency quartile.

These features have a lower value in the MT translations than in the post-edited translations, which indicates that either the search space of MT systems is not rich enough to contain these fluent hypotheses or that the weight of the LM in the scoring function used by MT systems to evaluate the quality of a translation hypothesis is not large enough. Additional experiments are required to decide which of these hypotheses is correct.

### 3.3 Difference between Post-Edited References and 'Free' References

We carried another experiment taking advantage of the fact that the TRACE corpus contains, for many sentences, both a post-edited reference and a reference that has been used in MT evaluation campaign either for training or testing. These 'free' references are produced without any constraints.

Using the experimental conditions presented in previous section, we have first tried to discriminate the references resulting from post-editing the MT output from the 'free' references. The performance on this task is somewhat better than the one achieved on the previous task: precision on the train set is 71% and 67% on the test set. As previously, several language model features are among the most important features (Figure 1), even if, for this task, several simple surface features are also relevant: SENLENGTH describes the sentence length, NUMPUNC the number of punctuation signs in the sentence and AVGTOKENLENGTH the average length of a token. The most relevant feature, T2SAVGNUMTRANS02, is derived from the alignment probability between the target and the source sentence estimated by IBM 1 model and quantifies the average number of words in the sentence for which the alignment probability is large[5]: post-edited translations, the vocabulary of which is close to the MT output, tend to use more the most frequent translations.

Another interesting comparison between post-edited and 'free' references is the difference in the way they re-order source sentence words. This comparison is motivated by our assumption that the reordering between a source and a reference results mainly from 'stylistic' reasons and does not change the meaning. If this hypothesis is true, post-edited references should exhibit less reordering than 'free' references.

Several metrics have been defined in the literature to quantify the difference in the words orders of a sentence and its translation. In the following, we will use the chunk fragmentation metric

[5] More precisely "large" corresponds to an alignment probability higher than 0.02

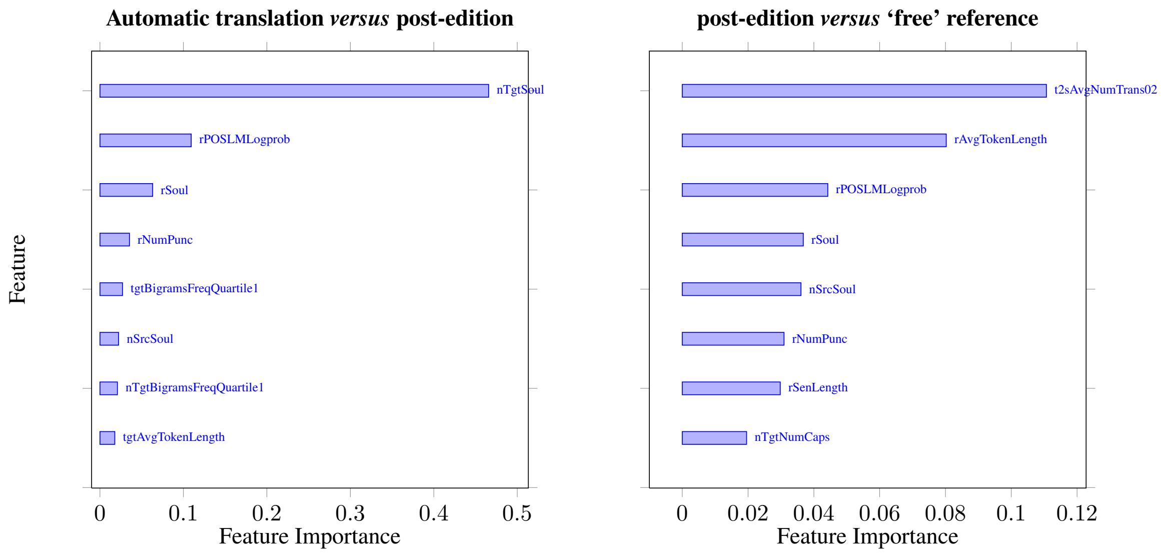

Figure 1: The most relevant features for distinguishing automatic translations from their post-edition (left) and post-edition from 'free' translations (right). Features starting with N are normalized by the sentence length; features starting with a R are made of the ratio between the value of the feature for the source sentence and its value for the target sentence.

defined in (Neubig et al., 2012). Intuitively, this metric quantifies the number of parts the source must be split into to reproduce the order of the target sentence, normalized by the sentence length; the higher this value, the more different the word order. Considering the part of the corpus for which two references are available, we compared their reordering as follows: the two references are aligned with the source using an IBM 4 model trained on the Europarl and NewsCommentary corpora (Callison-Burch et al., 2012); alignments are symmetrized using the GDFA heuristic and, based on these alignments, the chunk fragmentation (as defined in (Neubig et al., 2012), Equation 2) is computed for the two sentences.

For the English to French direction, almost 70% of the sentences the chunk fragmentation is higher in the 'free' reference than in the corresponding post-edited reference. This proportion is similar for the two systems considered in this work. While statistically significant, the difference between the chunk fragmentation in the post-edited and in the free references is quite small: it is, on average, of 0.08, meaning that there is roughly one more discontinuities every ten words. This experiment therefore shows that a large part of the re-ordering observed in today's corpora is not semantically motivated and should not be modeled.

## 4 Quality Estimation

Quality Estimation (QE) is the task of predicting the quality of a automatically computed output without knowledge of the true, expected, output. It is an important step in many Natural Language Processing applications and has recently gained interest in MT. Even if qualitative judgments about translation quality were not collected as, for instance, in the corpus of (Specia et al., 2010), the Trace corpus can still be used to develop and test a QE system, as shown in this section.

In a first experiment, we simply learned a regressor to predict the hTER score that can be interpreted as an (quantitative) estimation of the post-editing effort: it varies between 0 and 1 and quantifies the number of editions (normalized by the sentence length) required to transform the translation hypothesis into an acceptable output.

We also conducted a second experiment in which QE is cast as a classification task. The translations of the Trace corpus were divided arbitrarily into 3 classes according to their hTER scores: the first class contains all examples, the hTER of which is higher than 0.7 and corresponds to translations of 'poor' quality that require a significant editing effort; the second class corresponds to 'good' translations with a hTER score smaller than 0.3 that typically only require few editions; a third class gathers all other translations. The purpose of defining such coarse categories is to help translators focus on the most promising translations, a strategy which has been shown to significantly reduce post-edition time as well as the translator effort (Garcia, 2011). While this split into three categories is certainly questionable, we think that the corresponding classification results are easier to interpret than the MAE[6] score of the regression setting.

We used random forest in all our experiments and the features described in Section 3.2. Performance is estimated by computing the 95% confidence interval of either the MAE metric (for the regression setting) or the 0/1 score (for the classification setting) using bootstrap resampling: 20 splits of the corpus into a training set (80% of the data) and a test set (20%) were generated and a classifier was trained and tested on each of this split. Performance is then averaged over the different runs. Hyper-parameters of the random forest (the number of trees and the stopping criterion) were chosen by 3-fold cross-validation.

In the regression setting, the MAE of the hTER is $0.148 \pm 0.001$ when the translations of the two systems are mixed.[7] This value is however hard to link to the quality of a translation hypothesis as it is an average of an average (the MAE is averaged over the test set and the TER is normalized by the sentence length). That is why we prefer to consider the classification setting introduced above. Table 4 presents results obtained in this setting. Results are quite good: for most examples, the class is correctly predicted, showing that it is possible to automatically identify the translations of high quality. As expected, performance drops significantly when the translations of the two systems are mixed (training and test sets are more heterogeneous) and is slightly better for SYSRULE, the outputs of which are somewhat more 'regular'.

## 5 Inter-rater Agreement in Post-Editions

Another possible use of the TRACE corpus is to study the inter-rater agreement of post-edition: for each translation direction, $1,000$ automatic translations have been corrected twice, which allows us to compare post-editions and hTER scores. To the best of our knowledge, such studies were never reported in the scientific literature.

The similarity between the two post-editions can be estimated by the correlation between the hTER scores obtained when successively evaluating translations with respect to one of the post-edited references. This correlation is low: for the French to English direction, the Pearson coefficient between the hTER scores is $0.576$ and the Kendall $\tau$ is $0.447$, meaning that if translations were ranked according to their hTER scores, the translators would only rank one out of two pairs of arbitrary translations in the same order. More globally, only $12\%$ of the post-editions are the same.[8] The hTER between the two post-editions is $25.8\%$. Even if they are not directly comparable as one of them was not computed with respect to an 'adapted reference', this score is hardly smaller than the TER score obtained when evaluating SYSSTAT (see Section 2), which shows the limit of using TER as an evaluation metric. The most frequent editions of this transformations are substitutions ($52\%$ of the editions), followed by suppression and insertion ($18\%$ each); word shift only accounts for $12\%$ of the editions.

More qualitatively, Table 5 presents examples of the most different post-editions and illustrates possible justifications of these differences:

- difference in the sensitivity to literal translation : in many cases, one of the translator finds (part of) a MT translation to be comprehensible and grammatically correct even if no native speaker would ever 'produce' it and does not change it, while the other translator prefers to reformulate it (e.g. 5th example);
- unnecessary reformulation (without any obvious reasons) of the MT output, as in the 7th example in which "cette réglementation" is corrected in "le présent règlement" or in the 3rd example in which "ingrates" is replaced by one of its synonyms "ungrateful";
- ambiguity resulting from the lack of context, such as in the 1st example.

It must be noted that the quality of the initial automatic translation does not seem to have any impact on the consistency of the post-editions: post-editions are as different when the initial MT hy-

[6] Mean Average Error is a standard evaluation metric for regression defined as the average of the absolute errors $|f - y|$, where $f$ is the prediction and $y$ the true value.

[7] All results are for the English to French direction.

[8] Case and punctuation are not considered in comparison.

| System | Precision | Recall | $F_1$ score |
|---|---|---|---|
| SysStat | 86.11% ±6.10% | 83.36% ±7.84% | 81.01% ±10.02% |
| SysRule | 87.24% ±1.64% | 85.22% ±2.23% | 83.80% ±3.30% |
| both | 81.41% ±3.67% | 77.92% ±5.09% | 74.58% ±5.46% |

Table 4: Results achieved in the QE task using the classification setting; all reported scores correspond to 95% confidence interval estimated over 20 splits of the data in a train and test set.

pothesis is almost correct (4th and 7th example) as when it is completely wrong (6th example).

All these observations show the limit of the evaluation of MT by (h)TER: as post-editing is, at least, as subjective as translating, hTER scores are going to be as noisy as any other metric.

## 6 Conclusion

We have presented in this work a large corpus of post-editions and showed the different kind of analyses that it makes possible. These analyses are highly instructive: they show, in particular, the limits of the hTER metric and how error patterns in a MT system can be identified. Our future work aims at going further into these observations and at integrating these information in MT systems in order to improve translation quality.

## Acknowledgements

This work was partly supported by ANR projects Trace (ANR-09-CORD-023) and Transread (ANR-12-CORD-0015).

## References

Breiman, L. 2001. Random forests. *Mach. Learn.*, 45(1):5–32, October.

Callison-Burch, C., M. Osborne, and P. Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proc. of EACL*, pages 249–256, Genoa, Italy.

Callison-Burch, C., P. Koehn, C. Monz, M. Post, R. Soricut, and Lucia S. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proc. of WMT*, pages 10–51, Montréal, June. ACL.

Cettolo, M., C. Girardi, and M. Federico. 2012. WIT$^3$: Web inventory of transcribed and translated talks. In *Proc. of EAMT*, pages 261–268, Trento, Italy, May.

Garcia, I. 2011. Translating by post-editing: is it the way forward? *MT Journal*, 25:217–237.

Koehn, P. and C. Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proc. of WMT*, pages 102–121, New York City, June. ACL.

Kulesza, A. and S. M. Shieber. 2004. A learning approach to improving sentence-level mt evaluation. In *Proc. of TMI*.

Le, H.-S., I. Oparin, A. Allauzen, J.-L. Gauvain, and F. Yvon. 2011. Structured Output Layer Neural Network Language Model. In *Proc. of ICASSP*, pages 5524–5527, Prague.

Le, H.-S., T. Lavergne, A. Allauzen, M. Apidianaki, L. Gong, A. Max, A. Sokolov, G. Wisniewski, and F. Yvon. 2012. Limsi@wmt12. In *Proc. of WMT*, pages 330–337, Montréal, June. ACL.

Lefever, E. and V. Hoste. 2010. Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *Proc. of the 5th International Workshop on Semantic Evaluation*, pages 15–20, Uppsala, Sweden, July. ACL.

Neubig, G., T. Watanabe, and S. Mori. 2012. Inducing a discriminative parser to optimize machine translation reordering. In *Proc. of EMNLP*, pages 843–853, Jeju Island, Korea, July. ACL.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python . *JMLR*, 12:2825–2830.

Potet, M., E. Esperança-Rodier, L. Besacier, and H. Blanchon. 2012. Collection of a large database of french-english smt output corrections. In *Proc. of LREC*, Istanbul, Turkey, May. ELRA.

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proc. of AMTA*, pages 223–231.

Specia, L., N. Cancedda, and M. Dymetman. 2010. A dataset for assessing machine translation evaluation metrics. In *Proc. of LREC*, pages 3375–3378, Valletta, Malta.

Wisniewski, G., A. K. Singh, and F. Yvon. 2013. Quality estimation for machine translation: Some lessons learned. *accepted in Machine Translation*.

| | | |
|---|---|---|
| 1. | source | Elle roule, roule. |
| | automatic translation | She rolls, rolls. |
| | 1st post-edition | It roll**s**, roll**s**. |
| | 2nd post-edition | It roll**ed** and roll**ed**. |
| 2. | source | Mais plusieurs intervenants du milieu réclamaient au contraire une aide substantielle. |
| | automatic translation | But several participants of the environment demanded on the contrary a substantial assistance. |
| | 1st post-edition | But several **stakeholders** demanded, **to the contrary**, substantial assistance. |
| | 2nd post-edition | But several **players in the sector** called **on the contrary**, for substantial assistance. |
| 3. | source | Ingrats, les opérateurs ont pourtant trouvé à redire. |
| | automatic translation | Ungrateful, operators yet found anything objectionable. |
| | 1st post-edition | **Ingrates**, operators yet found anything objectionable. |
| | 2nd post-edition | **Ungrateful**, the operators nevertheless found more to say. |
| 4. | source | Rendez-vous était pris pour gonfler davantage les protestations populaires sur la place du Sol, à Madrid, alors que les décomptes des élections électorales et régionales commençaient. |
| | automatic translation | Rendez-vous was taken to inflate more popular protests about the place of Sol, Madrid, while the tallying of election and regional elections began. |
| | 1st post-edition | **A meeting was made** to inflate more **mass protests** in Sol Square, Madrid, while general and regional election **tallying** began. |
| | 2nd post-edition | **The call went out** to further swell the **popular protests** on the square of Puerta del Sol, Madrid, while **the tallying** of local and regional elections began. |
| 5. | source | Dear Valued Customer, please follow the steps below to have a troubleshooting. |
| | automatic translation | Cher valorisées à la clientèle, veuillez suivre les étapes ci-dessous pour avoir un dépannage. |
| | 1st correction | Cher client estimé, veuillez suivre les étapes ci-dessous **pour avoir un dépannage**. |
| | 2nd correction | Très cher client, veuillez suivre les étapes ci-dessous **pour être dépanné**. |
| 6. | source | I’m thinking this must be an ancient print date, right. |
| | automatic translation | Je retiens ce doit être une date imprimée antique. |
| | 1st correction | Je pense qu’il s’agit une **ancienne édition, c’est évident.** |
| | 2nd correction | Je pense que ça doit être une **ancienne date d’impression, n’est-ce pas.** |
| 7. | source | Each year, the Member States shall send the Commission a report on the evaluation of the execution and effectiveness of this regulation. |
| | automatic translation | Chaque année, les États membres transmettent à la Commission un rapport sur l’évaluation de l’exécution et l’efficacité de cette réglementation. |
| | 1st correction | Chaque année, les États membres **transmettent** à la Commission un rapport **sur l’évaluation de** l’exécution et l’efficacité **de cette réglementation.** |
| | 2nd correction | Chaque année, les États membres **communiquent** à la Commission un rapport **d’évaluation concernant** l’exécution et l’efficacité **du présent règlement.** |

Table 5: Examples of differences between post-editions found both in the French to English and the English to French corpora. Bold characters highlight the more striking differences