# What Can We Really Learn from Post-editing?

**Marcis Pinnis** marcis.pinnis@tilde.com
Tilde, Riga, Latvia
**Rihards Kalnins** rihards.kalnins@tilde.com
Tilde, Riga, Latvia
**Raivis Skadins** raivis.skadins@tilde.com
Tilde, Riga, Latvia
**Inguna Skadina** inguna.skadina@tilde.com
Tilde, Riga, Latvia

**Abstract**

This paper describes the findings of a large post-editing project in the medical domain carried out by Tilde. It analyzes the efficacy of post-editing highly technical texts in a specialized domain, and provides answers to questions important to localization service providers that are considering the introduction of post-editing in their translation workflows. The results show that by carefully analyzing post-editing projects, machine translation providers and language service providers can learn how to boost productivity in localization, save time and optimize resources in the language editing process, as well as leverage post-edits to improve machine translation engines through dynamic learning.

## 1. Introduction

In order to analyze the efficacy of post-editing highly technical texts in a specialized domain, Tilde embarked on a project to analyze a large post-editing effort in the medical domain. During the project, Tilde had the unique opportunity to take detailed logs of each activity performed by post-editors. Tilde then analyzed the post-editing results, allowing us to answer important questions like: (1) How effectively do post-editors really work with machine translation (MT)? (2) Do post-editors expend their efforts usefully on editing MT results? (3) How can MT be improved to meet the needs of localization companies that utilize post-editing to boost translation productivity? (4) How does the MT quality affect post-editing performance?

## 2. MT System

During the course of the project, post-editors used a statistical MT (SMT) system that was based on the phrase-based Moses SMT system (Koehn et al., 2007). The system was trained on the European Medicines Agency (EMEA) parallel corpus from OPUS corpus (Tiedemann, 2009) and latest documents from EMEA website (years 2009-2014) collected by Tilde on the Tilde MT platform (Vasiļjevs et al., 2012). The statistics of the training corpus before and after filtering are given in Table 1. The system's automatic evaluation results are given in Table 2.

| Corpus | Sentences before filtering | Sentences after filtering |
|---|---:|---:|
| Parallel | 378,869 | 325,332 |
| Monolingual | 378,869 | 332,652 |

*Table 1: Statistics of the training corpora used to train the SMT system*

**B. MT usage from free sources**
Free MT sources are numerous; we may quote WordLingo[3] and MyMemory.
The stats here is a whopping 45%. The figure is evenly distributed among all regions. The figure means that nearly half of all translators regularly use a free MT provider.

I will let everyone decide on the figures above, but here are my observations:

- Paid MT versus free MT. 15% translators using paid MT may seem low as of 2016. Note, however, that paid MT is an opt-in (adhering to paid MT is a deliberate act), while free MT is an opt-out (it is active by default, and can be opted out). Many translators run their tool stock. Like most car owners, they rarely open the hood, if ever. Also note that the 15% figure for paid MT usage was under 10% just 18 months ago, which reveals a fast growth rate: it projects into almost 50% by 2020. Free MT, however, remains relatively flat.

- The 15% figure includes an optical illusion, which is typical in statistics, and I will explain it here. Translators upload documents in different formats. Native formats (like DOC, PDF) are markers of an independent translator, dealing direct with clients; while pre-processed formats (XLIFF, TXML, and generally speaking, XML-based formats) indicate that the document was processed ahead of the translator, by a translation agency or corporate translation department. In that case, it is common, almost a rule, that those formats had MT injected at pre-processing time, in which case MT does not appear in the figures above. If high-tech is used ahead of translation, it is likely that artificial translation was used. Well over 75% of pre-processed formats are injected with a mix of Translation Memory and Machine Translation, with MT being more frequent than TM. With pre-processed formats making up nearly one half of the documents today, the real figure of MT use among translators, thus corrected, is above 20%.

## 9. Conclusion

We can safely say that Machine translation is now mainstream among translators. Concerning fully independent translators, the trend is still modest, but really present, and it is growing fast. As the younger generation steps in, and the emergent economies further develop, that trend can only intensify.

Translation used to be a luxury at prohibitive costs. It is now used at all levels in business and institutions. Two curves are predicting the advent of widepsread use and acceptance of MT at all levels of translation: one is the curve drawn by the need to lower costs in mass translation, the other is the slowly but steady rise in MT quality.

[3] WordLingo is not a free MT provider, but it is free for Wordfast users due to a special deal.

www.proz.com. The article has a belated entry for Machine Translation (stuck just behind "Other useful software"), which mentions that "*when coupled with terminology management, and post-editing services, MT can provide an attractive cost/benefit solution*". MT was still, as late as 2009, seen by translators as a minor, last-resort crutch for those who needed speed.

## 7. MT Acceptance Among Translators: Today

Hard statistics about translators' habits are hard to come by. My focus here is on translators defined as *individual* practitioners: freelance translators, and employed translators who have a say in their workflow. The reason statistics are hard to collect is that translators are very scattered. The profession is atomized into indidivual, isolated, practitioners.

To make things more difficult, agencies are shy about revealing their real practices, the technology they use, their prices. Prying reliable information out of translators and agencies is not easy, and will certainly be obsolete in a short few years.

One category of translation tools is the online Computer-Assisted Translation (CAT) tool: a browser-based alternative to the classical, installed CAT tool that translators love to hate. The online CAT tool is on the upswing, especially among two classes of translators: the younger generation, and translators in emergent markets. Whence a precaution about the following figures: the surveyed population is not characteristic of the entire population of translators, as of 2016. But biology and economics being what they are, that young and emergent population will inevitably become mainstream.

## 8. Today in Figures

Statistics are a difficult to handle properly, and can mean just about anything. Stats on the acceptance of MT by translators are difficult to form. We can only formally poll the use of MT among translators; as for acceptance, which is an attitude toward MT, we can only get clues.

The stats below are derived from two formal sources and one informal source. The two formal sources are an online translation tool (Wordfast Anywhere) with a community of 25,000 registered users, and over 3,000 regular users translating for over ten hours every month. In that situation, figures are reliable, as the tool provides detailed stats on the setup, as well as MT consumption, for each connected translator. The other source is derived from an installed tool (Wordfast Classic and Wordfast PRO), and the associated hotline, which registers the nature of hotline calls, and therefore has a good overview on MT usage.
The last source, an informal one, is the speaker's personal experience as a former translator and project manager, a trainer, a CEO in the translation industry, and a CAT evangelist. While not incorporated in the figures, that experience was used to perform sanity checks on the figures, and to offer an interpretation of the figures.

### A. MT usage from paid sources

Paid sources are basically subscription-based MT providers, the ubiquitous ones being Microsoft Translator and Google Translate, but there are others, like iTranslate4.eu. We should note that most paid sources cost literally nothing per month for a typical freelancer's consumption: about the price of a good beer. Still, the need to fill a form and provide credit card details ensures that users are 1. indeed professional translators, and 2. deliberately opt for MT.

The stats here is: 15% of translators use a paid source. The statistics in Wordfast Anywhere use IP numbers to track the approximate location of translators, and it appears that most of those using paid MT are in Europe (45% of the grand total), followed by North America (30%). The rest is evenly distributed around the world.
Stats in installed tools use email addresses, language code, and hotline call records to estimate location, and they concur with the above figures.