Postediting: an integrated part of a translation software program

By Jeff Allen

This article aims at redefining automatic translation within the translation workflow process, and how a new translation software product has been able to revolutionize the field of Machine Translation (MT) and MT Postediting.

The term Postediting (PE) is by far most commonly associated as a task that is related to Machine Translation (MT). In basic terms, the task of the posteditor is to edit, modify and/or correct pre-translated text that has been processed by a machine translation system from a source language into (a) target language(s). The notion of Postediting has been adequately defined as the "term used for the correction of machine translation output by human linguists / editors." (Senez, 1998). PE has also been briefly defined elsewhere as the "correction of a pre-translated text rather than translation from scratch." (Wagner, 1985).

Up until present, all PE production projects to my knowledge have run documents in batch mode through a machine translation system and then provide the raw output texts to human posteditors. These posteditors use existing third party desktop publishing software (e.g., Microsoft Word, Arbortext SGML editor, etc) to conduct the PE tasks in a process that is referred to as "passive postediting" because it requires the posteditor to adapt this new type of translation task to software that was unfortunately not designed for such a task. As a result, some efforts have been made to help improve the passive PE process by trying to outline and draw up a list of PE criteria. The Postediting Special Interest Group (Postediting SIG) of the Association for MT in the Americas (AMTA), a group that was set up in 1998 (Allen, 1999), has been concerned with the absence of publicly available postediting guidelines although MT integration has moving forward in significant ways during the past few years. Given this absence of specific PE guidelines, each MT implementation project has been forced to invent their own methods and strategies for conducting the PE tasks. All existing publicly available guidelines are often very general and basic, such as those proposed by Löffler-Laurian (1996:93-94): 1) criterion of situation and document type, 2) criterion of necessity, and 3) criterion of comprehensibility.

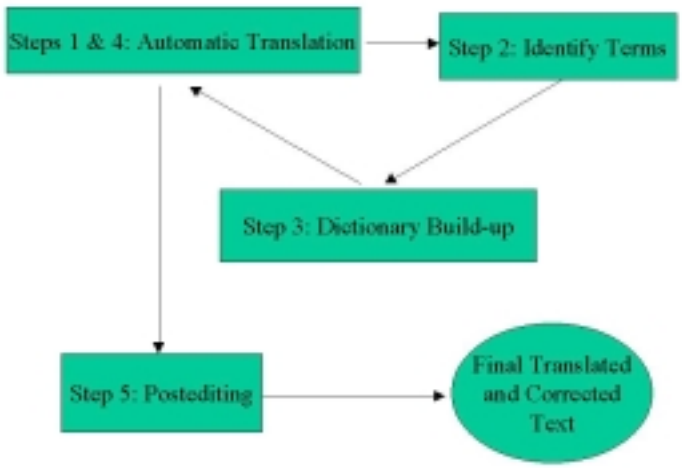
There have been distinctions made between Rapid PE, Minimal PE, and Full PE, yet these terms, and the tasks that accompany them, are often insufficient due to the lack of detailed criteria for the actual PE task. It is not uncommon that posteditors who are asked to conduct Rapid or Minimal PE actually end up making changes to a pre-translated document which are actually closer to the Full Postediting side of the spectrum. The simple reason for this is that everyone wants to produce the best translated document possible, and to retain translation jobs with the same client in the future, including PE jobs.

In general, attempts at implementing PE projects up until recently have focused on imposing a concept of minimal changes upon the translators in the translation workflow process, although without a specific defined set of criteria to follow. Since the raw output translated text has always been postedited within a general word-processor, this significantly slows down the on-screen PE process because all of the manual changes that are counted in terms of mouse clicks, drag-and-drop movements, and replacement keystrokes are not optimally manageable in such word-processors. It has seemed for years that more emphasis has been needed on studying user behavior and the measurement of changes and modifications that are required in the PE process to render MT output texts understandable and usable for different levels of end-user readers of the texts.

It is for this reason that the Paris-based company Softissimo has adopted a unique approach by providing translation software that not only includes an automated translation component, but also with an interface containing several advanced integrated features that significantly speed up the overall MT and PE translation processes. Softissimo's perspective is that Reverso is a complete software tool that allows a professional translator to get the translation job completed on time. It is a software program that provides the means for professional translators and posteditors to get the job done with more accuracy, more consistency, and to produce more translation volume in the same amount of time. From this perspective, Reverso software is a multilingual desktop publishing product that speeds up the translation process. Although there is an automated process of translating, this is only part of the overall process. For high quality translation jobs, it is important to correct the translation output, and the unique point of Reverso is that it provides a conducive environment for translation PE. This fully interactive translation environment is possible within a single translation software program. Rather than forcing translators to adapt their translation tasks to basic MT software that does nothing more than create raw output texts to be passively postedited in another word-processor, Softissimo has designed its Reverso Pro product as a translation software program that includes the PE task as an integral part of the translation process. Reverso Pro therefore has become a translation tool for advanced interactive postediting.

Softissimo has also developed its own set of steps that can easily be followed in using the Reverso Pro translation software. These several editing and postediting tasks are part of the entire translation process in using the translation software.

These steps are indicated in the accompanying flow chart.



Let’s take the example of a 10-page document.

Step 1: Automatic Translation

The initial step of automatic translation of the document by using Reverso Pro provides the translated text at 1-2 pages per second. With Reverso Pro, Step 1 presents the user with an

understandable translation in less than a minute. This is, in fact, a usable product for users who are only interested in content gisting of a foreign language document.

Step 2: Identify unknown and mistranslated terms

This next step consists of identifying the unknown, non-translatable, and mistranslated words and phrases. We have conducted tests on 4 different documents of approximately 10 pages in length. These documents come from completely different fields and of different types of page layout. As can be seen from the statistics provided in the accompanying table, there is an overall average of three occurrences per unknown word in each corpus: otherwise stated, there is an average of three "tokens" for each unknown world "type". These unknown words range from acronyms, to company names, to individual vocabulary words. The most striking fact with the multiple occurrences of unknown words across different types of documents is that once unknown words have been coded into a user dictionary, it is possible to statistically increase the translation accuracy by 300 percent for the unknown words and expressions. See Table 1.

For a document that contains an average of 40 unknown words to be coded, there will also be about 25 additional related expressions containing those words, and 15 other terms and phrases to code into the dictionary. Thus, approximately 80 potential dictionary entries can rapidly be identified. Approximately 30 minutes of time is needed to identify such a number of examples as preparation for the dictionary build-up step.

Table 1

Document number	# of words in source text (English)	# of words in raw MT output (French)	# of different unknown words	total number of unknown words (all occurrences)
Document 1	7,335 words	8,780 words	48	114
Document 2	2,618 words	2,923 words	17	54
Document 3	7,458 words	9,270 words	78	239
Document 4	2,307 words	2,630 words	13	34

Step 3: Dictionary Build-up

After identifying the various problematic terms, it is necessary to create dictionary entries within Reverso Pro for the unknown words, the words to preserve, and the mistranslated words and short expressions. Using the Dictionary Editor of Reverso Pro with an average of 1 term or expression entry per minute, this dictionary build-up stage takes approximately 1.5 hours of time for a 10-page document that contains approximately 40 unknown words + 40 other related expressions and phrases. Also note that the amount of data entry time decreases with experience on the software and with the advanced and improved analysis of the documents.

Table 2 and 3 give extracts from Reverso specialized dictionaries that show the format of user and specialized dictionaries ("n" stands for Noun):

Table 2, English to French Medical Dictionary

Entry headword/term	**word class**	**translation equivalent**
Acid-base balance	n	un équilibre acido-basique
Acute benign lymphoblastosis	n	la mononucléose infectieuse
Calcium content	n	la teneur en calcium
Notification of a disease	n	la déclaration d'une maladie
Outflow tract of right ventricle	n	la chambre artérielle du ventricule droit
Parathyroid insufficiency	n	une hypoparathroïdie
Person-to-person spread of disease	n	la contagion interhumaine
Shinbone	n	le tibia

Table 3, English to French Automotive Dictionary

Entry headword/term	**word class**	**translation equivalent**

4-wheel steering vehicule	n	le véhicule à quatre roues directrices
Auxiliary lift axle	n	un essieu relevable
Brake pad carrier	n	le logement de plaquette de frein
Driver air bag	n	un airbag conducteur
Headlight warning buzzer	n	un avertisseur de non-extension des phares
Multiple fuel injection pump	n	la pompe à injecteurs multiples de carburant
Power seatback recliner	n	le dossier de siège inclinable électriquement
Remote fuel-filler door release	n	la commande intérieure d'ouverture de trappe à carburant
Zero emission vehicle	n	le véhicule sans émissions polluantes

Step 4: Reprocess with Automatic Translation

Once the user dictionary has been augmented with all of the known problematic words and expressions, it is time to send the text through Reverso Pro again. This takes less than one minute of translation processing.

Up through Step 4, only approximately 2 hours of time have been invested into pre-translating the text, identifying and coding entries into the dictionary, and retranslating the text. The translation quality is improved by 300% with regard to terms that posed difficulty with translation quality at Step 1. At this point, an additional 30 minutes of time taken to make a few more changes on terminology and style will improve the translation to a level of quality that is appropriate for internal use within an organization.

By the end of Step 4, there are two resulting products:

- a translated document that is appropriate for internal use that has been completed in 1/3 the time that it would take to translate it via traditional translation methods;

- a user dictionary that is specifically adapted to the user's domain and which can also be used for future translations.

Step 5: Full Postediting

For documents that are intended for wide distribution and not simply for internal use within an organization, it is preferable to utilize the Full PE process to produce a high-quality final translation. At this stage, Reverso Pro presents additional advanced niceties like paragraph alignment, color-coding identification, and the ability to choose an alternative translation with a mouse click as specific features that allow for optimal interactive PE process within the software. With two hours of preparation work, and a few more hours of full postediting, a high-quality translated document can easily be completed. This is significantly less than 10 hours of time (an average of one page per hour) by using traditional translation methods.

Some PE Statistics

And now you may ask how fast the task is for high quality, full PE on marketing brochures. Even though other PE projects in the past have focused on technical documents, Reverso Pro can easily be used to increase your translation productivity on marketing information. Here are some statistics on a set of four marketing brochures that were prepared for a range of products during a period of a couple of weeks. These timed tests were specifically conducted on the PE process (from French to English) for translating all four of these related marketing brochures. A user dictionary was created starting with Brochure 1 and was completed at the end of Brochure 4.

Table 4

Brochure number	**# of words in source text (French)**	**# of words in raw output text (English)**	**PE time**	**# of total PE changes made**	**# of choices made for translation variants**
Brochure 1	540 words	559 words	22 minutes	56 changes	13 choices
Brochure 2	483 words	496 words	20 minutes	38 changes	8 choices
Brochure 3	472 words	486 words	20 minutes	57 changes	27 choices

Brochure 4	222 words	241 words	8 minutes	25 changes	10 choices

By the time these four product brochures were completed for a related range of products, a short French-to-English marketing dictionary specific to the company and its marketing expertise had been created. The user dictionary contained 26 verb/verbal expression entries, 15 adjective entries, and 145 nouns, multi-word nouns, and adverbial expressions. This 200 word dictionary represents an overall of 4 hours of dictionary entry work.

There are a couple of very important points to note with regard to the above-mentioned data. First, the processing time of automatically translating each of the four brochures texts and postediting them is the equivalent to the time necessary for a translator to simply type a text of the same number of words at an average of 25-30 words per minute -- not counting the time to conduct the cognitive process of translating. It is clear that a human translator working from scratch would barely have the time to simply retype the text in the amount of time it was possible to postedit such a text in the experiments above. Unless you can find a translator who can miraculously type at 75 words per minute, while translating at the same rate, it is not possible to arrive at the same level of productivity.

Secondly, in the case where specific terms in the marketing brochures already had a translation entry in the existing Reverso multi-domain dictionary, the translation within this domain was simply added to the dictionary and placed in priority with regard to the default translation. For example, the French word “capture” is given in English as “arrest” in Reverso’s multi-domain dictionary. However, in this specific field, the preferred English translation is “screen shot”. In the tests indicated above, the preferred translations for this domain were added to the dictionary entries and placed in priority above the original default translations. Thus, the general translation still appeared in all cases as an alternative translation. Yet, as we can imagine, it is improbable that “capture” in this specific field would never need to be translated as “arrest”. An additional customization process could be conducted on the 200 word dictionary (requiring approximately 15 minutes of time) to fine-tune this specific user dictionary so that only the preferred translations would appear in future translations. This would also then significantly reduce the number of mouse clicks made for the variant choices, thus reducing the total PE time even more.

These tests conducted in actual production mode for creating marketing brochures demonstrate clearly that using the Reverso Pro translation software product is significantly advantageous within translation production environments. If one can fully postedit texts for high-quality published documents in the same amount of time it takes to simply type the equivalent amount of text into a computer, then why should anyone want to spend the time conducting the translation process in their head?

By-Products

In addition to the translated document, there are several by-products created when following this process. First, the translator is able to produce much more translated material in the same amount of time. Secondly, a customized user dictionary allows translators to not only immediately improve the translation of a given document but also to improve the translation consistency of future documents. Thirdly, the source and final translated texts can be saved as

aligned bilingual texts. This is quite advantageous for combining with translation memory components.

You are invited to test out the basic Reverso translation engine at http://www.reverso.net which demonstrates the type of translation quality that can be obtained in Step 1 indicated able. The additional steps mentioned, and the related features, are part of the Reverso Pro translation software product. See http://www.softissimo.com for more details.

References:

Allen, Jeffrey. 1999. The Postediting Special Interest Group (SIG). In MT News International, No. #21, February 1999.

Allen, Jeffrey. 2001. Post-editing or no post-editing? Regular column article in the International Journal for Language and Documentation, Issue 8, December 2000/January 2001. pp. 41-42.

Allen, Jeffrey. (forthcoming). Post-editing. In Computers and Translation: A Handbook for Translators. Edited by Harold Somers. Amsterdam: John Benjamins.

Löffler-Laurian, Anne-Marie. 1996. La Traduction Automatique. Lille, France: Presses Universitaires du Septentrion.

Senez, Dorothy. 1998. The Machine Translation Help Desk and the Postediting service. In Terminologie & Traduction 1-1998, OPOCE, European Commission, pp. 289-295.

Wagner, Emma. 1985. Post-editing Systran – A challenge for Commission Translators, In Terminologie & Traduction 3-1985, OPOCE, European Commission.