January 15, 2014

# The great leap forward that never was

Uwe Muegge

# The Great Leap Forward that Never Was

*By Uwe Muegge*

**Ten years ago,** machine translation (MT) was very much a niche technology for translation professionals. MT tools were expensive, supported only a small number of language pairs, and did not play nice with many standard translation tools. Today, MT is available either for free or at very low cost, in thousands of language combinations, and many translation memory systems offer integrated MT functionality. In other words, MT has evolved by leaps and bounds.

As the number of translation professionals who post-edit raw MT has increased dramatically over the past 10 years, the big question is whether developments in post-editing have kept pace with those in MT.

## Post-Editing: The Basics

Before we proceed, let's review some of the basics. A general understanding of MT technology is a good place to start.

**Machine Translation (MT):** process of using software applications that automate the translation of text from one language (source language) into another language (target language), with no or minimal human intervention.

**Raw Machine Translation:** output generated by machine translation without human post-editing.

**Post-Editing:** process of revising raw machine translation by human linguists.

**Translation Memory:** software application that enables translation professionals to reuse their previous translations and perform other translation-related tasks efficiently.

**Revision:** bilingual editing of target content based on a comparison between the source content and the target content.

Simply put, during machine translation, an MT application takes the source files as input, and, depending on the type of MT technology (rule-based, statistical, or hybrid), uses specific algorithms to create target files. Depending on the intended use, these raw machine translations can either be used as is or be revised ("post-edited") by human linguists.

The convergence of translation memory and MT technologies has been critically important to language services providers and, as a result, globalization efforts as a whole. Essentially, translation memory systems have at their core a large database of aligned source and target segments (typically sentences) that automatically provides translation suggestions for sentences that have previously been translated or that are similar to previously translated sentences. While most translation professionals consider translation memories as primarily productivity tools, they are in fact first and foremost quality assurance tools. Even if linguists do not get a single match during a translation project, they always benefit from functions such as automatic completeness checks, automatic terminology recognition (if properly prepared), automatic tag/formatting checks, etc.

Post-editing of raw MT is the process where professionally trained human linguists systematically review and edit machine-generated translation content. Depending on the intended use, post-editing may range from only correcting terminology errors to comprehensive rewriting where the final text is indistinguishable from a human-generated translation.

## Where Was Post-Editing 10 Years Ago?

At the turn of the millennium, my professional life gravitated toward MT and post-editing. In 1998, I had gotten my first taste of this high-tech approach to translation while writing a master's thesis that involved post-editing raw MT. Frustrated with the tools for post-editing that existed at that time, I assembled the following suite of tools that enabled me to improve the efficiency of my post-editing efforts dramatically.

**Rule-Based MT Engine:** Before the advent of Google Translate, if freelance translation professionals wanted to use MT, they had to buy a commercial MT system. The only MT systems that freelancers could afford used the earlier rule-based approach to MT. Rule-based MT systems do not simply translate word for word. Instead, rule-based MT systems use a sophisticated repository of grammar rules for both source language analysis and target language generation. Rule-based MT systems also use one or more dictionaries. Rule-based MT uses a three-stage translation process:

1. Analysis: parses the source sentence to create a tree of the syntactic structure of that sentence.

2. Transfer: converts the syntactic tree for the source language into the corresponding tree for the target language.

3. Generation: populates the target tree with corresponding words to create a sentence in the target languages.

Most rule-based MT products were (and still are) targeted at the consumer market. However, a few of these rule-based systems, such as Promt, Systran, and Langenscheidt T1, offered professional-rate features.

One of the features of rule-based MT that I have always liked is the ease with which these systems can be tailored to meet the needs of a specific project. Rule-based MT systems typically allow users to specify language-specific settings regarding register and style that help optimize raw MT output. For instance, users can specify if the polite or the imperative form should be used in a translation.

Most rule-based MT products were (and still are) targeted at the consumer market.

Another great benefit rule-based MT products offer is the fact that they typically come with large domain-specific dictionaries that users can easily expand with their own entries. By taking advantage of the terminology management functionality of a rule-based MT system, users can ensure that even raw MT contains only the client's preferred terminology.

**Automatic Terminology Extraction:** It is a little known fact that many rule-based MT systems offer an unknown word function. The unknown word function creates a list of words in a source text that are not included in any of the dictionaries submitted to the MT system. This function enables users to provide rule-based MT systems with the complete bilingual dictionaries that are required for the best translation results.

**Integrated Translation Memory System:** Even back in the late 20th century, some rule-based MT products featured a built-in translation memory. However, at the time, I chose to use an external translation memory tool, Trados 3. Using my standard translation memory tool gave me easy access to my previous translations. During post-editing, these previous (human) translations were available as fuzzy matches in addition to the raw MT alone that the translation memory of an MT system would provide.

**Custom Word Macros:** One of biggest problems in post-editing, at least in my opinion, is the fact that commercial post-editing tools do not offer much support for syntactical or morphological changes. Consider the following scenario. A word that is not the first word in a sentence needs to be moved to the beginning of a sentence to make that sentence more readable. This editing task typically involves the following steps:

1. Select the word to be moved.

2. Move the selected word from its current position to the beginning of the sentence.

3. Change the case of the first letter of the selected word to upper case.

4. Change the case of the first letter of what was the first word in the sentence to lower case.

Fortunately, there is a rather simple solution for this type of problem: the macro functionality in Microsoft Word. As earlier versions of Trados were themselves sets of Word macros, using custom macros was an obvious choice. By the way, all it takes to create Word macros is to click "Start recording" at the beginning of the process that is to be automated and "Stop recording" at the end. With macros, any complex editing task, such as the nominalization of a verb or changing the case/inflection of a word, can be reduced to pressing a simple hotkey combination.

## Where Is Post-Editing Today?

In 2007, Google launched a free post-editing environment, Google Translator Toolkit. Translator Toolkit is a cloud-based service that

provides translation professionals with a translation memory environment for post-editing raw MT created in Google's statistical MT, Google Translate. Today, Google Translator Toolkit is probably the most popular system designed specifically for post-editing.

**What is great about Google Translator Toolkit?** Google Translator Toolkit became an instant success because this service offers a number of very compelling benefits:

- A free post-editing environment for free statistical raw MT.
- Support for more than 70 languages and more than 5,000 language combinations.
- Many key features such as terminology management, translation memory management, and collaborative translation/sharing of translation memories and dictionaries.
- A simple, very user-friendly system.
- A cloud-based service: no software to install; runs on Windows, MacOS, Linux, iOS, Android, and many other operating systems.

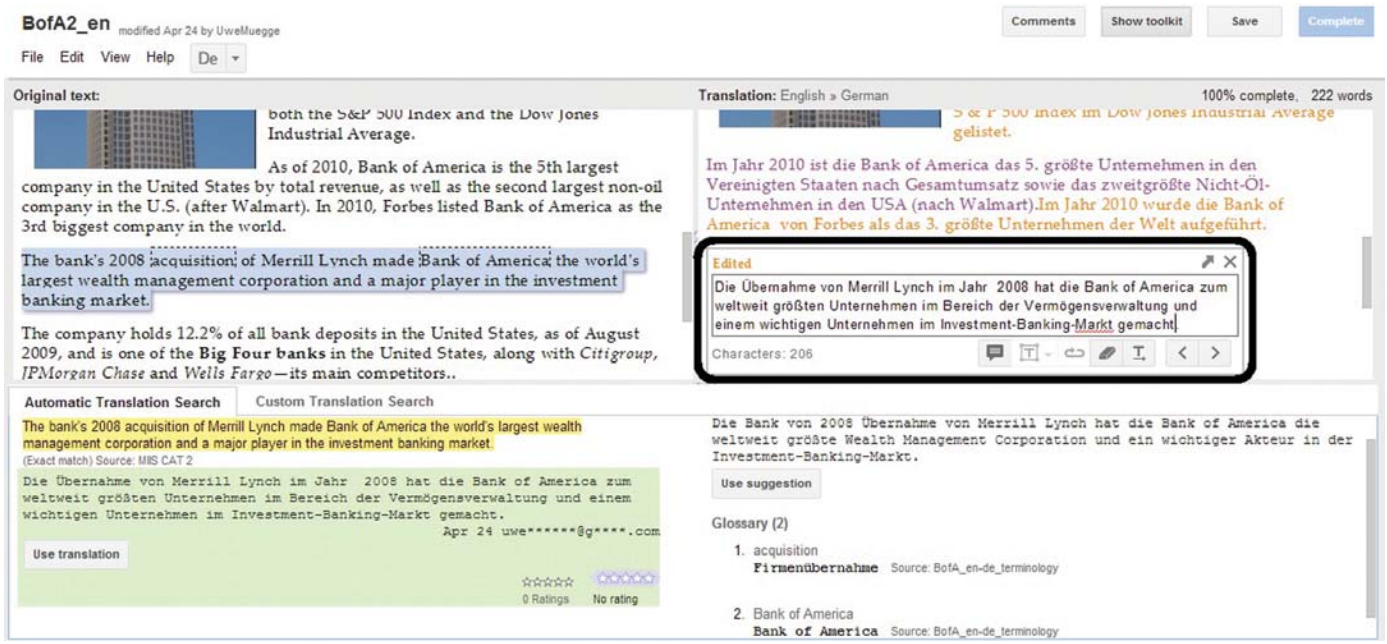
**What is not so great about Google Translator Toolkit?** First, I want to draw attention to the fact that Google did not create Translator Toolkit for altruistic reasons. Translator Toolkit was primarily designed to provide Google with training material for improving the translation quality of Google's statistical MT system, Google Translate. Understanding Google's motivation for creating Translator Toolkit explains why after all these years the feature set of this post-editing/translation memory system is still very rudimentary. (See Figure 1 below for an example.)

Google has received its share of criticism for the lack of privacy in Translator Toolkit. By default, all users of Translator Toolkit share their translation memories not only with the developers of Google Translate, but with all other users of the system as well. While it is possible for users to create "private" translation memories, there is no easy fix for the following issues I have with Translator Toolkit:

- Users have no way of customizing the raw MT Google Translate creates. There is only one specification users can make when submitting text for translation in Translator Toolkit: selecting the language pair.
- While users can upload their own bilingual dictionaries, these dictionaries will not be used for translation, as Translator Toolkit limits their use to the post-editing phase.
- Translator Toolkit does not offer specific functions for changing a) the syntax of a sentence, or b) the morphology of individual words.

By the way, most if not all of these limitations also apply to many translation memory environments that pull raw MT from Google Translate or Bing (Microsoft's statistical MT service).

**Figure 1: The editor (highlighted) in Google Translator Toolkit offers only very rudimentary functions for post-editing MT.**



## Three Things You Can Do to Make Post-Editing More Efficient

In my humble opinion, the most popular environment for post-editing raw MT leaves a lot to be desired in terms of offering task-specific functionality. However, the good news is that there are a number of things that translation professionals can do themselves to improve the efficiency of the post-editing process.

**Manage Terminology:** Using the right terminology consistently is very important in almost any translation project. In post-editing projects, even though the end user may be willing to accept less than brilliant style, incorrectly translated terms are typically not acceptable. Therefore, it is a good idea for post-editors to create comprehensive, project-specific, multilingual glossaries prior to each post-editing project. If the client does not provide comprehensive glossaries, I recommend using one of the many automatic terminology extraction tools and services that help post-editors create multilingual glossaries quickly and inexpensively. And it goes without saying that for the sake of terminology management alone, all post-editing should be performed in a translation memory environment.

**Customize the MT System:** One of the most powerful ways to improve the efficiency of post-editing is, of course, improving the quality of the raw MT. Earlier, I described how translation professionals can customize a rule-based MT system: by selecting built-in domain-specific dictionaries, uploading client glossaries, providing translations for all unknown words, and applying project-specific style settings. But what about statistical MT? Can users customize those? Absolutely! One of the most exciting developments in the area of MT is the advent of do-it-yourself (DIY) MT. In a DIY MT system or service, users build their own statistical MT system using their own translation memories. One example of this new breed of tools is Microsoft Translator Hub. The Microsoft Translator Hub is a free service that anyone can use to a) create customized MT engines, and b) use these MT engines to create high-quality raw MT. Raw MT from DIY MT systems is typically available in multiple file formats, including TMX and XLIFF for easy import into standard post-editing environments. Note that DIY MT services typically require a minimum of 10,000 sentences of parallel text/translation memory for customization.

**Customize the Translation Memory System:** None of the standard commercial post-editing tools available to freelance translation professionals support language-specific post-editing functions. As mentioned above, using the macro-recording function in Microsoft Word is an easy way of simplifying complex editing tasks such as changing the inflection of a word. While almost all translation memory tools now come as stand-alone tools, a few like MetaTexis for Word and Wordfast Classic still use Microsoft Word as an editing platform. Those translation professionals who are looking for the most efficient post-editing platform and are willing to invest a few hours recording macros should give Word-based translation memories a close look.

## Post-Editing Can Be a Much Easier Task

By all indications, more translation professionals than ever are involved in post-editing raw MT. While the technology and the economics of MT have evolved dramatically, making high-quality raw MT available to almost every translation professional, commercial post-editing environments are still relatively primitive. The good news is that there are a number of strategies that translation professionals looking for an improved post-editing experience can use. Through managing terminology, customizing the MT engine, and customizing the (MS Word-based) translation memory system, linguists can improve their post-editing efficiency dramatically. It is certainly true that each of these strategies involves a considerable and, with the exception of customizing the translation memory, ongoing effort. However, for any but the casual post-editor, the benefits of making these improvements in their tools and processes should be immediate.

## Resources

**Google Translator Toolkit**
http://translate.google.com/toolkit

**MetaTexis for Word**
www.metatexis.com/mxword.htm

**Wordfast Classic**
www.wordfast.com/products_wordfast.html

**Microsoft Translator Hub**
http://hub.microsofttranslator.com

## Additional Reading


Muegge, Uwe. "Do-It-Yourself MT: Taking (Statistical) Machine Translation to the Next Level," http://owl.li/rPxY6.

Muegge, Uwe. "Dispelling the Myths of Machine Translation," http://owl.li/rPxTo.

Muegge, Uwe. "Ten Things You Should Know About Automatic Terminology Extraction," *The ATA Chronicle* (September 2012), 24-27, http://owl.li/rPyhy.

Translation Automation User Society Machine Translation Post-editing Guidelines (2010), http://owl.li/rPy7v. ■