



next level
globalization

Tell the world.

Cleaning up and Harmonizing your Translation Memory

**Inconsistent, polluted translation memories? It might be time for a TM spring cleaning.
Here is how you will achieve the best results.**

Michael Oettli
Managing Director, NLG

July, 2017

Cleaning up and Harmonizing your Translation Memory

Translation memories (TMs) continually grow over time with increasing localization activities. At the same time, corporate terminology is often created and managed in a parallel process, not in sync with the TM development. This situation inevitably leads to significant inconsistencies between a company's content assets and can jeopardize quality while increasing cost and process time. So how we can overcome this hurdle?

It's quite common that companies dealing with high volume TM assets consider their translation memories to be polluted with inconsistencies. But even if they know the problem, they often hesitate to take action, fearing the resulting costs.

What are the challenges?

Two main issues can cause inconsistent translation memories:

1. Inconsistencies in target translations

The translation memory contains multiple and different translations for exactly the same source sentence. This often occurs when multiple translation suppliers are used in an uncontrolled and decentralized setting and no effective TM maintenance program is employed. The current trend of mergers and acquisitions in some industries, such as Life Sciences, has led to an increase in companies centralizing translation management and consolidating translation assets, which has also aggravated this problem.

2. Inconsistencies between validated corporate terminology and the terms used in the TMs

This seems to be the most common and most important issue, as the use of the proper terminology is very important for every company and often tightly connected with the company's brand identity. Companies invest time and money in building a solid and validated termbase, but in most cases this is out of sync with their growing TMs. TMs were also often started long before a systematic terminology approach was initiated.

The dilemma begins when new content is created or updated by reusing previously translated and approved content. Should the new deltas follow the approved terminology, leading to inconsistent use of terms between new content and legacy approved content? Or should the approved terminology be ignored to create consistent documents, which results in losing the meaning and power of a validated corporate terminology?

What is the solution?

The problem is obvious and the impact on quality, cost, and timelines can be significant. Many companies desire to clean up translation memories in order to eliminate inconsistent and multiple target entries, and to bring the terminology used in the TMs in sync with the approved company's terminology. But what are the best viable options to accomplish this?

When it comes to terminology mismatches, in many cases companies select a project-based TM cleanup approach. With every new translation project, they reopen pre-translated content, by adding a penalty to the reused context and 100% matches, asking the Language Service Provider to review this content and to harmonize the terms based on the approved corporate terminology. However, depending on the volumes and scope, this can be a never ending and time-consuming process with ongoing and substantial costs.

Imagine having to translate high-volume product documentation based on previous documents reusing 80 percent of the content. This will result in significantly added costs and could even jeopardize the company's product release date due to the additional time needed to review all the content.

Would it not be more appropriate and efficient to conduct a universal cleanup project, which may require a higher one-off investment, but leads to the desired result in a much shorter timeframe?

The suggested approach

At NLG we have successfully followed this approach for enterprise clients who wanted to benefit from clean translation memories in a shorter timeframe, because many of their internal processes were dependent on clean translation assets. What follows is a detailed description of this type of approach.

Before starting, we need to clarify certain details regarding your terminology management and translation process as well as your current TM management and maintenance process.

Ask yourself the following questions:

- How many years have you used translation memories?
- How many languages do you translate into and for how many do you maintain translation memories?
- How the TMs are updated? Do you add new features or overwrite existing ones?
- Do your TMs contain metadata such as status, project name, product type, etc.?
- If you have in-country review, is the reviewed content imported in the TM? Do you use a specific status for this?
- Do your TMs contain display messages, software items, broken segments, etc.? If yes, are these somehow tagged or marked in the source documentation?

- Have you ever migrated content to your TMs that was created in a different translation tool than the one you are currently using?
- Do you maintain separate TMs by company division/department or do you have one master TM?
- Do you manage TM in-house or do you outsource TM maintenance?

This gives a better indication of which actions to take and whether there are specific corporate and content characteristics that must be considered moving forward. Also, to conduct such a project, you should work closely with a language engineer who understands terminology and the techniques of manipulating data in the translation memory databases.

The first goal is to analyze your TM entries at a general level with regard to the use of each translation unit (segment). Very often translation memories grow over time, while content evolves, and technology and language also change. This can result in a large amount of translation units remaining unused for years. A proper TM analysis will illustrate the frequency of use, as well as the dates of the last usage of a translation unit. You can then make an informed decision whether to keep these old units or remove any translation units that haven't been used for a certain amount of time. This helps you to initially reduce the volume of the TM, making further analysis easier.

Correcting TM inconsistencies

As a next step, the TM is analyzed to identify multiple target segments for the same source segment. To do this, you first normalize source and target segments. Normalization means fixing anything that could make identical segments appear inconsistent due to the use of different punctuation, such as quotation marks, dashes (short vs. long), apostrophes etc. or double spaces.

After normalization, harmonization and automated correction processes can take place. This is where the multiple target segments are identified and replaced by one target segment. Typically we **replace** translation units, rather than **remove** them, as these may carry metadata and contextual information important for proper use. Don't just rely on your LSP to replace one translation unit with another. You might know the history of the development of a translation memory better than the LSP. In some cases it might come down to the date of the entry, meaning that multiple target segments get replaced by the most recent entry. In other cases the defining criteria might be in the supplier metadata, meaning that the multiple target segments will be replaced by a segment coming from LSP "A". Or the segment might depend on the site, product line, status, etc.

Source	Target
Inconsistency in Target	
Please refer to local country requirements for these areas.	Consulte los requisitos locales del país para estos ámbitos.
Please refer to local country requirements for these areas.	Consulte los requerimientos locales del país para estos ámbitos.
Inconsistency in Target	
Product availability is subject to regulatory approval in the respective markets.	La disponibilidad de los productos está sujeta a la aprobación normativa en cada mercado.
Product availability is subject to regulatory approval in the respective markets.	La disponibilidad de productos depende de la aprobación conforme a la reglamentación nacional en cada mercado.
Product availability is subject to regulatory approval in the respective markets.	La disponibilidad de los productos depende de la aprobación conforme a la reglamentación nacional en cada mercado.
Inconsistency in Target	
Catalogue number	Número de catálogo
Catalogue number	Número de referencia en el catálogo
Inconsistency in Target	
These can be adjusted in Settings.	Estos mensajes pueden configurarse en Ajustes.
These can be adjusted in Settings.	Estos mensajes son configurables en Ajustes.

Picture 1: Inconsistent target segments

There may be different techniques and tools that fit the purpose of the above-described normalization and harmonization processes. At NLG we have seen the best results achieved by developing and implementing scripts and applying rules with Perl, a programming language originally developed for text manipulation.

Correcting Terminology Inconsistencies

Once the inconsistent translation memory is harmonized, you can move to the next process step, the correction of inconsistent terminology used in the translation memory against the approved corporate terminology.

For this step, you first run a terminology consistency analysis to identify any key term mismatches in the TM. To do this, you can export a reduced TMX file containing only unique segments and analyze using an off-the-shelf quality assurance tool such as QA Distiller, XBenCh, ErrorSpy, Verifika, etc.

These QA tools analyze the translation memory against the validated corporate terminology and generate a mismatch report, which should be reviewed to eliminate false errors.

Key Term Mismatch (Visual status indicators / Optische Statusanzeigen)	
Audible and visual status indicators See figure	Akustische und visuelle Statusanzeigen Siehe Abbildung
Audible and visual status indicators See figure 3	Akustische und visuelle Statusanzeigen Siehe Abbildung 3
Audible and visual status indicators See figure 3	Akustische und visuelle Statusanzeigen Siehe Abbildung 3
Your processor is equipped with audible and visual status indicators (beeps and LEDs).	Ihr Prozessor ist mit akustischen und visuellen Statusanzeigen ausgestattet (Signaltöne und LEDs).

Picture 2: Inconsistent Terminology

To correct any key term mismatches in the translation memory at this point, you engage with the translators. First, you export a new, further reduced TMX containing only translation units with identified mismatches. The translators receive this TMX for correction, together with the mismatch report from the QA tool. This correction can be performed either in a Computer Aided Translation (CAT) tool or using a TMX Editor. To ensure a thorough understanding of the requirements, to set expectations and goals of the process, and to explain the task, it is recommended to arrange a training session with the translators before they start the TMX editing. Once you receive the corrected TMX you should run basic QA checks (spell check and inline tag check) to avoid introducing any errors the translators might have produced during their editing step.

Once the TMX is corrected and spell-checked, it can be merged back into the original master TM. At this point we recommend exporting the master TM from the Translation Management System (TMS) again, and to run the original scripts and rules to normalize, harmonize and correct the translation memory from multiple target segments, especially if there has been any translation activity during the course of the clean-up project. This ensures that any newly introduced TM inconsistencies due to multiple target segments are addressed. Lastly, before importing the merged TMX back into the system, it is advised to run a validation of the master translation memory to avoid any import complications.

Results

After cleaning up the translation memory and aligning its terms with your approved corporate terminology, you will see immediate positive results within your next localization activity. The newly cleaned TM will produce improved leveraging due to "better" matches, thus also reducing costs.

The harmonized terminology is now the basis for consistent and high-quality translations. Your brand identity expressed by the unique use of key terms is now perfectly reflected in your content translated in any language. Translator questions regarding the dilemma whether to follow the termbase or the pre-translations coming from the TM will be eliminated, and the whole translation process will be streamlined. QA checks on the LSP side can be performed smoother and faster with significantly reduced error findings, enabling a considerable decrease in the turnaround time for large volume translations.

Keep it clean

Now, as a last and very important step, it is recommended to implement a Standard Operating Procedure for managing your TM and for dealing with approved terminology, including standard QA procedures for every LSP working on your content. This will ensure that the TM will not be polluted again, and there will be no need for additional TM cleanup projects anytime soon. If you have an in-country review process in place, this procedure must be

extended to the reviewers as well to guarantee that they follow the approved terminology and avoid making unnecessary term changes that might have a negative impact on your translation assets.

If you're working with a TMS – and if the process allows it – it is also advisable to change the setting from “add as new segment” to “overwrite segment”. This ensures that you are not creating TMs with multiple target segments again.

Conclusion

Discussions with different clients highlight the fact that TM and terminology inconsistency is a common issue of which they are perfectly aware. However, many companies are very hesitant to explore their options of cleaning their TMs. Some lack the necessary budget, while others will tell you that the TMs will just be polluted again anyway within six months, so what is the point? The majority of these companies are not working in a controlled translation management environment, but rather in a decentralized manner and with multiple LSPs.

If you are working in one of those global corporations where localization is still seen as a necessary evil, just devouring money, it might be very difficult to procure the additional budget to clean translation assets.

However, for organizations with a certain translation maturity, clean and consistent transition assets are essential. These companies often have a structured terminology management process. They view terminology not only as an asset facilitating higher quality, but also as one that contributes to brand identity, differentiating their “lingo” from the competition in marketing tools as well as in technical documentation. For these companies, the only way to go is to clean up translation memories and to establish an ongoing structured and controlled procedure to avoid the future pollution of translation assets. Building a solid case with data and ROI information to secure the necessary budget should be easy.

About NLG

NLG is a market leading, full-service provider of language related services partnering with global companies in the areas of life sciences, automotive, engineering and power generation. Far from a typical translation agency, we're a process partner for our clients. NLG takes service and technology to “the next level” and our clients benefit from more than 15 years of experience providing best-in-class services around translation and localization, workflow management, process design, consulting, and development & implementation of custom tools and applications. Learn more at NLGworldwide.com