boost
your content
to success

Loosely based on Marie Kondo: The magic of cleaning – tidying up with [itl]-clean

How Festo got much closer to the wealth of data in its translation memories

Success-Story

Where has the treasure been hidden?

The contents of translation memories are without a doubt a hugely valuable asset for international companies. They hold large chunks of texts from various parts of the organization in different languages and conform to corporate terminology specifications. They guarantee efficient and cost-effective translation process. They are the holy grail for companies. They are "big data".

Yet the data in translation memories (TMs) is very often unstructured, past its use-by date and full of data that is kept "just in case", making it look more like an attic stuffed full of things than a treasure chest.

There is one thing that stuffed attics and large TMs have in common: you know they should be tidied up. But every attempt to do this is so off-putting that all you can do is quickly close the door again.

But there comes a point when you can no longer avoid the inevitable.

You may be moving, you may need more space for something else or you actually want to make use of whatever you have stored. You may be changing systems, the space needed for storage is too large and unwieldy or you may want to use the database for training a machine translation process. That's when you no longer have a choice and you have to get down to work.

If it is too big a job to do on your own, then – taking the example of the attic full of stuff – there are people who can help you with removals or consultants to help you organize your "stuff". And for managing the clean-up of content, there is itl.

The world of content creation and languages has been impacted by accessible, affordable and above all efficient solutions that artificial intelligence and machine learning are offering. Yet at itl we have not just limited ourselves to the obvious topic of machine translation; instead, we have made the numerous opportunities opened up by *intelligent solutions* as a new area of expertise also part of our core activities.

The importance of the content of translation memories as big data has not gone unnoticed by specialists. By working together closely with colleagues in *global translation* as well as with our long-standing clients, software developers soon became aware of the pain points of localization.

Festo and the treasure chest in the attic

Festo SE & Co. KG A is company that has its own professional in-house localization team, is always keeping in touch with the latest trends, and also aims at optimizing processes. We count ourselves lucky that Festo has been our client for more than ten years now. Festo challenges and encourages us to suggest, introduce and try out innovative solutions and put them into action by working together as partners.

The latest example of this cooperation can truly be called a success story and the protagonist is [itl]-clean.

Like so many other successful and international organizations, Festo finds itself faced with the challenge of having to clean database memories that are chock-full. Yet people are often fearful that valuable content may get thrown out and have the feeling that this will be a never-ending task that will cost time, money and be stressful.

Some of the data in the translation memory is already more than 20 years' old and has even lived through one CAT tool change. At the time, files created at different periods by different service providers were merged and migrated. Despite best efforts and intentions, this made it increasingly unlikely that the data could be cleaned up in an efficient and useful way.

In addition, terminology was becoming an ever more central part of the localization process and much effort was invested in keeping it up to date. This in turn highlighted the discrepancy with an outdated translation memory database.

In autumn 2021, the Festo localization team at itl, headed up by team leader Stephanie Morjan, was kept up to date during their regular meetings about the progress made in the development of a TM clean-up tool. This was something that could be very useful for Festo in order to get to grips with the problems that both parties were aware of. At the same time, Bastian Hileman, Head of Software Development (aka *intelligent solutions*) at itl, organized a few gatherings to exchange information about this very topic. This was soon followed by the first of a series of meetings with Festo to discuss their expectations.

The requirements from Festo were clear and challenging:

- The clean-up would have to be carried out within an acceptable timeframe so that the data would quickly be available again for use, but also – and more importantly – no large amounts of new data should be accumulated during the clean-up.
- The data that had been removed would have to be accessible so that any fears that content, which had been removed might be thrown out even if it was needed at a later stage, could be overcome.
- The costs of the clean-up would have to be reasonable and efficient in terms of resources. Any manual processing was out of the question.
- The level of quality of the content achieved with the clean-up would need to be maintained as part of a continuous quality assurance process.

[itl]-clean is the right tool for the job

The tool for the automated and long-lasting clean-up of large, multilingual content systems which itl has been working on for some time, is [itl]-clean, a scalable cloud solution.

In a nutshell, it automatically removes those text segments from the database that contain recognisable errors and stores them in a separate database.

We have been able to fine-tune [itl]-clean in cooperation with our experienced contacts in the Translation and Content Services team at Festo and were finally able to use the tool for a large project, to the great satisfaction of everyone involved.

One benefit of such a clean-up is the reduction in the amount of data that need to be managed so that they require less storage space and less effort is needed to migrate and move it.

Another benefit can be found in the optimization of the data quality. Clean and clear databases improve the quality of translations, simplify quality assurance and reduce sources of error, thus also minimizing negative feedback and laborious complaint management.

So what exactly can [itl]-clean do and what was achieved in the project with Festo that got the client to wax lyrical

Just a few technical details

[itl]-clean is based on a set of scalable AI algorithms that objectively target recognisable and measurable characteristics of multilingual content. The way these are selected, configured and combined is the result of a clever mix of expertise in the localization industry that itl has built up over many years, best practices in the area of localization by well-known R&D institutions as well as the competencies and up-to-date knowledge of software development of itl's *intelligent solutions* team.

Here is an overview of the algorithms that were actually used with a few typical examples.

Language	Terminology	Translation instruction	Formal errors	Meta data
Language recognition: translation unit does not contain the expected source or target language	Multilingual terminology checks in line with the term bank or glossary	Difference in segment lengths between source and target	Characters that can be expected are missing (punctuation, missing brackets)	Outdated segments (date it was entered)
	Granular configuration in terms of identification, e.g. identification of forbidden terms, department-specific terms, terms without target language, etc.	Segments with one word or segments that are far too long for training an MT engine	Normalization of empty spaces	Entries by specific departments, authors, etc.
		Empty segments		
		Segments only with symbols or characters		
		Tag-heavy segments		

The working method is as follows:

- The algorithms are defined
  - > All checks can be individually adjusted dependent on the goal (migration or MT training) and language (punctuation is not always the same in the source and the target language)
- Translation units that do not conform are removed and stored in a separate database
- The result is a clean database and a database(s) with the "rejects" (removed segments)

Most European and Asian languages, including a few national language variants, can already be checked now.

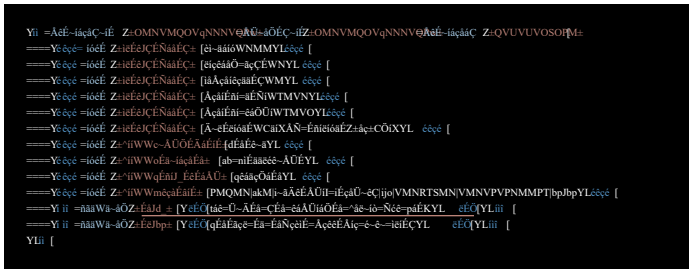


Image: Example of language recognition. The German text has been saved as part of a UK-EN entry

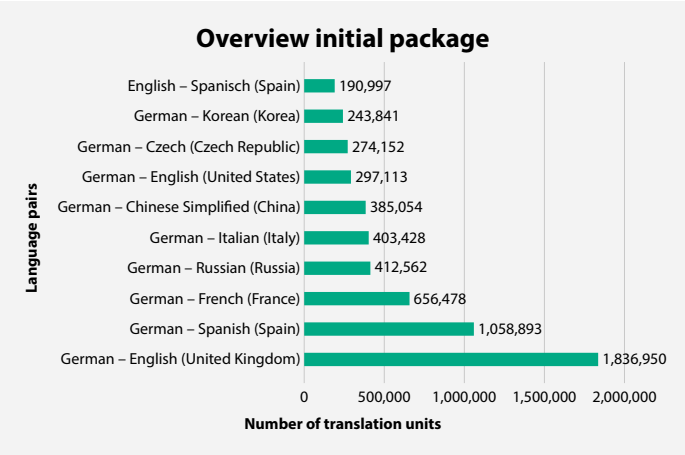
A convinced customer

Once the expectations had been discussed and a few tests had been carried out, Festo eventually decided in favour of [itl]-clean. Other than the requirement that the clean-up should run as anticipated, the main criteria were:

- A separate database with the data that had been removed
- The trouble-free re-import of the cleaned-up data (the tool does not change the structure of the tmx-files in any way)
- A report of the entire process

Project facts and figures

Once the first round of the clean-up process had been carried out, 10 language combinations with a total of 5.8 million translation units had been tidied up.



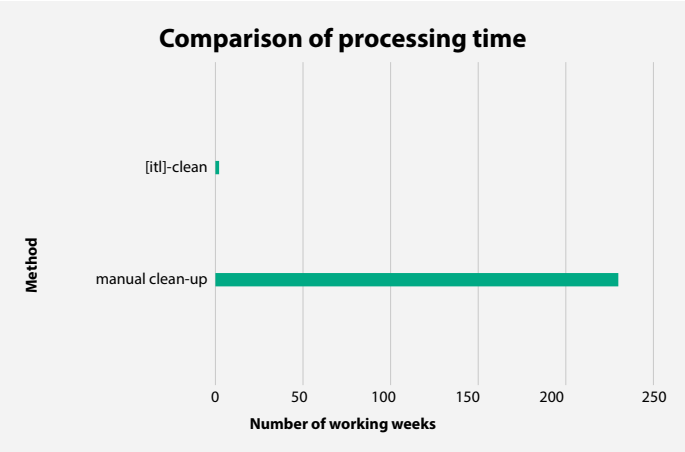
Processing time

It only took two weeks to automatically check all the above-mentioned 5,759,448 translation units. The only manual work needed was just a few hours at the start to determine how the data was going to be selected and agree on the parameters to be checked.

This means that only a small data volume was added during the clean-up process. This was easy to identify using the dates of the units and so could be included in the next clean-up.

If this process had been done manually, assuming that, in an ideal scenario, one translator per language would work on this task full time, it would take roughly around four years.

Incidentally, the software architects at itl are nowhere near satisfied yet with the current performance of [itl]-clean and are already refining it to make it quicker.

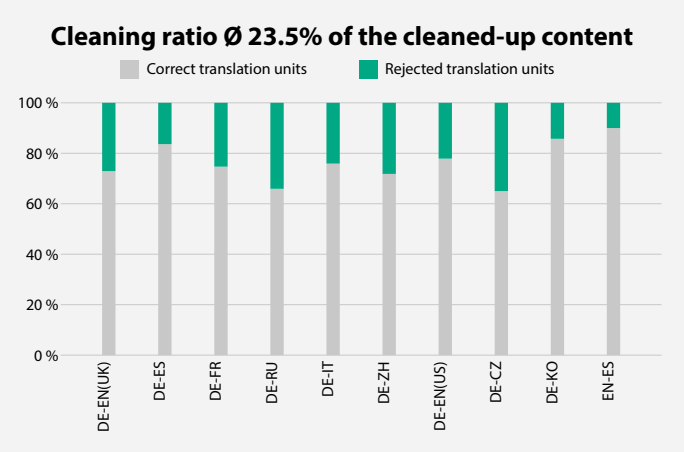


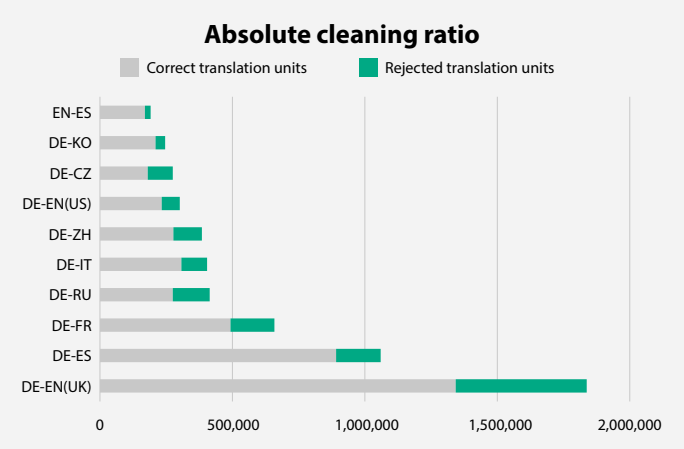
Report of the results

After the completion of the clean-up, the exciting question was of course: how much had been removed and, more importantly, what?

How much?

It came as no surprise that 23.5% of the content was removed; given that the amount of data was quite considerable and that some of it was quite old, this was perfectly understandable. The values for the individual language combinations were between 10% for the smallest translation memory (English-Spanish) and 34% and 35% respectively for German-Russian and German-Czech.





What?

Since this was the first time that [itl]-clean was used, and used extensively, it was only logical that Festo insisted on reviewing the results and check the contents. To do this, random samples were taken from all databases in which the removed segments had been stored.

The results did indeed meet expectations. The most frequently occurring errors were incorrectly allocated languages and terminology errors.

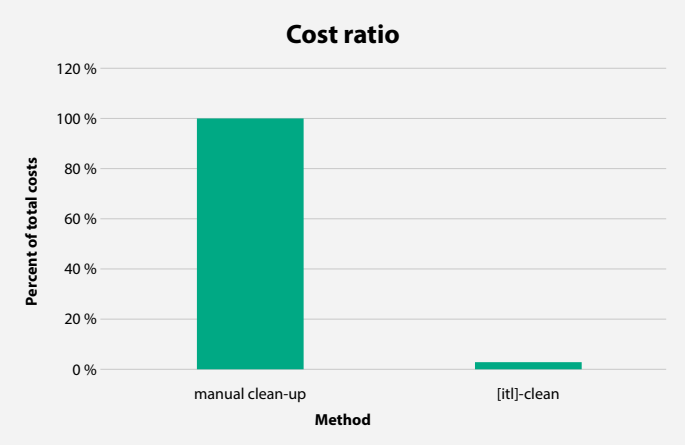
A fearful moment

Then, we suddenly got feedback at itl that proofreaders at Festo had discovered a number of wrong translation units with incorrect terminology in the material that had been cleaned up. The itl developers were immediately called in and they set to work straightaway. What would happen if a mistake in the system would call into question what had already been achieved?

Thankfully the all-clear was given after just a few hours. It was all down to people and the instruction they had given the system. One-word segments should by default not be cleaned up, as it is possible that the source text needs to contain the same wording as the target text – think for example of the word "server". As a translation unit with just one word, this was a segment that had been marked as "incorrect". Thanks to the high performance of the [itl]-clean software, the parameters and test runs are very easy to adjust.

The budget

Similar to the comparison of processing times mentioned earlier, it is not difficult see the cost benefits. Even when the costs for a review by a human translator are low, the size of the TM content can quickly lead to spiralling costs without any cost/benefit ratio whatsoever.



Long-term benefit

Now that the efficiency of cleaning large databases is much improved, which in itself is a real breakthrough, it is also possible to implement these processes as part of a regular quality assurance cycle. This would avoid creating databases that are so large that a thorough clean is needed, while the intervals between clean-ups will be shorter and the effort will be noticeably reduced.

What's next?

In the meantime, Festo has sent us other translation memory databases that need to be cleaned up, while the "clean" ones are used once again for ongoing translation projects.

Since the data has been updated and the data packages are now smaller, the company can take the next step in making the most of the precious data in its translation memories.

The experts in *intelligent solutions* and *Localization Engineering* at itl are now on the hunt for the next trend that would help customers and itl to use the latest technologies from the world of AI, cloud native and system connectors to best effect. This would give those in charge of translations, project managers, translators and everyone else more time to dedicate to the language, the processes and quality – instead of sorting out, moving around or preparing data.

About Festo:

The Festo Group is a global player in the field of digitalization. As a worldwide leader of automation technology and technical education, Festo is firmly focused on digitalization and its products and services are aimed at enabling the smart production of the future. To achieve this, the company is using artificial intelligence and machine learning as key tools.

Having been founded in 1925, this independent family-owned company with headquarters in Esslingen am Neckar has been setting trends in automation for over 60 years, while its unique offer has made it the world leader in technical training and education. 300,000 customers in factory and process automation rely on the company's pneumatic and electric drive solutions.

About itl:

Since 1982, our goal has been to help manufacturers impress their customers with great technical documentation. Worldwide. As a full service provider, we don't just focus on the text and graphics in the final print or online product – we're mindful of the background infrastructure, software, and workflows too. At itl, we help you to optimize every single aspect while always ensuring outstanding quality.

We are a reliable partner for our customers and combine classic technical documentation with a start-up spirit that characterizes the development scene. Our answers to the current demands of industry are:

- Concepts based on images and digital documentation
- Machine translation
- [i]-match, our in-house developed language management platform
- Development of automatic interfaces between customer and service provider

Our core competencies are translation, editing, graphics production, publishing and consulting, with special emphasis on optimising and automating all processes within the documentation life cycle.

itl Institut für technische Literatur AG
Elsenheimerstraße 65-67, D-80687 Munich
+49 89 892623-610
vertrieb@itl.eu