Translation Memory

A White Paper introducing a key component of the Johnson & Johnson Global Translation Center
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation Memory</td>
<td>2</td>
</tr>
<tr>
<td>Contents</td>
<td>3</td>
</tr>
<tr>
<td>An Introduction to Translation Memory</td>
<td>4</td>
</tr>
<tr>
<td>How Translation Memory Works</td>
<td>5</td>
</tr>
<tr>
<td>Exact Match (100%)</td>
<td>7</td>
</tr>
<tr>
<td>Context Match (101%)</td>
<td>7</td>
</tr>
<tr>
<td>Fuzzy Match (&lt;99%)</td>
<td>7</td>
</tr>
<tr>
<td>Fuzzy Matching</td>
<td>7</td>
</tr>
<tr>
<td>Types Of Matches</td>
<td>7</td>
</tr>
<tr>
<td>Matching Algorithms</td>
<td>8</td>
</tr>
<tr>
<td>Stop Words</td>
<td>8</td>
</tr>
<tr>
<td>Punctuation</td>
<td>8</td>
</tr>
<tr>
<td>Word Order</td>
<td>8</td>
</tr>
<tr>
<td>Attributes</td>
<td>9</td>
</tr>
<tr>
<td>TM Analysis Reports</td>
<td>9</td>
</tr>
<tr>
<td>Control and Reporting</td>
<td>9</td>
</tr>
<tr>
<td>Pros and Cons of Translation Memory</td>
<td>10</td>
</tr>
<tr>
<td>Functional Considerations</td>
<td>11</td>
</tr>
<tr>
<td>The Johnson &amp; Johnson Global Translation Center</td>
<td>12</td>
</tr>
<tr>
<td>Conclusion</td>
<td>13</td>
</tr>
</tbody>
</table>
An Introduction to Translation Memory

Introducing a key component of the Johnson & Johnson Global Translation Center

One of the biggest challenges facing global business today is the orchestration and management of information. As the amount of information produced by companies has grown, so has the need to translate this information for the markets they serve. The desire to increase translation productivity, to deal with the huge increase in translatable information, has led to the development of Computer-Aided Translation (CAT) tools such as Translation Memory.

This paper aims to explain what Translation Memory is, how it works in the context of the Johnson & Johnson Global Translation Center, and the benefits it brings to both users and the Johnson & Johnson family of companies as a whole.

Academic research during the 1970s and early 1980s that focused on computer modeling and “Automated Language Processing Systems” led to the idea of using machines to facilitate the automatic processing of a source text into a target language.

The idea of a Translation Memory is often attributed to a 1978 article by Peter Arthern where he describes the concept as follows. “Any new text would be typed into a word processing station, and as it was being typed, the system would check this text against the earlier texts stored in its memory, together with its translation into all the other official languages [of the European Community].”

Translation Memory has evolved into a sophisticated tool used by companies across the world, enabling them to translate information into a vast range of languages, faster and cheaper than ever before. The rest of this paper will describe how Translation Memory works, its pros and cons and how it works at the heart of the Johnson & Johnson Global Translation Center.
A Translation Memory (TM) is a bilingual database containing source and previously translated texts. A Translation Memory enables a translator to reuse, or leverage, previously translated text in new translations, avoiding the need to translate new text from scratch.

Data organized and contained in a Translation Memory comprises two types of text; source texts that are original texts from which translations have been made, and target texts that contain the translated versions of the source texts. A Translation Memory always includes a language pair. An EN_DE Translation Memory, for example, would include English source texts and German translations.

In order to be stored in a Translation Memory, source and translated text are first broken into small segments. Segments are small chunks of text that are automatically defined based on formatting or punctuation, such as sentences, titles or bullet points. Each source segment and its corresponding translated segment are then connected together to form Translation Units.

---

**Chinese, the most spoken language around the world, has over 50,000 characters. But you only need to know 2,000 of those characters to be able to read the newspaper.**
If a text has been translated without the use of a Computer Aided Translation (CAT) tool, and both source text and its corresponding translation are available, a translation memory can be created for use in future translation projects, through a process called post-translation alignment. This process is often used to build a new TM from an existing corpus of source and target texts.

Segments are small chunks of text that are automatically defined based on formatting or punctuation, such as sentences, titles or bullet points. Each source segment and its corresponding translated segment are then connected together through a process called post-translation alignment, to form Translation Units. This process is often used to build a new TM from an existing corpus of source and target texts.

When a translator begins a new project, he or she loads both the source text and Translation Memory into a Computer Aided Translation (CAT) tool. The CAT tool starts by automatically dividing the source text into segments. Once the text is segmented, the CAT tool then begins comparing the source segments from the new text to source segments stored in the Translation Memory.

If it finds a segment that exactly matches one stored in the Translation Memory, it retrieves the corresponding Translation Unit and inserts the previously translated text into the new translation. This process enables translators to quickly accept or modify previous translations for use in the new text.

For source segments where a match with the Translation Memory has not been found, the translator translates the source text from scratch, creating a new Translation Unit which is stored in the Translation Memory. As the Translation Memory is updated and augmented in real-time, new Translation Units can be immediately reused in other areas of the translation project, further improving efficiency.

Some advanced CAT tools even allow multiple translators to work with a shared Translation Memory, updating it in real-time and using each other's translations the moment they become available.

There are over 231 completely extinct languages and 2,400 of the world’s languages are considered to be in danger of dying out.
Fuzzy Matching

Leveraging previously translated content to improve translation efficiency

By using CAT tools and a Translation Memory together in the process outlined above, it’s quick and easy for translators to locate and automatically translate exact matches. Depending on the type of translation project and the size of the Translation Memory the amount of 100% matches can vary. For technical documentation translation projects, where text is repeated often, or where text is added to or updated incrementally, the number of 100% matches between source text and Translation Memory can be high. However, CAT tools are also "smart" enough to identify partial matches through a process known as Fuzzy Matching. For projects where the number of 100% matches are lower, Fuzzy Matching employs a computer algorithm to identify segments in the source text that are similar to, but not exactly the same as those stored in the Translation Memory.

The ability to identify partial, or Fuzzy Matches, allows translators to further increase efficiency, leveraging more segments from the Translation Memory and ensuring existing translated text can be reused as much as possible.

Types Of Matches

**Exact Match (100%)**

An exact match (or 100% match) happens when a source segment matches, word for word and character for character, a segment stored in a Translation Memory.

**Example:**

The cat sat on the mat. (source)

The cat sat on the mat. (TM)

**Context Match (101%)**

Translation Memories also look at preceding and sometimes following segments to determine whether a source segment and Translation Memory segment share the same context.

**Example:**

The dog sat by the door. The cat sat on the mat. It was raining outside. (source)

The dog sat by the door. The cat sat on the mat. It was raining outside. (TM)

**Fuzzy Match (<99%)**

A fuzzy match is a likelihood, expressed as a percentage, that a source segment matches a segment found in a Translation Memory.

**Example:**

The cat sat on the mat. (source)

The dog sat on the mat. (TM)
Matching Algorithms

How translation tools calculate matches and reuse previously translated content

Although CAT tool vendors do not release the algorithms they use to calculate fuzzy matches some common factors that affect the matching score include:

**Stop Words**

Giving stop words like “and”, “of” and “the” the same weight as other words can cause potential issues. For example;

*Description of the Service and Definitions. (source)*

*Ownership of the Services and Marks. (TM)*

In the example above 3 out of 6 words are the same, and one word has one additional character (Service vs. Services). By treating stop words in the same way as regular words, a fuzzy match score of 66% would be returned.

4 out of 6 or 4/6 = 66% match. If we ignore stop words, however, it quickly becomes apparent that the above two segments are not a suitable match.

**Punctuation**

Punctuation is a difficult issue since sometimes it can affect meaning, and other times it is not important when deciding if a translator can leverage a segment from a Translation Memory. For example;

The cat sat on the mat.

the cat sat on the mat

However, punctuation can also drastically change the meaning of a sentence, and ignoring it may have unintended consequences. For example:

Private. No visitors allowed.

Private? No! Visitors allowed.

**Word Order**

When calculating a Fuzzy Match it is important to take into consideration word order in addition to identifying matching words and punctuation. If word order was not considered, the following two segments would return an exact match:

I had cleaned my car

I had my car cleaned

Although each segment contains the same words and punctuation, it is clear they both have very different meanings.
Control and Reporting

Managing Translation Memories and monitoring their efficiency

Attributes

In addition to source and translated text, most Translation Memories also store information called “attributes” with each Translation Unit. The most common attributes stored include creation date, the name of the translator, client, project and subject matter domain.

Storing additional attributes gives translators and project managers much more granular control over how a Translation Memory is used. For example, in the case of medical translations for packaging or instructions for drugs, regulatory oversight dictates that translations are certified and adhere to guidelines relative to the country where the drug is to be used.

After these translations are certified the corresponding Translation Units in the Translation Memory can be tagged with attributes to identify them as such. For future translations of packaging or instructions where certification is required, a translator or project manager can select only to use Translation Units that have the “certified” tag. This ensures that only “certified” translated text is leveraged from the Translation Memory, helping to mitigate any future issues when new projects are submitted for regulatory review.

TM Analysis Reports

Before a translation project takes place, the files for translation are segmented and compared with the Translation Memory to generate a Translation Memory analysis report.

The Translation Memory analysis report shows how many exact and fuzzy matches have been found, indicating how much the Translation Memory can be leveraged and how much new translation effort is required. Matches are normally organized as follows:

- 100%
- 95% to 99%
- 85% to 94%
- 75% to 84%

Since translation is priced per word, 100% matches are usually removed from the total word count, reducing overall costs. Percentage matches are then priced on a sliding scale, the higher the match, the lower the cost per word. Anything under 75% is considered a non-match and a new translation is required, charged at the full per-word rate.
Pros and Cons of Translation Memory

Practical considerations when using Translation Memory technology

A Translation Memory is designed to be a tool to aid human translators, not replace them. Therefore it is important to understand the considerations and constraints to be able to use Translation Memory efficiently and cost effectively across large translation projects.

Since a Translation Memory is merely a storage and retrieval system, the quality and efficiency of the Translation Memory is subject to the scope and quality of translations it is provided. Translation Memory works best when translating text that has a high level of repetition across content types, or text is updated on a regular basis. Good examples where Translation Memory can be used effectively are Technical Documentation, Product Catalogues, User Interfaces and Instruction Manuals. Texts that are written about highly specialized subjects are also good candidates for a Translation Memory especially if the subject matter and format do not vary greatly, and the translators have good subject matter expertise.

Understanding the subject matter and context of source texts is important to translators when deciding to reuse segments in new translations. One of the main challenges when using Translation Memory is what’s known as de-contextualization. Since segments are treated as individual blocks of text and are leveraged using an algorithm, translated segments proposed by the matching algorithm might not be suitable when the meaning of the entire source text is taken into account.

Translated segments are often produced by a team of translators, each with their own subtly different styles or word choices. In these instances, matched segments created by different translators can sometimes read very strangely when placed in the final text. To solve this problem, attributes can be used to ensure translator, subject matter, and project type can be selected and applied to the matching process, ensuring much greater continuity and consistency in the final translated text.
Functional Considerations

Translation memory is remarkably precise in the way that it categorizes and stores Translation Units. This makes it sensitive to changes in source and translated text over time.

For example, if multiple versions of a source text are translated using a Translation Memory, with each subsequent source version having text added to it, a Translation Project Manager would expect each subsequent new source version to leverage the old translations at 100% match. If a source file is added to, but the original text is not changed, the Translation Memory will recognize the old segments, match them 100% and then attempt to find matches or fuzzy matches for the added content.

However, if a new source version is reviewed and edited from beginning to end, before being translated, this can have a significant impact on the number of exact matches returned using the same Translation Memory.

The reason this happens is because Translation Memory software finds matches between segments by recognizing sequences of characters and identifying patterns. Using pattern based recognition to match segments, while extremely effective, means the software will not recognize matches between segments that are written differently, but have the same meaning.

For example;

The Acme product is both streamlined and lightweight and can be used in many situations. (TM)

Acme products are streamlined and lightweight, enabling their use in a variety of situations. (Edited source)

While both segments mean the same thing, they are now very different in terms of structure, (the sequence of characters is drastically different). Adding or removing words or punctuation when editing source text causes differences in both segmentation and alignment to be introduced. This means that when the edited source text is analyzed and compared against the Translation Memory, it could result in the edited segment returning a fuzzy match lower than 75%, requiring a new translation.

It is therefore important to keep source text structure as consistent as possible when using a Translation Memory to ensure the highest level of leveraging and cost savings.

Of the population of the United States, 21 percent of citizens five years and older speak another language at home. Of that 21 percent, 62 percent speak Spanish.
Translation Memory forms the core of the Johnson & Johnson Global Translation Center, enabling users and suppliers to benefit from the cost and time savings that the technology offers. The Global Translation Center acts as a single repository for Translation Memories from all parts of the organization, storing them for easy retrieval and use in translation projects. Johnson & Johnson users and suppliers can upload and download Translation Memories by language pair and use the search and filter functionality to find Translation Memories that meet specific project criteria.

Rather than saving Translation Memories individually the Global Translation Center saves all Translation Units from all Translation Memories in a single repository. When a user downloads a Translation Memory it is created from all Translation Units stored in the central repository that fit the attributes the user requires. This enables, for the first time, sharing of Translation Memories across teams, business units and approved suppliers.

For example, two remote teams working on very similar product information, translating using the same language pairs, have the ability to quickly leverage a common Translation Memory. Something that was once hindered by issues of version control and file management is now simple. The two teams and their respective suppliers can now each upload their existing Translation Memories and download new versions comprising a combination of both, leading to significant cost and time benefits.

Attributes give users and suppliers more control over how Translation Memories are used. For example, if a Translation Memory needs to remain confidential, or is only suitable for use in a certain subject area or business unit, access can be restricted. An example might be a Translation Memory used for EN_DE packaging translations, where the DE translations have been certified by a legal representative. This Translation Memory would need to be locked to avoid anyone else changing the certified Translation Units contained in it.

The flexibility and power of the Translation Memory repository at the heart of the Global Translation Center gives Johnson & Johnson users and suppliers significant control that was previously unavailable at an enterprise level, all while remaining simple to use. This in turn leads to a huge benefit across the organization since leveraging a single centralized Translation Memory repository offers cost and time savings far in excess of what can be achieved using multiple Translation Memories distributed across teams, departments, business units and suppliers.
Conclusion

Translation Memory is an important technology for any company engaged in large-scale translation projects. By recycling previous translations for use in new content, Translation Memories drastically reduce the time it takes to bring new products and services to global markets.

As Translation Memories grow with new translations added on an ongoing basis, they become more accurate and deliver significant efficiencies. By centralizing multiple Translation Memories into one platform, the potential cost and time savings are enormous.

Johnson & Johnson has chosen to invest in the technology that will make centralization and sharing of Translation Memories easier than ever before and users of the Global Translation Center will see immediate benefits. Translation Memory is set to help Johnson & Johnson’s products and services reach more patients, practitioners and consumers, in more markets more efficiently than ever before.