Integrating Machine Translation into Translation Memory Systems

Matthias Heyn

Abstract

Within the last few years, there had been a remarkable change within the use of tools at the desktop of professional translators. Whereas, traditionally the keyword for the automation of the translation process had been *machine translation* (MT) this has significantly changed in the last few years towards the usage of *translation memory systems*. On the other side, MT-systems are more and more targeting their genuine market. Non-professional users' main interest lies in "quick information translation".

This general development doesn't mean that MT is not used any more at translator's desktops. It rather means, that the role of MT in a professional environment has significantly changed. MT for professional translators means that MT is one software component among other ones within a central translation memory system. MT is reduced to a "proposal machine" for worse case situations: if no information at all is accessible or if the source is simple enough.

Integration of MT into translation memory systems can be done by several architectures and this paper will investigate the different possibilities and their pros and cons. Existing integrations with the TRADOS Translator's Workbench for Windows will be discussed.

Matthias Heyn

Matthias Heyn studied at the universities of Heidelberg and Stuttgart *linguistics and information science*. He worked as a lecturer at the University of Heidelberg and published in the fields of lexicology and computational linguistics. He specialised in corpus research, alignment strategies and computational lexicology / terminology before he started in 1992 to work with TRADOS Gmbh in Stuttgart. In 1994 he got manager of the research and development department of TRADOS and founded in 1995 TRADOS Benelux S.A. in Brussels. He is currently managing director of TRADOS Benelux S.A.

Trados GmbH

TRADOS is widely acknowledged in developing software products in the field of translation tools. TRADOS exists now 12 years and therefore is one of the most experienced companies in the field. With more than 8000 users of TRADOS products - to a large extent terminology database systems and translation memory systems - TRADOS has in-depth knowledge of user needs and requirements, the current market situation and linguistic knowledge within specialised languages. TRADOS is recognised to be one of the most successful commercially driven companies in the field of CAT-tools. TRADOS International Network has offices in Germany, Belgium, Great Britain, Spain, Sweden and Switzerland and resellers all over the world.

Matthias Heyn
Trados Benelux SA/NV
303, Avenue de Tervuren, B-1150 Brussels, Belgium

Tel: +32 2 775 84 70, Fax: +32 2 775 84 80
E-mail: matthias@trados.com

1. Introduction

Over the last few years, the use of tools at the professional translator´s desktop has significantly changed. Whereas the keyword for the automation of the translation process used to be *machine translation (MT),* the dominant notion for **language professionals** is nowadays *translation memory systems (TMS)*. This development is due to several factors: generally speaking, language professionals are experts, dealing with semantics and pragmatics much better than any machine can do. They do not need to bother with imperfect machine translations, but they do need substantial aid in the organization of their work concerning terminology and retrieval of existing human translations. A TMS now takes over this part: the machine does what a machine can do best.

On the other hand, MT has proven its worth in informative translation, helping non-professionals to understand the rough meaning of documents. Therefore, MT vendors start now targeting the market of the standard application user by means of marketing and pricing.

This general development does not necessarily mean that MT has no future at the translator's desktop. It rather means that the role of MT in a professional environment has changed significantly and hence must be redefined.

In this paper, we will have a closer look at that new role of MT within a language professional's environment and see that the integration of MT into TMS can sometimes lead to fruitful synergies. We will investigate the conditions for the use of MT and discuss several integration architectures. Furthermore, we will give a few examples of existing integrations in the TRADOS Translator's Workbench for Windows.

2. Shifting towards Translation Memory Systems

There are some general reasons for the recent success of translation oriented software applications:

- The general tendency towards computerisation of text flow, gives translators more and more access to machine readable source documents.
- The processing power of modern (desktop) computers enables functionalities that were not available in the past and that are crucial to the successful implementation of translation tools on standard machines. Generally speaking, all improvements to the hardware are very welcome in this specific application area.
- The integration of translation software into the translator's software environment has improved considerably.
- Translation software has met the overall quality standards of the industry software with regard to user friendliness and software ergonomics.
- The knowledge of translators about the benefits of computerising their work is steadily growing.

Within this general trend towards the use of translation software, we can distinguish between two approaches: MT and TMS. Both approaches are defined as follows:

A TMS stores in a computer all translations made by a translator. In case of re-translation, these translations are retrieved automatically.

An MT system applies grammatical rules and information from dictionaries to a given source sentence in order to translate it.

These two approaches to translation are quite contradictory. MT tries to model the translation process, so to speak replace the translator; whereas a TMS supports the translator by making the individual translation process reproducible.

2.1 What comes out of the system?

Whereas a TMS can be described as a system where *all output is based on human input;* an MT system can be described as a system where *all output is performed by a machine process*. TMS avoids generative capacities whereas MT relies on them. In a professional environment, this means that a translator can blindly rely on any TMS, if he or she trusts the translator who has previously worked with this system. In contrast to working with a TMS, an MT translator can never trust the output and has to proceed to a time-consuming and boring revision- (or better: repair-) phase.

2.2 What does the system learn?

Another interesting feature of TMS is the "learning"-factor: a sentence has to be translated once and never to be translated again, whereas with an MT system, a translation, with its possible errors, is always *re-generated*. In other words: repetitions are only learned once within a TMS, where MT always re-generates them.

2.3 How to get a better profit from the tool?

A very important aspect for language specialists is the "tuning" of a system. In the case of TMS, this is very simple. The only thing a translator has to do is translating with this system. Since a TMS "learns" in the background the introduced translations, it improves automatically. There is no other specialist knowledge required but good professional translation skills.

Improving MT output is a rather tedious work. Documents can be preprocessed by using controlled authoring or controlled language mechanisms; output has to be postprocessed and revised; the dictionary component of the MT system can be updated; grammar rules can be adapted etc.

Updating the MT dictionaries is always very time consuming and requires specialized knowledge at different linguistic levels. Updating the generative core component of an MT system (the "grammar base") is difficult and may yield side-effects that are almost always uncontrollable. In short, tuning an MT system is rather complicated and time-consuming and requires skills beyond standard application user knowledge and beside the standard skills of a language professional.

2.4 Psychology of a tool

MT tries to *replace* the translator, a TMS is doing the opposite: it tries to *support* the translator. A translator who works with MT looks like working most of his time against the machine because of operations like error-prevention and error-repair. A translator working with a TMS is feeling even more responsible about her or his translations because their work is going to be "re-used".

To summarise: a TMS frees translators from boring and repetitive tasks and lets them concentrate on what they do better over machines, i.e.: handling the semantics and the pragmatics. This generally leads to a broader acceptance of TMS by language professionals.

2.5 What is needed by language professionals?

Professional translators do not have problems with morphology and syntax but with semantics and pragmatics. In most cases this has to do with lack of knowledge about the subject area or, in other words, with lack of terminology and specialised language collocations. A professional translator does not need a system that handles syntax and morphology, but a reliable term bank or a translation memory covering the subject field.

Maybe an analogy can help to clarify the relationship of MT and TMS as software products: there is software for the benefit of professional users and software for simulating the professional. We can think of a software that tries to replace partly an accountant and software that is used by accountants; or else: software that tries to simulate the skills of an architect and software that is used by architects to facilitate their job. MT tries to simulate a translator and a TMS software is used by a translator to do a better job.

3. MT as an “add-on”

From the above mentioned discussion we can redefine the role of MT in the field of professional translation. First of all, we can narrow down the scope of conditions for a successful use of MT. If a translator is confronted with a sentence and:

- this sentence or a sufficiently similar sentence cannot be retrieved by a TMS;
- the sentence lies syntactically more or less within the scope of the capabilities of the MT system;
- there is a certain coverage of the MT dictionaries of the required subject area;
- the MT system is capable to preserve the formatting;
- the MT system is a keystroke away and responds quickly (or has already prepared a translation over a previous batch process);
- the MT system uses the terminology of the private term bank system of the translator,

then probably a good proposal of the MT component can speed up editing time. The proposal can be corrected and next time the memory is in charge of the sentence!

Therefore, we can describe MT in a professional context as "a proposal machine” that can be switched on and off - dependent on the conditions of the text to be translated. MT is not a core component, but plays a subordinate role as part of a set of useful tools a translator can choose from, like spell-checkers, electronic dictionaries etc..

4. Integrating MT into TMS

In principle there are two possible architectures to integrate MT into TMS. It can be integrated using batch processing or interactive integration. Already in the earliest

development phase of the object-oriented class system of the TRADOS Translator's Workbench for Windows, was decided for the implementation of a neutral layer for machine translation integration. This enables to integrate an existing MT system seamlessly either in batch or as an interactive component.

4.1 Batch integration

Batch integration means that a text passes first through an analysis process of a TMS which sorts out all sentences (translation units) that are unknown to the TMS. These sentences are then passed on to the MT whereafter the results are reimported into the TMS. The access to the TMS yields now results that are probably MT output and not former human input. It is self-evident, that these entries have to be marked and treated separately.

Batch integration can be implemented easily: MT and TMS only have to communicate over a common file exchange format. The disadvantages of this approach are the following:

- the required file processing adds an additional preprocessing phase to the translation process;
- it disables the translator from making interactive decisions such as expanding or shrinking phrases, correcting on the fly errors in the source text (typos!), adding terms to the termbank, switching to a different MT lexicon (or changing the access sequence of MT lexicons), adding an abbreviation in order to avoid segmentation faults, etc.. This kind of changes can only be respected by the MT system after manually restarting the complete batch process and are therefore in practice left out most of the time.

On the other hand, even slow MT systems can be integrated over batch, since the proposals are quickly accessed over the TMS.

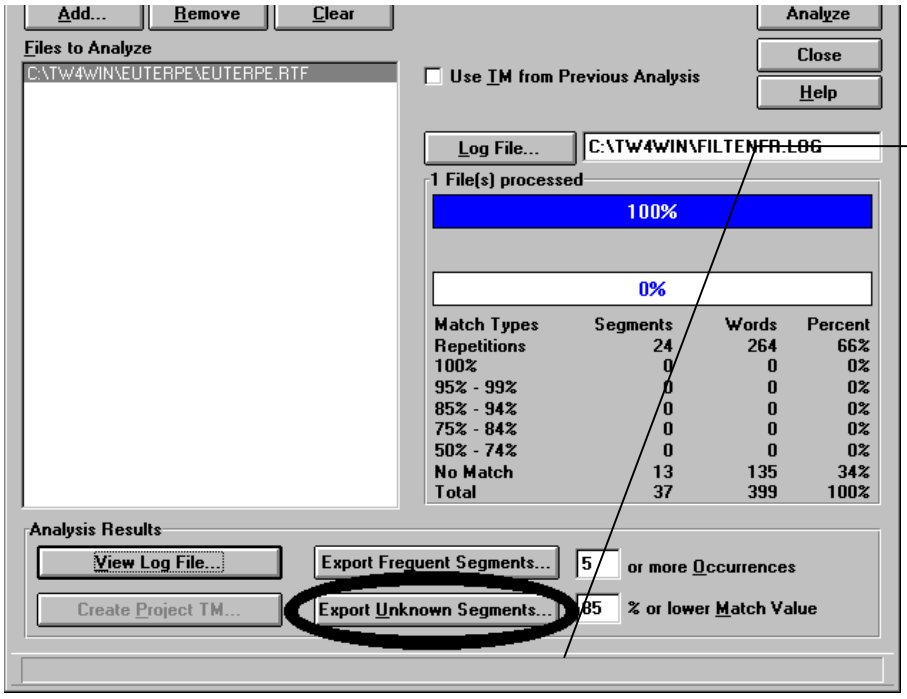
Batch integration is often the only possibility to integrate an MT into a TMS. This is the case e.g. if the MT is not running on the same platform as the TMS or if the MT is too slow for interactive integration or if the exchange protocols are not fit for adaptation.

4.2 Example batch integration

4.2.1 LOGOS Machine Translation combined with the TRADOS Translator's Workbench for Windows

The TRADOS Translator's Workbench for Windows integrates the LOGOS MT system using a typical batch processing environment.

A given document (RTF or WordPerfect format) is analysed in order to detect all segments (sentences) that are unknown to the current translation memory of the TRADOS Translator's Workbench for Windows. Fig 1. shows the menu where this operation will be performed.



By pressing this button, all segments with a similarity lower than 86% compared to the translation memory, are written to an export file, and then passed to LOGOS for translation.

Fig. 1: The “Analyse dialog” of the TRADOS Translator’s Workbench for Windows

This file is then passed to the LOGOS machine translation, which translates the segments. The result of the machine translation is then reimported into the TRADOS Translator’s Workbench for Windows. During this process all formatting information of the source text will be respected by LOGOS and the formatting is later stored within the TMS.

This approach has the advantage that the segmentation process is performed by the TMS, which ensures uniform integration into the later translation process while preserving the same segmentation.

After this preparatory work, the translator proceeds in the familiar way: from the well-known word processing environment (e.g. WinWord or WordPerfect) the TMS is interactively consulted. All proposals from the machine translation are marked as machine translations. In addition, the user has the option to set penalty values to "punish" machine translation entries (see Fig. 2 of the TRADOS Translator’s Workbench for Windows “Translation Memory Options” dialog window).

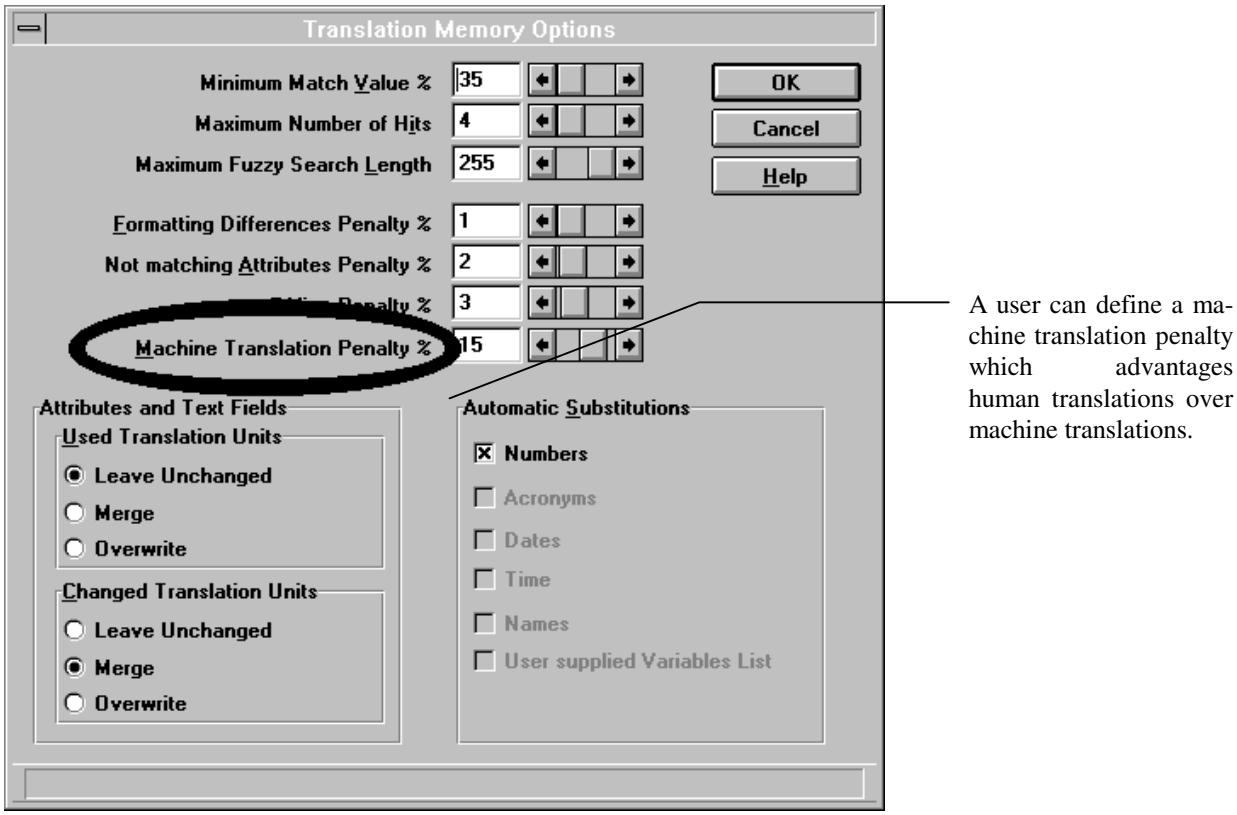


Fig. 2: The “Translation Memory Options” dialog

This mechanism ensures that human translations are proposed before machine translations to the user.

4.3 Interactive Integration

Interactive integration means that the user can interactively decide whether he sends a given segment from the source document to the MT and that the MT thereupon sends back a result. After correcting the errors of the MT system, the segment, as usual, is then stored in the TMS and retrieved in case of similiar or equal sentences.

Interactive integration is more difficult to implement and requires the same platforms as MT and TMS (or sufficiently powerful exchange protocols) and a quick response time from the MT.

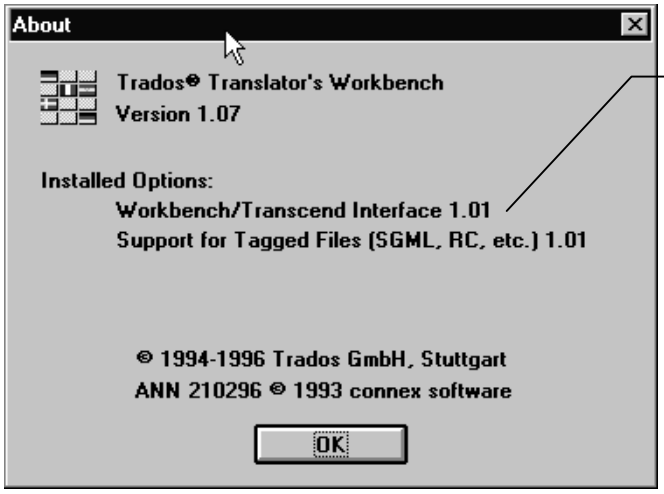
The big advantage of this solution lies in the flexibility for the user. Interactive decision-making like changing the segment sizes, correcting source text errors etc. can be performed without interfering with the MT.

On the other hand, interactive access is by nature slower than the batch access and must therefore rely on appropriate hardware and efficient MT systems.

4.4 Example Interactive Integration

4.4.1 Intergraph TRANSCEND and TRADOS Translator's Workbench for Windows

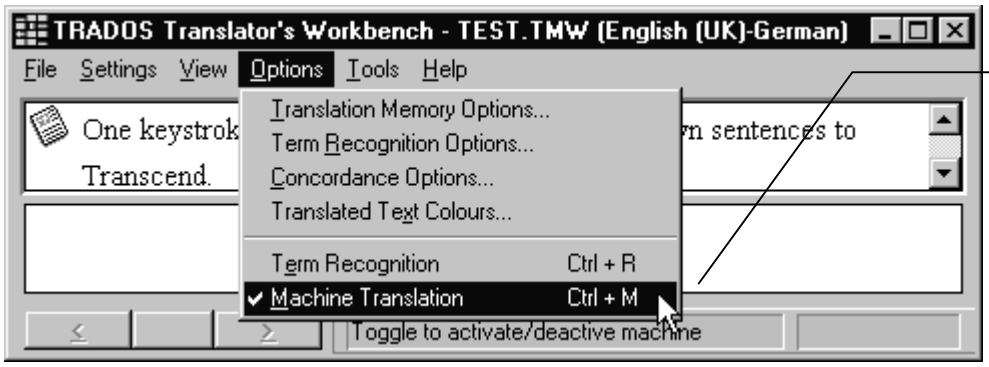
The interface to TRANSCEND is an extension to the TRADOS Translator's Workbench for Windows, which is automatically installed (See Fig. 3: TRADOS Translator's Workbench "About" dialog).



Under "Installed Options" additional plug-ins of the TRADOS Translator's Workbench are indicated. Automatically installed options are support for MT and Tagged file support.

Fig. 3: TRADOS Translator's Workbench "About Menu"

The TRANSCEND MT has to be loaded into memory in order to be accessible over the TRADOS Translator's Workbench for Windows. One keystroke (see Fig. 4) activates an option that makes all unknown sentences pass from the TRADOS Translator's Workbench to TRANSCEND.



One keystroke switches the MT on and off.

Fig. 4: Activating Machine Translation within the TRADOS Translator's Workbench for Windows

An example is given by Fig. 5, where the heading *After the wash* was taken from an instruction manual of a washing machine and could not be retrieved from the translation memory. Now, TRANSCEND English/French machine translation comes up with the proposal *Après le se laver.* The proposal of the machine translation is clearly distinguished by colours (a grey frame). If the translator now corrects the wrong proposal of the MT system and stores it in the TM, the next time the

translation is needed, the corrected version of the sentence will automatically be presented by the translation memory.

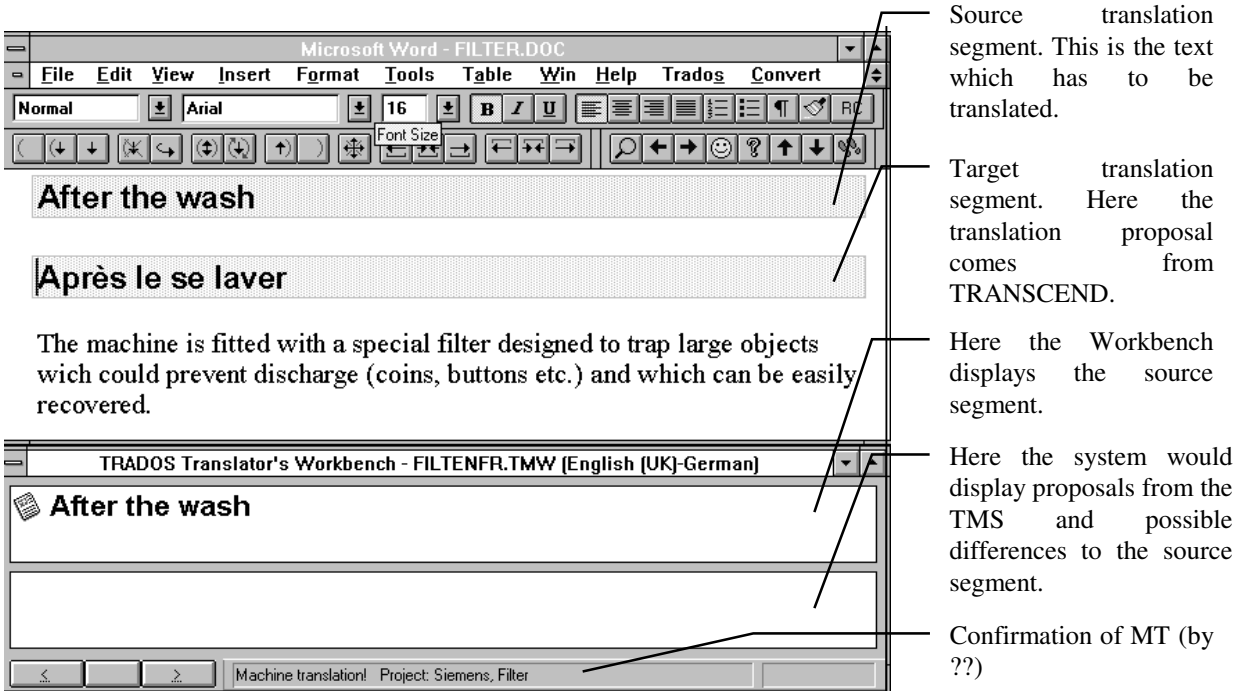


Fig. 5: TRANSCEND has been activated after unsuccessful search in the TM

5. MT lexicon versus TMS term bank

A general problem with the combination of a MT system and a TMS is that both systems provide lexicons. That means that probably two competitive sources have to work together. A possible solution could be the storage of the TMS information in the MT or vice versa or the passing on of information from the MT to the TMS or vice versa. In the context of MT used by language professionals, there seems one solution preferable.

MT dictionaries are specialised dictionaries for the explicit storage in a formal way of information from different linguistic levels. The more sophisticated the MT the more elaborate the dictionary structure. Scientific prototypes of MT are confirming this tendency with complex feature structure lexicons and specialised editors for these dictionaries.

The more advanced the system the higher has to be the (theoretical) language and information science competence of the user coding an MT dictionary. But, for all MT systems it is true that the coding time for one dictionary entry is rather high.

On the other hand, TMS are offering professional translators the possibility to encode their own terminology. This is necessary since one of the biggest problems for professional translators is to find wordings that are **not found** in standard dictionaries. Good TMS are offering sophisticated term bank systems that can be freely configured for appropriate terminographical work. In practice, considering

time constraints and production stress in a professional translator's environment, the available time for coding terminology is rather limited. That means often that under production circumstances only quick term-list equations are possible.

Now, let's look to both dictionaries from two angles:

5.1 Use of the MT lexicon by the TMS

MT lexicons are normally constructed round a core lexicon that covers the standard lexicon of a given language. This is not at all of interest for professional translator's. If there are specialised dictionaries available in MT systems these could be interesting for manual consultation by the translator. Precondition is a suitable access for "human" users or even better an import into the TMS term bank system.

5.2 Use of the TMS lexicon by the MT

In this perspective, it is very important that the MT respects the terminology of the translator. There are again two possibilities:

1. The contents of the term bank are transferred to the MT dictionary. In general, this involves the manual adaptation of the lacking linguistic information - which consumes a lot of time and effort.
2. The TMS is not only passing to the MT a particular sentence that has to be translated but also all known terminology of that sentence found in the term bank of the TMS.

The second possibility is certainly more appropriate, since a terminology database is more frequently updated and more easily maintained than an MT lexicon.

One can argue that the MT produces more errors if there is not enough linguistic information found in the lexicon, but in the environment of professional translators it has to be stressed that **the syntactical correctness is not important compared to semantic and pragmatic correctness**. The only reason for maintaining minimal linguistic information within term banks that have to be passed on to MT is that frequent morphosyntactical deviations are slowing down the editing process.

5.3 MultiTerm and TRANSCEND

An example solution for the passing on of terminological information from the term bank system of a TMS to an MT has been implemented within the TRADOS Translator's Workbench. The term bank component of the TRADOS Translator's Workbench - MultiTerm - is a fully fledged term bank system. MultiTerm allows for free database definitions and can be used in rather sophisticated terminology driven environments or within pragmatic production driven environments.

The TRADOS Translator's Workbench performs the automatic detection of terminology stored within MultiTerm ("term recognition") and points out this information to the user. Term recognition is in itself a rather complicated function which involves non-trivial tasks like the decomposition of complex compound phrases or the handling of separable verb-prefix constructions etc. Recognised terms are then passed to TRANSCEND. Besides the passing on of the terminology, the TRADOS Translator's Workbench also provides a protocol for passing on a few basic morphosyntactic features to TRANSCEND.

The complete MultiTerm entry of Fig. 6 can be written with 14 keystrokes. After adding this entry to MultiTerm and reopening the heading of the washing machine operation manual again, TRANSCEND produces the translation given with Fig. 7.

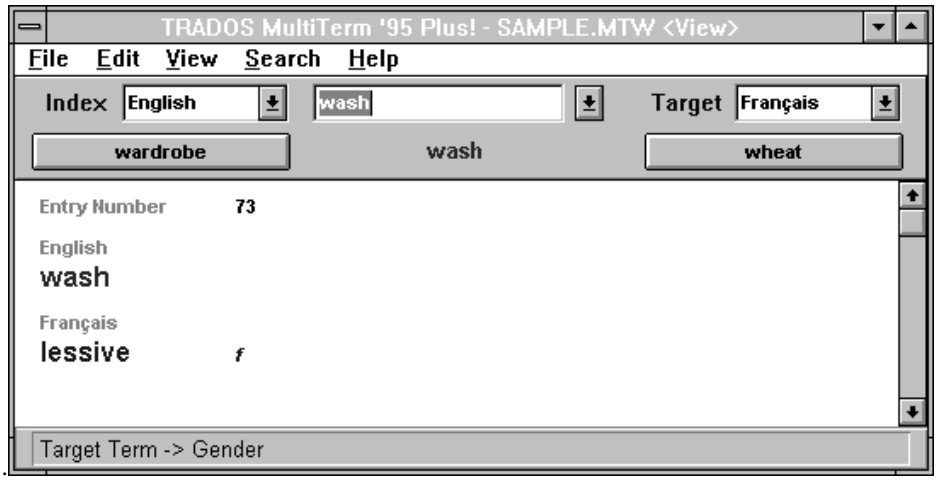


Fig. 6. MultiTerm database entry for English "wash"

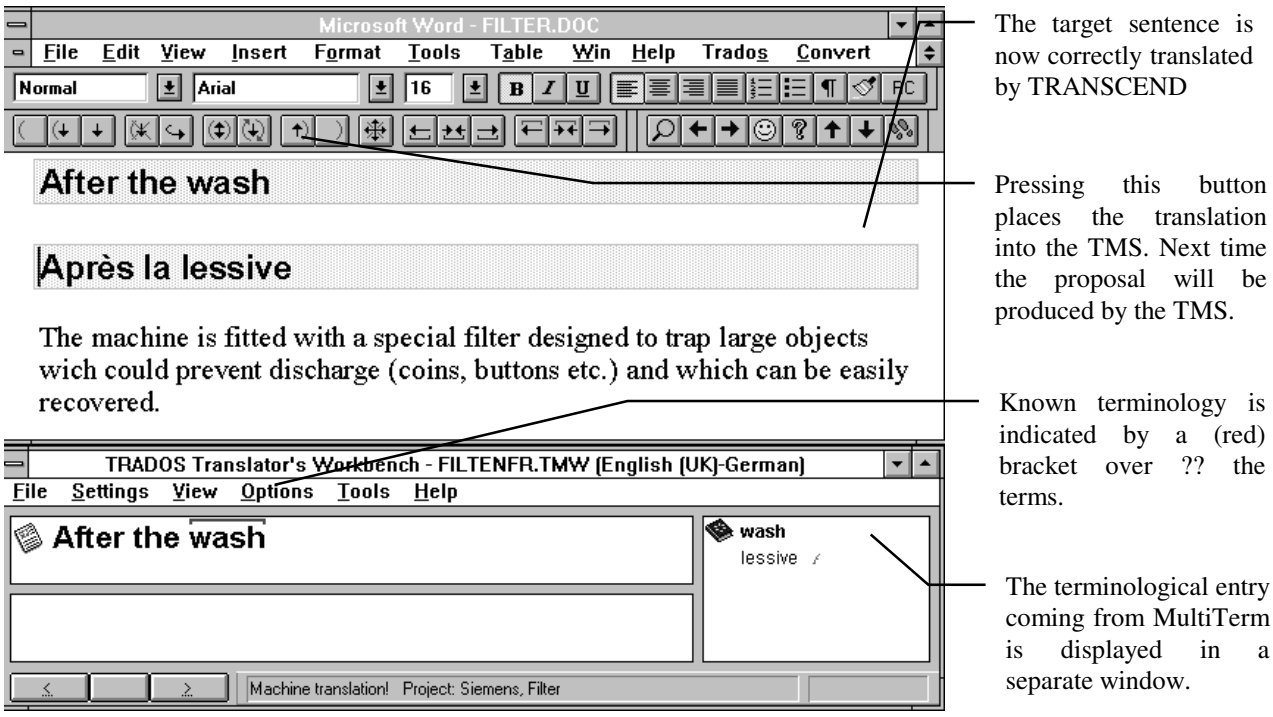


Fig. 7. Retranslation of TRANSCEND using a term bank entry of MultiTerm

When the translator now confirms the translation, it will be stored in the translation memory.

The next sentence in the text (*The machine is fitted ...)* is translated by TRANSCEND with: "*La machine est ajustée avec un filtre spécial a conçu pour prendre au piège de grand wich d'objet pourrait empêcher la décharge (les pièces, les boutons etc.) et*

qui peut facilement retrouvé." The translation is disturbed by a typographical error in the source (...large objects could...). The translator can correct the source error and restart TRANSCEND with two keystrokes. The result changes only slightly: "*La machine est ajustée avec un filtre spécial a conçu pour prendre au piège de grand d'objet pourrait empêcher la décharge (les pièces, les boutons etc.) et qui peut facilement retrouvé.*" In this case the translator does not benefit from the MT (or can even be hindered by it) and will certainly propose a different translation, such as e.g. the one given in Fig. 8.

Fig. 8. gives an example where the TMS retrieves a former manual translation in the case of a retranslation.

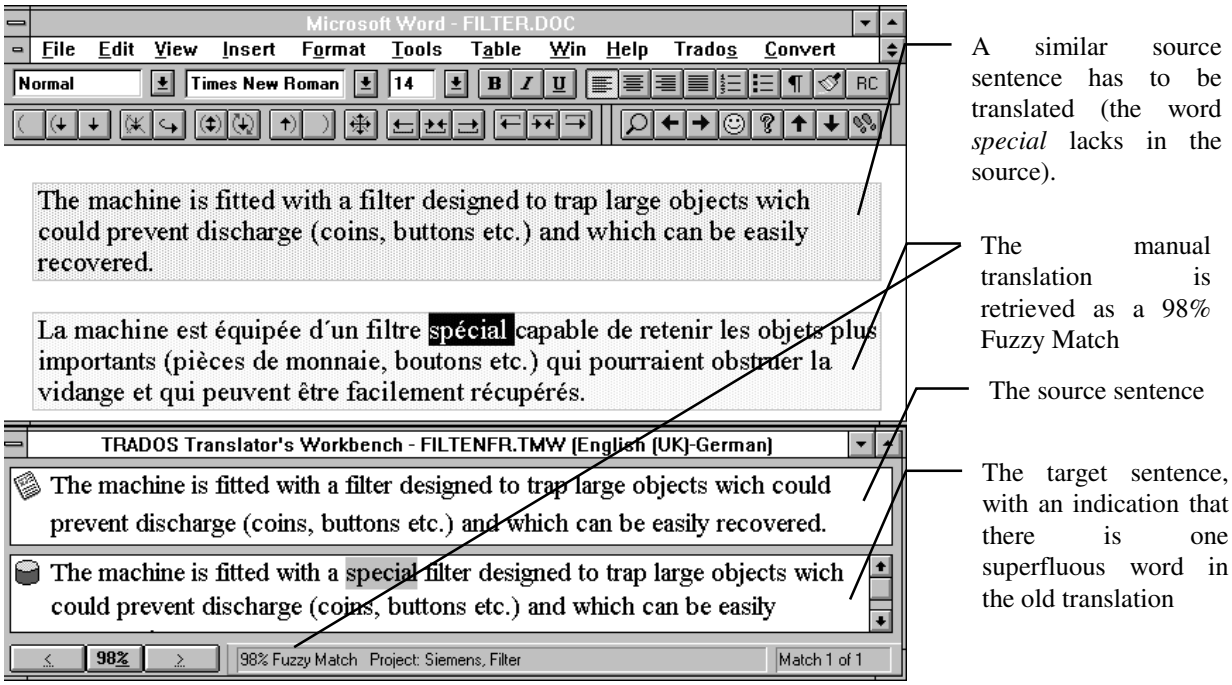


Fig. 8 Retrieval of a human translation within the TRADOS Translator's Workbench

A side effect of working with the TMS is that all translations are immediately retrievable in form of concordance searches. If a translator for example searches for *discharge prevention*, she or he will get the results shown in Fig. 9.

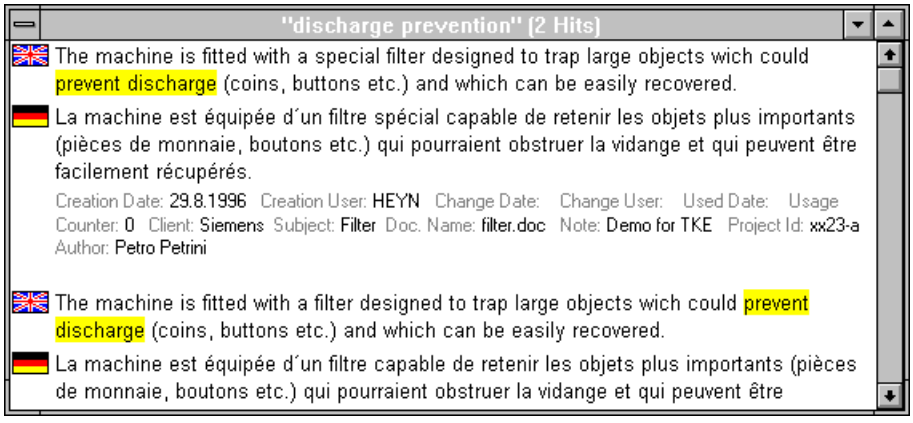


Fig. 9. Concordance search for *discharge prevention*.

Concordance searches are in many cases more important for reconstructing the semantics and pragmatics of a given translation task than using MT. At least this is true for the language professional.

6. Summary

By investigating the relationship of machine translation systems (MT) and translation memory systems (TMS) in the context of (technical) translations by language professionals, we first had to redefine the role of machine translation. The MT's main application lies within information translation and consists in helping non-language specialists to overcome language barriers. MT tries to simulate the skills of a translator and this is sufficiently successful for certain application fields in the mass market.

On the other hand, the professional translator does not need an "less skilled" electronic colleague, but reliable professional software helping to do a professional job. Specialised software for translators today is TMS, which takes over the parts in the translation process that can be successfully delegated to a machine. Therefore, it is evident that the role of MT in the context of professional translations has to be redefined as an optional "add-on" tool within the TMS. If certain conditions prevail, MT can speed up the editing process. Preconditions are: a seamless integration - preferably an interactive integration - and sufficiently powerful links of the term bank system to the MT.

Successful translation process automation as one part of an overall document production flow means for the future a better harmonisation of the involved technical solutions. We are faced with the problem to integrate authoring tools, document retrieval software, workflow solutions, translation memory tools and last but not least machine translation. A key role for satisfying solutions will be the "interconnectivity of software modules" combinable to holistic solutions.

Tendencies in information science towards distributed objects and in general towards object orientation are very important in this respect.