Translation Memory Database in the Translation Process

Sanja Seljan, Ph.D.
Faculty of Humanities and Social Sciences, Department of Information Sciences
Ivana Lučića 3, 10 000 Zagreb, Croatia, tel/fax: +385 1 600 24 31
sseljan@ffzg.hr

Damir Pavuna, M.Sc.
Integra d.o.o., A. Stipančića 18, 10 000 Zagreb, Croatia, tel:+385 1 38 33 447
damir.pavuna@integra.hr

Abstract*. Translation memory (TM) today widely used Computer-Assisted Translation (CAT) tool, based on matching of source and target language segments is intended for future reuse of already translated domain restricted type of texts. It is consistent, fast and cost-saving but the biggest lacks are non-existence of language knowledge and context insensitivity. Often integrated with machine translation (MT) software, TMs have been implemented in various multinational companies and EU.*

The neural network approach is judged and confronted to the knowledge based approach. Experience of TMs use in Croatian organizations is presented and suggestions for further integration into document processing workflow given.

Keywords. Translation Memory (TM), database, Computer-Assisted Translation (CAT) Tools, translation unit, cost-saving, speed, consistency, terminology, alignment, languages.

1. Introduction

Everyday need for quick access to information and documentation, translation into other languages, increasingly short time of product life and documentation production, translation needs in business, education, science and culture have caused use Computer-Assisted Translation (CAT) tools in the translation process.

Translation memory (TM), in its simplest form a database, is today widely used Computer-Assisted Translation (CAT) tool intended for future reuse of already translated texts. Based on matching source and target language segments, TM does not translate anything by itself and differs from the Machine Translation (MT) software.

In the TM new segments are compared to the database content, and the resulting output (exact, fuzzy or no match) is reviewed and finalised by the translator.

Primary reasons for implementing TM technology are consistency, speed and cost-saving, but under condition to be used on certain type of text of considerable volume. It is consistent, fast but the biggest lacks of TMs are non-existence of knowledge of the language to be translated and context insensitivity.

In spite of this TMs have been implemented in various multinational companies and EU, often with MT software and other CAT tools, where the translation process is integrated into document processing workflow.

In the paper the experience in TM use is presented in translation to Croatian on the text of the restricted domain, with respect to 10 years experience and ever growing TM. Necessity of using some kind of language knowledge is discussed and solutions are proposed.

2. Translation demands in EU

Translation demands in EU surpass human capacities from several sides (huge number of pages to be translated every day, 20 official languages, need for consistency, for data-sharing, extremely short deadlines, insufficient number of translators, considerable volumes of text, translation costs, specialized training courses, requirements for every new member state, etc. ref [18] this is the area where systematic use of CAT and MT tools could facilitate written communication by offering fast and cost-effective raw translations.

With the last enlargement of EU consisting of 25 member countries and 20 languages (Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Slovak, Slovene, Spanish, Swedish), the complexity of translation goes up to 20 x 19 = 380 language combinations. With such an overload of translation work, use of CAT tools, among which translation memories (TMs), has been recommended by Directorate-General for Translation of the European Commission.

Need to overcome language barriers that exist not only in EU, but also in national government institutions, agencies, scientific, business and educational institutions, in industry, multinational

associations, etc. show that use of technology in the translation process of the domain specialized text have become necessary

3. Integrated TMs

The main body responsible for translation and written communication, Directorate-General for Translation (DGT) of the European Commission (EC) employs more than 2000 professionals has recommended use of CAT/MT tools and methods in order to cope with overload of translation work:

a) Administration and document tools – for electronic submission translation requests and integration into electronic archiving system
b) Use of CAT tools such as:
   - Terminology tools
   - TM at two levels: local and central
   - Machine translation
c) Voice recognition

According del Pino [4], six institutions and bodies of the EU (Council, Commission, Court of Auditors, Economic and Social Committee, Committee of the Regions, Translation Centre) have published a tender in 1996 for the acquisition of a software package providing "integrated translation support tool", i.e. translation memory software integrated with word-processing programs, with text alignment utility and a terminology management program.

For the reasons of short deadlines, expenses, large volumes and need for consistency EU has integrated all modules of the translation process and had added the translation memory into EURAMIS system. This type of integration into "Translation Workstation" discovers the accelerated need for integration of communication tools, especially translation tools into organisational workflow.

The acronym EURAMIS, standing for European Advanced Multilingual Information System used in the Commission's Translation Service is e-mail base client-server application that gives access to a number of language tools and services and is based on the following principles [2]:

- Storage of linguistic resources of general interest in one place (Linguistic Resources Database, LRD) in order to make them available to all users in the translation service
- One internal format because of easier exchange of results
- Mass treatment of linguistic data on a server by modular programs

Integration of machine translation software (SYSTRAN) with translation memory (TWB – TRADOS) combines both types of advantages depending on the users' preferences and on the type of document to be translated, while post-editing can be done in a Word format.

Automatic lookup of the terminology in Systran goes first to very powerful morphological analyser that is used as a lemmatiser and the result of this analysis goes to EURODICAUTOM, the central terminology database which is redesigned to a modern relational database management system. Retrieved terminology can be also used in TWB environment using MULTITERM terminology base.

The best results are obtained for the language pairs that combine TM (TWB) and Systran [19] such as En-Fr, En-Es, En-It, Fr-En, Fr-Es, Fr-It, Es-En, Es-Fr, with the main idea of using MT system Systran when there is no satisfactory proposition from the TM (TWB) database.

4. Translation Memory

In the TM the efficient translation is achieved through three main functions:

MAT tool breaks texts into segments (sentences or sentence fragments) and presents the segments in a convenient way, to make translating easier and faster. In some tools each segment is presented in a special box and the user can accept the offered translation, modify it or enter the new one.

- Translation of each segment is saved together with the source text. Source text and translation are treated as translation units (TUs). One can return to a segment at any time to check the translation. There are special functions which help navigating through the text and finding segments which need to be translated or revised (quality control).
- The main function of a MAT tool is to save the translation unit in a database, called TM which can be re-used for any other text or even in the same text. Through special "fuzzy search" features the search functions of MAT tools can also find segments which do not match 100%. This saves time and effort and helps the translator to use terminology more consistently.
- Along with storing of aligned segments of source and target languages, translation memory grows, and the result of the segment to be translated can be offered in the form of exact match, fuzzy match or when no match is found.

4.1. Reasons to use

Need for consistent use of terminology, data-sharing of common resources, re-use of already translated and revised text suggest used of Translation Memory, in its simplest form a database.

TM is organised as a database, where each source language segment is stored together with target language segment, which is called "translation unit".

TMs can be organised for numerous language pairs and most often are integrated with terminology bases, alignment system and word-processor, but also with machine translation systems, as in EU translation workflow. TMs are intended to augment efficiency, consistency, reduce costs, enable data-sharing, and not to replace the human.

4.2. Output

TM does not translate anything by itself and differs from the MT software. In the TM new segments are compared to the database content and any new sentence that has to be translated is firstly searched in the database and the matched value is calculated (it can be defined by the user). The resulting output (exact, fuzzy or no match) is reviewed and finalised by the translator. If the mach value is 100% (i.e. exact mach) then the translation is inserted from the database into the text. If the match value is below 100% but above the defined percentage (i.e. "fuzzy mach") the user can accept the offered translation and correct it. Language segments whose values are below the defined margin have to be translated manually and then new proposals will be stored for future use.

In the case of already existing source and target language texts, the text can be aligned on the level of word, phrase, clause, sentence or even paragraph and then integrated into TM dbase. In the TM system, text information system is separated form the file formatting code, so the original page layout can be saved.

4.3. Type of text

Primary reasons for implementing TM technology are consistency, speed and cost-saving, but under certain conditions. Besides the basic condition for the source text to be in electronic form, and cleaned from miss spellings, only certain types of texts are suitable for translation using TM. There are several basic conditions all of whom should be confirmed in order to proceed with TM translation:

- Repetitive type of documentation of the same domain that tends to appear regularly or at least several times, so the old versions could be reused
- Larger volumes of text (e.g. manuals, decisions, technology reports, catalogues, etc.)
- Style of the text having many repetitions (terms, phrases or sentences repeat and could be re-used in new translation of the same domain) containing specialized vocabulary, short and simply structured sentences with not many pronouns and adjectives, standardized terminology, universal tables and graphics
- Need for consistency of terminology
- Short deadlines
- When professional classic type of translation takes too long

4.4. Format

There are different practical solutions offered by different companies, such as "Translator's Workbench" by Trados used together with the corresponding terminology management solution "MultiTerm", then "Transit" TM system by Star used with "Termstar", then "Déjà vu" by Atril, "TM2" by IBM, "XL8" by Globalware, among which Trados merged with SDL into SDLX currently has competitive advantage.

In order to allow better interoperability between TM systems, the special group of the LISA (Localisation Industry Standards Association) named OSCAR group (Open Standards for Container/ Content Allowing Re-use) has defined standards for the development and maintenance of open standards [15] for the language industry which include:

- TMX (Translation Memory eXchange) format – the standard for the exchange of translation memory data
- TBX – the standard for exchange of structured terminological data
- SRX – the standard for exchange of information about how translation tools segment text

TMX is a certifiable standard format (since 1998), created to allow easier exchange of TM data between tools and/or translation vendors with little or no loss of critical data during the process. TMX provides a standard XML format for the representation of Translation Memory (TM) data. According to OSCAR research group, TMs database represent strategic corporate assets, of worth for international business and TMX is actually a way to protects them against market and technology changes. It also indicates that although TMX is free and open format, only software products that have been certified as conforming to TMX by LISA's testing lab are licensed to carry TMX logo.

5. Pro and con of TMs

As translation work done by TM is much faster, the resulting profitability of the TM system is measured through lower costs, shorter time, terminology consistency and saved layout of the page, and one of the key advantages is that translation units in TM can be organised as pairs among number of languages.

Despite of mentioned savings, the hidden costs should be also mentioned, such as costs of software, maintenance, education for work with TM system, building of glossaries and TM revision. The biggest lacks of TMs are non-existence of knowledge of the language to be translated and context insensitivity.

Among obstacles in use of TMs, several ones could appear [10]:

- It is possible that TM tools do not fit into existing translation or localization process regarding approval of translation changes
- Customization of the TM systems and sufficient training
- Significant investment (purchase of software, importing of past translations into TM database (i.e. alignment process), training, if necessary additional terminology tools, maintenance costs

(upgrading with memory, fast network card, hard disk)
- Protection of TM investment by developing proper strategy for maintaining TM database (data on frequency of updated, regular distribution, backup), ownership of the TM (regulated by agreement), confidentiality, support of the TMX format
- Legalities – whether the intellectual property of the TM belongs to freelancer, vendor or end-client

6. TMs in organizational workflow

When considering introduction of large-scale of TM software, even if the organisation is entirely based on translation experience and organisationally independent, according del Pinto [4] the following should be taken into consideration:

After considering mentioned demands for type of text (volume, repetitiveness, consistency, cost-effectiveness, time), the next stage would be building of TM database from the existing data after considering the following questions:

- Use of existing data – are they usable with a reasonable degree of effort, particularly whether original text and translation exist in electronic form and in the same format. If not, is scanning of some key documents a reasonable alternative and if yes does the translation organisation have a convenient system for retrieval of all necessary documents?
- Can alignment be lightly edited or must be of perfect quality? How much time does it take or is it better to subcontract it?
- If suitable data do not exist or are not worth converting, will translators accept working with TMs without any significant benefits at the beginning and will TM be result of their interactive work within reasonable time?

According to our experience another very important fact is the text repetition. The organisation which has a lot of similar translations can seriously consider buying TM product. For instance technical manuals or agreement revisions are ideal for TM use because every new document has many same or slightly changed sentences and savings are significant. On the other hand TM has noting to do with translating classical literature. It is very important because no translator is going to make this big effort to implement TM in his work if there is no evident savings in his work to be done.

When working with TM the following should be considered, the main question is a way to organise preparation of interactive work, which might include:

- Decision on the most suitable TM, if there are several of them and does the TM contain necessary material (previous versions, related documents, etc.)
- How to organise data-sharing of translation project among several translators during interactive work (simultaneously, updating TM and approved changes)
- Scanning the document for references pointing to extra documents
- Can translators cope with augmented on-screen information flow?
- As integration of revisions into the TM workflow is of special importance for the translation in question and for next ones, the following questions should be considered:
- Will reviser update TM used by translator asking him for permission
- Who takes responsibility for the final version of the TM and of the document
- Necessity to include labels when changes in the document or TM are made (person, date, document – most TM tools have all these data in the structure of every change made in TM database – they have the trace of changes)
- How is the backup performed (it is very important to have periodical backup and labelled TM history)

Our experience in doing revisions it is that revised TM becomes Read Only TM and every new translation must pass revision to become Read Only TM that has the permission to be used. For every language there is one translator that is in charge for TM and does the revisions and approves the TM to become read only.

7. TMs in Croatian organizations

Unfortunately there is no reliable information about the use of TM and CAT tools in Croatia. Implementation of CAT tools requires organizational, educational and professional changes, so not many companies have implemented it jet. According to information that Integra d.o.o. has from their long experience in the field, and as Trados exclusive dealer for Croatia, there are no more then 10 companies that have more then one TM workstation and 200 freelance translators that use TM to help them producing translations. MAT tools are rather often in use but mostly as a little bit more intelligent dictionaries and not as tool that produces translations!

According to our experience in using CAT tools (more the 10 years of using CAT tools, mostly Trados) especially TM, this is a very useful tool that can really drop down your costs, increase your productivity and quality of translation. When talking about technical manuals and their translation (e.g. localisation of IT industry manuals) our analysis have shown us that even with a new field (e.g. media products) we have very fast - after 3. updating of products reached 30% of translation saving. On the other hand with our old customers we can, on some very similar products, reach 50% of savings! This seems to be some kind of TM threshold in savings.

At the moment we are considering steps to use TM tools with some kind of knowledge about Croatian language, and we hope that it can increase our translation savings, productivity and quality. The biggest lack is that TM does not have any knowledge about the language to be translated. TM is neither sensitive to context problems nor to cultural or symbolic differences of language pairs. Necessity of using some kind of language knowledge is obvious.

8. Conclusion

Although TM saves time during translation by using existing translations, it also creates some new tasks (management, revision, preparation, post-editing). It the document is translated into several languages, some additional gains can arise. TM brings also additional demands regarding protection of copyright, property of TM and additional expenses when requiring additional linguistic resources, as well as significant investment of in-house training. One of important benefits of TM could be complementarities among translators who could benefit from the previous work of their colleagues, but knowing that the price of translation with TM differs from the professional translation. As the language changes, TM could be also considered as a linguistic archive.

Introduction of TM into organisational workflow imposed numerous questions, such as: Has the introduction of TM ensured kind of balance between data-sharing and confidence of data, between job sharing, data quality and heavy management. Are acquired data in the TM valuable for future translation projects? Is TM database valuable for new translation tasks and how is this new type of workflow accepted among translators? etc.

Although there are potential risks and inconveniences, TM and MT represent different types of CAT tools that will augment efficiency (consistency, data-sharing) and bring some savings (cost, time, effort). Although TM can be seen as a pure database intended for archiving of language pairs, it can be enriched by linguistic component related to the proper language.

TM does respond to another aspect of the human intelligence: its capacity and pragmatism integrating technology and language for the purpose of automation and computerisation of the translation process, its integration in the translation workstation and organisational workflow with the main aim: tool for realisation positive human needs.

References:

[1] Andersen, Poul. Translation Tools for the CEEC Candidates for EU Membership – an Overview.http://ec.europa.eu/comm/translation/reading/articles/pdf/1998_01_tt_andersen.pdf [10/10/2005]

[2] Blatt, Achim. EURAMIS: Added Value by Integration. Terminologie et traduction 1, pp. 59-73, 1998.

[3] Cavalli-Sforza, Violetta ; Brown, Ralf; Carbonell, Jaime; Jansen, Peter J; Kim, Jae D. Challenges in Using an Example-Based MT System for a Transnational Digital Government Project. Proceedings of EAMT, 2004.

[4] Del Pino, Santiago. Using Translation Memory Software (TMS): An Organisational Checklist. Terminologie et Traduction, pp. 132-139; 1998.

[5] Directorate-General for Translation of the European Commission. Translation Tools and Workflow; 2005. http://europa.eu.int/comm/dgs/translation/bookshelf,/toolsandworkflowen.pdf [10/1/2006]

[6] Directorate-General for Translation of the European Commission. Translation Tools and Workflow; 2005. http://europa.eu.int/comm/dgs/ translation/bookshelf,/toolsandworkflowen.pdf [10/1/2006]

[7] Directorate-General for Translation of the European Commission. Translating for a Multilingual Community; 2005. http://europa.eu.int/comm/dgs/translation/bookshelf/brochure_en.pdf [10/1/2006]

[8] Dovedan, Zdravko; Seljan, Sanja; Vučković Kristina. Strojno prevođenje u procesu komunikacije, pp. 283-291. Informatologia 35 (4); 2002.

[9] EAGLES Evaluation Working Group. Benchmarking translation memories. http://www.issco.unige.ch/ewg95/node157.html#SECTION001043000000000000000

[10] Heuberger, Andres. What you Need to Know about Translation Memories? Multilingual webmaster.com http://www.multilingualwebmaster.com/library/trmemories.html [12/2/2006]

[11] Hodasz, Gabor; Grobler, Tamas; Kis, Balazs. Translation Memory as a Robust Example-based Translation System. Proceedings of EAMT, 2004.

[12] Hutchins, John. The State of Machine Translation in Europe and Future Prospects. HLT Central; 2002. http://ccl.pku.edu.cn/doubtfire/NLP/Machine_Translation/Overview/Article%20-%20MT_John_Hutchins.htm [20/10/2005]

[13] Hutchins, John. Towards a new vision for MT. Proceedings of MT Summit; 2001 Sept 18-22; Santiago de Compostela,

Spain. http://www.translationdirectory.com /article402.htm [15/10/2005]

[14] Hutchins, John. Computer-based translation tools, terminology and documentation workflow: report from recent EAMT workshops. 4th Infoterm Symposium "Terminology work and knowledge transfer", Vienna, 1998.

[15] LISA (Localisation Industry Standards Association): OSCAR (Open Standards for Container/ Content Allowing Re-use) http://www.lisa.org/sigs/oscar/ [12/2/2006]

[16] Loffler-Laurian, Anne-Maire. La traduction automatique. Villeneuve d'Ascq: Presse Universitaire du Septentrion; 1996.

[17] Petrits, Angeliki. EC Systran: The Commission's Machine Translation System; 2001. http://europa.eu.int/ comm/translation/reading/articles/pdf/2001 _mt_mtfullen.pdf [15/1/2006]

[18] Seljan, Sanja; Pavuna, Damir. Why Machine-Assisted Translation (MAT) Tools for Croatian? Proceedings of ITI International Conference, pp. 469-474. Cavtat, 2006.

[19] Ulrich, Heidi. La mise en place du Translator's Workbench (TWB): Concurrence avec SYSTRAN et élément humain. Terminologie et traduction 1, pp. 102-116; 1998.

[20] Zetzsche, Jost. Translation Memories: The Discovery of Assets: Recognizing opportunities and overcoming obstacles to TM sharing. MultiLingual Computing & Technology 72, vol. 16 (4), pp. 43-45; 2005.