Using Translation Memory to Speed up Translation Process

Marija Brkić
Department of Informatics, University of Rijeka
Omladinska 14, 51000 Rijeka, Croatia
mbrkic@uniri.hr

Sanja Seljan
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
sanja.seljan@ffzg.hr

Božena Bašić Mikulić
Primary school "Turnić"
Franje Čandeka 20, 51000 Rijeka, Croatia
bozena.basic@gmail.com

Summary

Translation process is one aspect of human creativity. Due to globalization, EU accession negotiations, and the need for information exchange, the amount of translation work increases on a daily basis. The translation process is hindered by the fact that the languages involved differ culturally, stylistically, syntactically and lexically. This paper explores the benefits and limitations of TMs (translation memories). TMs are not used for replacing humans in the translation process, but rather for enhancing the human translation process. In this paper, a detailed analysis of Atril's Déjà Vu X system is presented, along with its time-saving implications, which are based on the reuse of previously stored segments. Excerpts from three different digital camera user manuals are translated from English into Croatian. Evaluation is performed by measuring the time difference between human and TM-based translation speeds in preparation, translation, and revision phases, and with regard to six different parameters.

Key words: Translation Memory (TM), Déjà Vu, Computer-Assisted Translation (CAT), language pair, translation unit (TU), translation speed

Introduction

The EU relies on the principles of open access to documents, multilingualism and democracy. Therefore, the EU legislation needs to be translated into each of the official languages. On the other hand, the legislation of each particular member state needs to be translated into one of the EU’s official languages (Seljan & Pavuna, 2006b). To sum up, translation demands in the EU surpass human capacities. There are 23 official languages with 23 x 22 = 506 language pairs. A huge number of pages need to be translated on a daily basis. Short deadlines, demands for consistency and data-sharing, and insufficient number of translators further impede the translation process, particularly for newly admitted countries (Seljan & Pavuna, 2006a). Nowadays, fortunately, translators have fully automatic MT systems and CAT tools at their disposal (Valderrábanos, 2003). Moreover, the usage of such tools has been recommended by the Directorate General for Translation of the European Commission (DGT), which is European Commission’s in-house translation department (Seljan & Pavuna, 2006a). Depending on their needs, translators can opt for MT or CAT tools.

MT was first conceived as a technology that significantly speeds up the translation process and offers human-like quality translations. Soon, it became clear that such a goal is far-fetched (Valderrábanos, 2003), which led to the development of TM technology. Current computational models of MT are limited to tasks for which rough translations are adequate, tasks where human post-editors are used and tasks limited to sub-domains in which fully automatic high quality translations are achievable (FAHQT) (Jurafsky & Martin, 2009). TMs, on the other hand, exploit machine memory for storing translated segments in order to reuse them in future translations. Their usability, therefore, increases with the size of the stored data.

As Croatia’s EU accession negotiations are underway, it is high time for the development of Croatian language tools and resources. This paper explores the benefits of using TMs, in particular Atril’s Déjà Vu X (DVX) system, and presents the results of a study in which English and Croatian are source and target languages, respectively.

Translation memory

TM technology is based on the notion of reuse of previously translated segments. It is usually integrated into a system which has a terminology management module and a lexicon (Valderrábanos, 2003). This technology does not aim at replacing humans, but rather at enhancing the human translation process. A TM can be defined as a database which stores corresponding source and target language translations, called translation units (TUs).

Approaches

There are two TM implementation approaches. Despite the differences in implementation, TMs are designed with the common purpose of storing previously translated material in an organized way, in order to present it to the user in future translations (Gow, 2003).

Sentence-based approach

A sentence-based approach divides source and target language texts into corresponding TUs, which can be sentences, titles, subtitles or list entries. TUs are stored in a database and retrieved in future translations in cases of identical or similar TUs in new source texts. Sentence units are easy to identify if they start with capital letters and end with full stops (Gow, 2003). However, abbreviations or full stops which are not at the end of sentences pose problems. These problems can be solved by defining new sentence delimitation rules (Déjà Vu, 2009). The main benefit of the sentence-based approach, compared to a character-string-in-bitext-based approach, is that exact matches are more likely to be relevant because sentence-based TMs represent an extreme form of high precision, low recall search (Simard & Langlais, 2000). Fuzzy matching algorithms, on the other hand, are based on statistical models of similarity. Since these models are only loose approximations, the matching algorithms sometimes create useless matches, known as 'noise', or fail to generate matches, the phenomenon known as 'silence' (Bowker, 2002 in Gow, 2003).

Character-string-in-bitext(CSB)-based approach

A CSB-based approach involves storing of source texts and corresponding translations in a database. The resulting texts are called bitexts. Bitexts can be used for preparatory background reading. In this approach, identical character strings of any length are recognized and reused (Gow, 2003). Working with sentence segments, instead of entire sentences, has its advantages (Simard & Langlais, 2000). It enables identification of identical sentence segments or even several consecutive identical sentences at once (Macklovitch & Russell, 2000 in Gow, 2003). 'Noise' phenomenon is still present, but this time as a result of finding unreliably small matches. 'Silence', on the other hand, occurs because there is no support for fuzzy matching. One of the disadvantages of this approach is that internal repetitions have to be recycled exclusively through terminology databases, because only entire translations are added to databases (Gow, 2003).

Advantages

Two major advantages of TM technology are consistency and speed. Consistency is of crucial importance in non-literary texts, for example software and hardware manuals (Valderrábanos, 2003), or business, legal, scientific and technical texts (Gow, 2003). These texts are highly repetitive. The longer they are, the more likely they are to contain repetitive content (Austermühl, 2001 in Gow, 2003). Repetition can occur, not only internally, but also across several texts in the same domain (Gow, 2003). Moreover, software documentation, besides being highly repetitive, is subject to frequent version updates. It is thus an ideal candidate for exploiting TM benefits (Bruckner, 2001). Consistency is especially important in cases where several translators work on the same project and share the same TM on the network (O’Brien, 1998 in Gow, 2003).

Speed is important, regardless of the domain, because globalization has brought forward endless translation demands (Valderrábanos, 2003). Using TMs in the translation process implies cost reduction. For example, the translation process can be started as soon as the first draft of the document to be translated is obtained. Furthermore, translation vendors can lower prices and thus earn more contracts (Gordon, 1997 in Gow, 2003). On the other hand, freelance translators can save up their valuable time or increase their earnings by increasing the translation speed (Gordon, 1996 in Gow, 2003).

In addition, using TMs preserves original page layouts in translated documents, because formatting information is hidden in embedded codes (Seljan & Pavuna, 2006a).

Finally, TMs are usually integrated into systems which have tools for building dictionaries (Webb, 1998) and reporting detailed statistics on internal and external repetition and word counts. These tools can help project managers in scheduling localization products (Esselink, 2000).

Disadvantages

Although TMs bring a lot of advantages, there are also some limitations of their usage.

First of all, using TMs implies initial decrease in productivity because translators need to master the environment (Webb, 1998). The odds of finding quality matches increase with the size of TMs (Gow, 2003). Moreover, the beneficiary effects of using TMs are felt only on repetitive texts (Valderrábanos, 2003). Therefore, cost-effects of investing additional time into importing existent translations through the process of alignment should be calculated (Seljan & Pavuna, 2006a).

Although, according to Esselink (2000), TMs indisputably save time, regular database maintenance is time-consuming (Austermühle, 2001 in Gow, 2003).

Furthermore, source texts need to be in digital form (Gow, 2003) and suitable file formats because not all formats are supported since TM systems require filters to preserve formatting. As an effect, they are usually bundled only with filters for most commonly used formats (Esselink, 2000).

TMs affect quality of the translation because, by using them, translators tend to avoid using anaphoric or cataphoric references in order to make segments more ‘universal’.

Seljan and Pavuna (2006a) add lack of language knowledge and context insensitivity to the list of drawbacks and point out additional software, maintenance and education costs.

There are also other concerns with regard to using TMs. For example, it is questionable whether translators should be paid differently for identical and fuzzy matches recovered from TMs. Furthermore, the ownership of final TMs is also unclear.

Déjà Vu X

DVX is a very powerful and adaptive CAT system which integrates several CAT tools – lexicon, terminology database, TM, alignment module, etc. The first version of this system appeared in 1993. DVX is an example of a hybrid approach. Matches are ranked into the following categories: perfect/exact match, fuzzy match, guaranteed match (there is overlapping of neighbouring TUs as well) and assembled from portions match. It features auto-search, assembling, propagation and pre-translation. Besides, it gives a detailed statistical analysis of source and target language texts. Terminology database is the only database that needs to be manually filled and it allows linguistic enrichment of inserted terms. A list of lexicon entries is automatically built. However, entries which are to be included need to be manually translated. DVX combines modern TM technology with example-based machine translation (EBMT). EBMT implies translation by analogy and enables combining several segments into one translation segment (Déjà Vu, 2009).

According to a survey, which included 699 translators from 50 different countries, DVX was the second TM on the list of popularity, meaning that 61% of translators were acquainted with the system. On the usage list, it was the fourth. To be precise, there were 23% of translators using the system (Laugodaki, 2006). The statistics show that the system is preferred by translators with higher level of information literacy. DVX scored better than competitors' systems in functionality, efficiency, speed, reliability, price, and usability. According to the survey, it also had better customer support.

Experimental study

The feature to be examined in this study was the speed of the translation process, with the goal of measuring the time difference between human and TM-based translation speeds (Bruckner, 2001). The following parameters were taken into account:

- Measure (minutes),
- Evaluation procedure (comparing times needed by each translator to deliver translation),
- Score (time needed for the translation process),
- Metric (faster / slower),
- Languages (English – source language, Croatian – target language),
- Text type (hardware documentation), and
- System used (Déjà Vu).

The study was conducted by two expert translators, who translated three excerpts from Kodak's, Nokia's and Canon's digital camera user manuals, respectively. Each excerpt was two standard pages long and contained information on battery and relating equipment care and maintenance. Therefore, each excerpt contained terms and phrases from the same domain. The excerpts were translated from English into Croatian.

The experimental study was conducted in the following phases:

1. Text selection phase
2. Preparation phase (lexical analysis of the source language texts and their technical registers)
3. Translation phase (setting up environment, translating, building up lexicon, filling terminology database)
4. Revision phase (post-editing)

Preparation phase

For all the excerpts, the preparation phase was performed jointly by the two translators. The lexical analysis of the source language texts (English) lasted 45, 10 and 18 minutes, respectively.

Translation phase

Prior to the translation process, the translator using the TM spent 5 minutes setting up the environment. The translator had some previous experience with DVX translation memory. After the text was inserted into DVX, the system calculated that the internal repetition was 6 per cent.

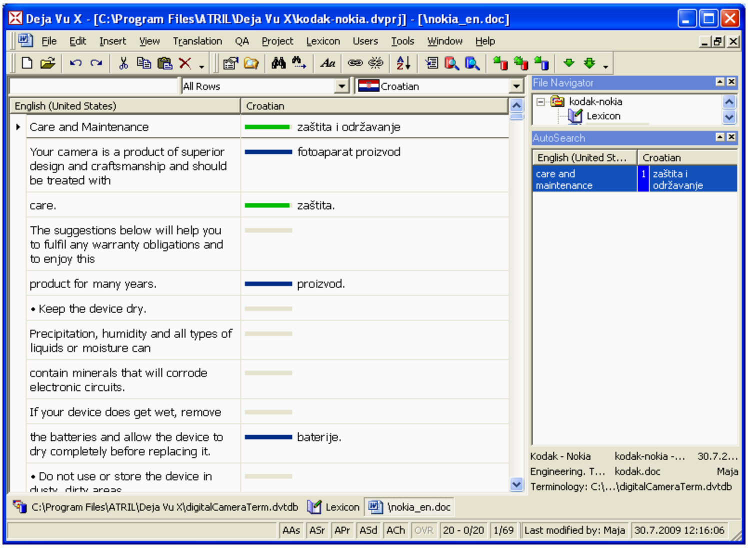
The translation speed in the first translation (Kodak) was identical for both translators (37 minutes).

After the first text was translated, the TM translator had to manually build up the system's lexicon, which lasted 15 minutes and included 68 entries. The inserted entries were mostly nouns in nominative singular and plural, and masculine adjectives in singular. Filling the terminology database lasted 10 minutes. Only 18 phrases were added due to Croatian rich morphological system.

The second excerpt (Nokia) first underwent the pre-translation process. Besides the exact matches, fuzzy matches and parts assembled from portions were also allowed. After processing 69 source language sentences, the system found 6 sentences with 1 or more fuzzy matches and 31 sentences assembled from portions, some of which are presented in Figure 1.

The traditional translation process for the second text lasted 31 minutes, while, with the help of the TM, it lasted 30 minutes. After the second text translation, 111 new entries were added to the lexicon and 13 new phrases were added to the terminology database. These processes lasted 15 and 10 minutes, respectively.

The third excerpt (Canon) also underwent the process of pre-translation prior to the translation. The system processed 41 source sentences and found 1 sentence with a unique exact match and 32 sentences assembled from portions. The system found substantially more matches than it had in the second text (Figure 2).


Deja Vu X - [C:\Program Files\ATRIL\Deja Vu X\kodak-nokia.dvprj] - [\nokia_en.doc]
File Edit Insert View Translation QA Project Lexicon Users Tools Window Help
All Rows
Croatian
File Navigator
kodak-nokia
Lexicon
AutoSearch
C:\Program Files\ATRIL\Deja Vu X\digitalCameraTerm.dvtdb
Lexicon
\nokia_en.doc
AAs ASr APr ASd ACh OVR 20 - 0/20 1/69

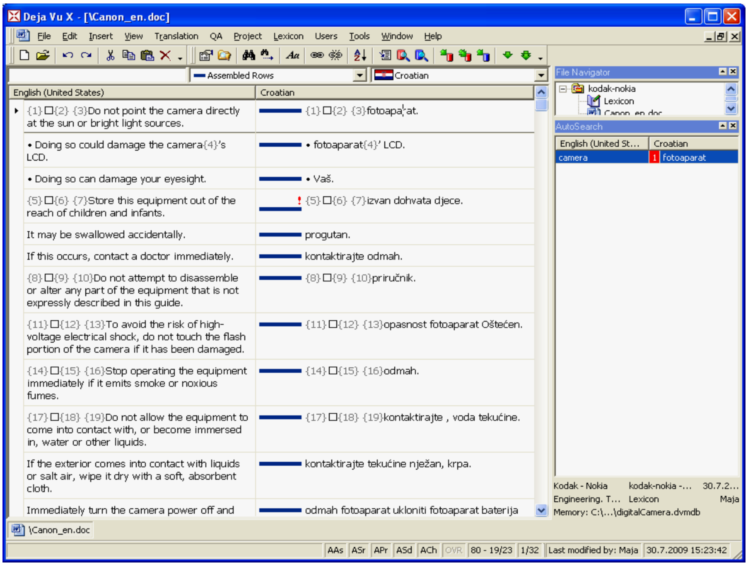
English (United States)	Croatian
Care and Maintenance	zaštita i održavanje
Your camera is a product of superior design and craftsmanship and should be treated with	fotoaparat proizvod
care.	zaštita.
The suggestions below will help you to fulfil any warranty obligations and to enjoy this	
product for many years.	proizvod.
• Keep the device dry.	
Precipitation, humidity and all types of liquids or moisture can	
contain minerals that will corrode electronic circuits.	
If your device does get wet, remove	
the batteries and allow the device to dry completely before replacing it.	baterije.
• Do not use or store the device in dusty, dirty areas	

English (United St…		Croatian
care and maintenance	1	zaštita i održavanje

Kodak - Nokia kodak-nokia -… 30.7.2…
Engineering. T… kodak.doc Maja
Terminology: C:\…\digitalCameraTerm.dvtdb

Last modified by: Maja 30.7.2009 12:16:06

Figure 1: Matches found by DVX when the second source text was inserted


Deja Vu X - [\Canon_en.doc]
File Edit Insert View Translation QA Project Lexicon Users Tools Window Help
Assembled Rows
Croatian
File Navigator
kodak-nokia
Lexicon
Canon_en.doc
AutoSearch
\Canon_en.doc
AAs ASr APr ASd ACh OVR 80 - 19/23 1/32

English (United States)	Croatian
{1}□{2} {3}Do not point the camera directly at the sun or bright light sources.	{1}□{2} {3}fotoapa¦at.
• Doing so could damage the camera{4}'s LCD.	• fotoaparat{4}' LCD.
• Doing so can damage your eyesight.	• Vaš.
{5}□{6} {7}Store this equipment out of the reach of children and infants.	! {5}□{6} {7}izvan dohvata djece.
It may be swallowed accidentally.	progutan.
If this occurs, contact a doctor immediately.	kontaktirajte odmah.
{8}□{9} {10}Do not attempt to disassemble or alter any part of the equipment that is not expressly described in this guide.	{8}□{9} {10}priručnik.
{11}□{12} {13}To avoid the risk of high-voltage electrical shock, do not touch the flash portion of the camera if it has been damaged.	{11}□{12} {13}opasnost fotoaparat Oštećen.
{14}□{15} {16}Stop operating the equipment immediately if it emits smoke or noxious fumes.	{14}□{15} {16}odmah.
{17}□{18} {19}Do not allow the equipment to come into contact with, or become immersed in, water or other liquids.	{17}□{18} {19}kontaktirajte , voda tekućine.
If the exterior comes into contact with liquids or salt air, wipe it dry with a soft, absorbent cloth.	kontaktirajte tekućine nježan, krpa.
Immediately turn the camera power off and	odmah fotoaparat ukloniti fotoaparat baterija

English (United St…		Croatian
camera	1	fotoaparat

Kodak - Nokia kodak-nokia -… 30.7.2…
Engineering. T… Lexicon Maja
Memory: C:\…\digitalCamera.dvmdb

Last modified by: Maja 30.7.2009 15:23:42

Figure 2: Matches found by DVX after the third source text was inserted

The translation phase for the third text lasted 29 minutes for the traditional translation and 23 minutes with the TM. It took 5 minutes to add 45 new entries to the lexicon, and 3 minutes to add 7 new phrases to the terminology database.

Revision Phase

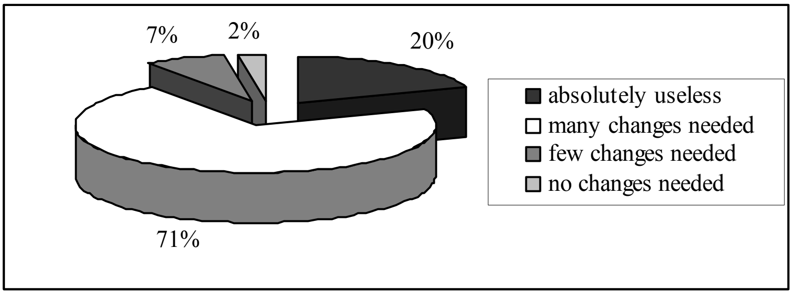
The revision phase for the three translations in the traditional translation process lasted 5, 7, and 5 minutes, respectively, while in the TM translation process it lasted 5, 8, and 5 minutes, respectively.

Evaluation

According to the automation level, evaluation methods can be divided into automatic and manual. Although time-consuming, manual methods better suit real-life system application. They can be implemented in two different ways. One way is that a translator scores each segment on a 1-4 scale from “absolutely useless” to “no changes needed”. The other way includes modifying each resulting segment into acceptable translation and counting post-editing steps needed (Hodász, 2006).

The results of a 1-4 scale test performed on the third translation are presented in Chart 1. Since the TM is still under development, these are only initial results. With the growth of the TM, the increase in the percentage of segments classified as ‘few changes needed’ or ‘no changes needed’ can be expected.

Chart 1: Effectiveness of TM



Discussion

The search process gives the highest priority to the TM matches and the lowest to the lexicon matches, with the terminology database matches in between. Nevertheless, the user is supplied with all the matches in a separate window, which enables them to choose the most appropriate match for the given context.

Most of the problems with DVX regard morphology and word order. A word extracted from the lexicon which is a masculine adjective needs to be changed into feminine or neuter in order to match the context syntactically. The same is valid for verbs, since the lexicon contains their infinitive form.

Furthermore, words are extracted in order of appearance and they often need to be rearranged because of the syntactic rules of the target language.

There are also capitalization issues which need to be resolved. For example, if there is a word which was the first word in a previously stored segment, it remains capitalized regardless of its position in subsequent occurrences.

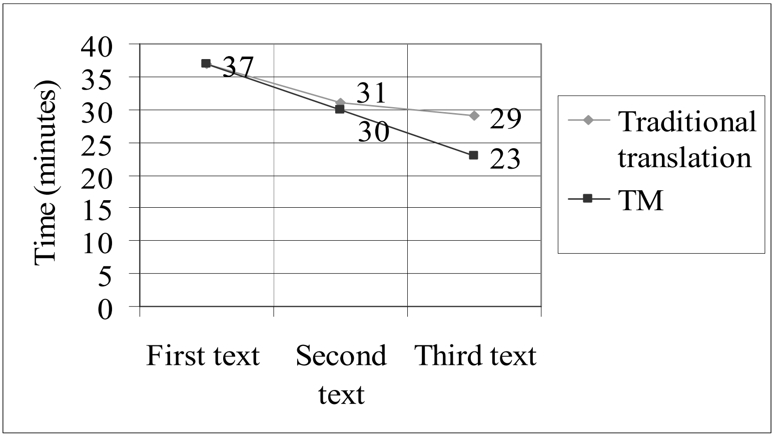
Additionally, punctuation is also retained if the word found in one of the resources is followed by a punctuation mark.
However, one of the main advantages of using the TM is the user interface, which enables the translator to see parallel sentences or units in the same row. That eases the translation process because the translator does not have to spend a lot of time inserting or deleting units of the source text and scanning through both texts.

Speed of translation process
Since the main advantage of any TM system should be speeding up the translation process, here follow the results of this case study. Table 1 presents time (expressed in minutes) needed for each phase (preparation phase has been omitted because it was done jointly by the two translators).

Table 1: Time spent for each phase

		Translation phase (minutes)	Lexicon and terminology database filling (TM) (minutes)	Revision phase (minutes)	Total (minutes)
1st text	Traditional	37	/	5	42
	TM	37	25	7	69
2nd text	Traditional	31	/	5	36
	TM	30	25	8	63
3rd text	Traditional	29	/	5	34
	TM	23	8	5	36

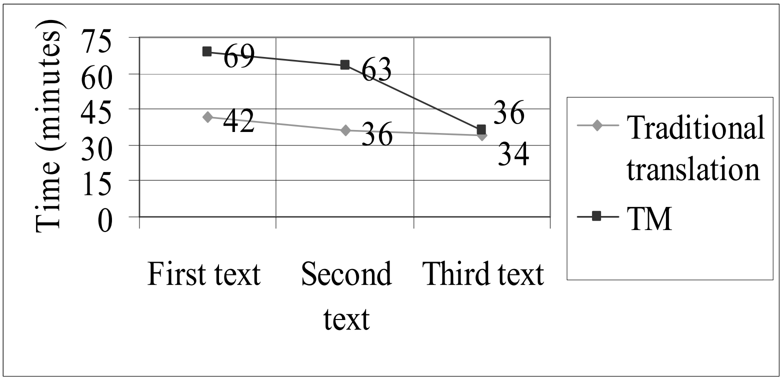
Chart 2: Time spent in translation process



It is evident that the TM translator becomes faster than the traditional translator in the second, and even more, in the third translation. Since the TM was empty prior to the first translation, the translation phase of the first excerpt was of the same length for both translators. As DVX grows in size, the TM translator becomes increasingly faster than the traditional one (Chart 2). The difference in speed is the most obvious in the third translation, where the TM translator is 6 minutes faster.

Taking into account the time which the TM translator needs to spend in the process of filling the lexicon and the terminology database, it is evident that using TM systems can be somewhat time-consuming in the beginning (Chart 3).

Chart 3: Total time for traditional translation and TM



Nevertheless, it is quite plausible that the time which the TM translator spends in filling the lexicon and terminology database gradually decreases because both databases have smaller number of entries to be added in, while the time needed for human translation remains the same. Therefore, the time difference for the TM translator and the traditional translator becomes less significant.

Conclusion

This paper presents a detailed analysis of the benefits of the Atril’s Déjà Vu X translation memory system, with its time-saving implications based on the reuse of previously stored translation units.

In this case study, segments from three different digital camera user manuals are translated, with the aim of presenting how TMs ensure consistency in non-literary texts. After measuring the time difference between human and TM-based translation speeds and taking into account 6 parameters, those being measure, evaluation procedure, score, metric, languages and text type, it can be concluded that TMs speed up the translation process, especially in later phases, i.e. when texts show certain level of local or global repetition. Nevertheless, even though TMs unquestionably save time, regular lexicon and terminology database maintenance is still time-consuming. However, this task does not take as much of translator’s time as databases grow in size. On the other hand, lack of knowledge asks for intensive work in the revision phase. Even so, TMs represent a valuable resource, especially when several translators work in the same domain, and aim to produce fast, consistent and professional-quality translations.

Acknowledgments

This work has been supported by the Ministry of Science, Education and Sports of the Republic of Croatia, under the grants No.130-1300646-0909 and No.009-0361935-0852.

References

Bruckner, Christine; Plitt, Mirko. Evaluating the Operational Benefit of Using Machine Translation Output as Translation Memory Input. // *Proceedings of the VIII MT Summit*. Spain, Santiago de Compostela, 2001

Déjà Vu – Translation memory and productivity system. http://www.atril.com/ (July 25, 2009)

Esselink, Bert. A Practical Guide to Localization. Amsterdam/Philadelphia : John Benjamins Publishing Company, 2000.

Gow, Francie. Metrics for Evaluating Translation Memory Software. M.A. thesis. University of Ottawa, Canada, 2003. http://www.chandos.ca/Metrics_for_Evaluating_Translation_Memory_Software.pdf (August 3, 2009)

Hodász, Gábor. Evaluation Methods of a Linguistically Enriched Translation Memory System. // *Proceedings of the Fifth International Conference on Language Resources and Evaluation*. Italy, Genoa, 2006, 2044-2047

Jurafsky, Daniel; Martin, James H. Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition. New Jersey : Pearson education, 2009

Lagoudaki, Elina. Translation Memories Survey 2006: Users' perceptions around TM use // *Proceedings of the ASLIB International Conference Translating & the Computer 28*. UK, London, 2006.

Seljan, Sanja; Pavuna, Damir. Translation Memory Database in the Translation Process. // *Proceedings of the 17th International Conference on Information and Intelligent Systems IIS 2006* / Aurer, B.; Bača, M. (ed.). Croatia, Varaždin : FOI, 2006a, 327-332

Seljan, Sanja; Pavuna, Damir. Why Machine-Assisted Translation (MAT) Tools for Croatian? // *Proceedings of the 28th Conference on Information Technology Interfaces – ITI 2006*. Croatia, Cavtat, 2006b, 469-474

Simard, Michel; Langlais, Philippe. Sub-sentential Exploitation of Translation Memories. // *LREC 2000 Second International Conference on Language Resources and Evaluation*. Greece, Athens, 2000

Valderrábanos, Antonio S.; Esteban, José; Iraola, Luis. TransType2 - A New Paradigm for Translation Automation. // *Proceedings of the MT Summit IX*. New Orleans, Louisiana, 2003, 498-501

Webb, Lynn E. Advantages and Disadvantages of Translation Memory : A Cost/Benefit Analysis. M.A. thesis. Monterey Institute of International Studies, Monterey, California, 1998. http://www.tradulex.org/Bibliography/Webb.htm (October 31, 1998)