Seljan, Sanja; Pavuna, Damir. Why Machine-Assisted Translation (MAT) Tools for Croatian? // Proceedings of 28th International Information Technology Interfaces Conference – ITI. Cavtat, 2006. pp. 469-475.

Why Machine-Assisted Translation (MAT) Tools for Croatian?

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. *Necessity for Machine-Assisted Translation (MAT) has become obvious in multinational companies, professional societies, in government agencies, in EU etc. Amount of text is rapidly growing and need for fast translation is increasing. Answering to needs of information society, business, education, science and culture, it is necessary to make on one side overview of MAT tools used in EU, translation needs and costs and on the other side to carry a survey of the situation in Croatia in order to proceed with further concrete activities regarding organizational, professional and educational changes, creation of language resources and joining the European projects and standards.*

Key words: MAT Tools, Translation Memory, Machine Translation, speed, consistency, Croatian, EU, standards.

1. Introduction

In the time of emerging information society, i.e. knowledge society, need for electronic accessibility of language resources has become obvious. In countries that approach the EU, necessity for Machine-Assisted Translation (MAT) has become obvious in a number of aspects: from translation needs in multinational companies, in professional societies, agencies, ministries, up to need for information search in different languages in order to lower costs, reduce time and augment speed and consistency.

The cooperation with EU imposes urgent need: translation of tens of thousands of pages from English to Croatian and vice versa in a short time, the problem of terminology consistency, localization, data-sharing and reuse of already translated texts. As the EU relies on the principles of open access to documents, multilingualism and democracy, the EU legislation should be translated to all Union's official languages and the legislation of the member state into EU official language. In 2005 about 1,3 million pages were translated on 20 official languages (Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Slovak, Slovene, Spanish, Swedish) and it is predicted that after the next enlargement it will be about 2,5 million pages a year. With insufficient number of translators and interpreters, extremely short deadlines, with average length of the document decreased from 30 to 15 pages, this is the area where machine-assisted translation (MAT) or/ and machine translation (MT) could facilitate written communication by offering fast and cost-effective raw translations. Besides electronic dictionaries, the professional software, translation memories, terminological databases and localization principles have been applied in the EU translation industry.

The present situation gives rise to several important questions: creating infrastructure, building up translation resources according EU standards, integration of Croatian modules within other language pairs, data-sharing, education for new translation jobs, cooperation between academy and industry, government support and financing, international cooperation, national priorities and strategy development. All mentioned questions speak in favor of emerging industry connecting, technology and language, science and industry, politics and culture.

In the paper special attention is given to EU demands (translation needs, costs and tools), as well as need for creation of Croatian language resources, educational and organizational changes and cooperation because of integration into European environment.

2. Why MAT Tools?

The term MAT stands for "Machine-Assisted/Aided Translation Tools" suggesting use of computer tools in everyday translation work (from word processors, online dictionaries, terminology bases, translation memories, etc.). The term TM stands for "Translation Memory", i.e. type of database helping to work more efficiently, where 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 TU 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.

The industry standard for the exchange of translation memories is TMX (Translation Memory eXchange) which has been adopted, according LISA standards, between translation suppliers, enabling importing and exporting of TMs.

3. Translation and EU

In the EU, consisting of 25 countries and 20 official languages, plus candidate countries, need for quick translation is obvious every hour of the day. Therefore, in order to compensate demands regarding capacity of translator's work (large volumes of texts, consistency, short time and quick multilingual communication).

According to EC calls for action, the EU promotes multilingualism and underlines importance of language diversity, pointing out three areas in the EU (society, economy and the Commission's own relations with EU citizens) where language knowledge is today seen as a desirable life-skill for every EU citizen.

According to Loffler-Laurian [7] the translation market at the beginning of the 1990s was as follows:

Table 1. Number of translated pages 1990s

	Human Translation	Machine Translation
Europe & the United States	300 million pages	2.5 million pages
Japan	150 million pages	3.5 million pages

In 1990s in Europe and the United States 2,5 million pages were translated through a decade, while today the same number of pages has been translated only in a year and half for needs of EU. It is expected that with next enlargement in few years 2,5 million pages will be translated in a year.

In the 1990s, it can be seen that only 6 million pages were translated by machine translation which compared to 450 million pages translated by humans' gives 1.3% of the total.

According to Angeliki Petrits [8], MT statistics over last 10 years show that number of users is steadily increasing, not only in the EC, but in other EU institutions and member states, as well as by the growing public use of MT technology on the web. In the report for EC Systran from 2001 it is indicated that in 1993. there were around 11.000 requests and in 2000. about 97.199 requests. At the beginning of 1990s demand was 2.000 pages per month and today ranges from 45-50.000 monthly.

In 2000. a total of 546.248 pages were machine translated for needs of EU (77% for the EC and the remaining 23% were shared between other EU institutions and member states).

Regarding translation costs, in 2003, before enlargement, expenditure on translation came to 1.45 €/per citizen for all EU institutions, i.e. 0.60 € (for EC only).

After enlargement, the annual translation costs set to rise to 1.78 €, i.e. 0,70€ per citizen.

Before 2004 enlargement, translation in all institutions accounted for about 0.55% of the total EU budget and about 9% of the EU administrative expenditure. After the 2004 enlargement, the translation costs are estimated to be around 0.8%, and it comes to around 13% of the administrative expenditure of the EU institutions.

4. Demand for translators

Whenever a country joins the EU, it is necessary to translate the existing EU laws - the "acquis communautaire" of around 80 000 pages at the time of the enlargement (Treaties and secondary legislation) into the language of that country. Translation of the EU law is the responsibility of the new member state government, while the Community institutions – the Council and the Commission – are responsible for finalizing, approving and publishing of the translated texts in the special edition of the Official Journal of the European Union. In addition, the legislative documentation of the new member state has to be translated into EU official language, e.g. English. Operating in 20 official languages, the EU parliamentary sessions are conducted in 20 languages simultaneously, requiring also some 60 interpreters helping politicians from 25 member states to communicate [9].

Every new member country should have 90 full-time professional translators, assistants and other support staff and about 120-130 interpreters. The largest translation office is the Commission's Translation Service (fr. Service de Traduction, SdT) with some 1.500 translators divided between Luxembourg (one third) and Brussels (two thirds) but also every EU institution (Council, European Parliament, European Economic and Social Committee and the Committee of Regions, Court of Justice, Court of Auditors) having its own translation service. Translators and interpreters need to undergo specialized training courses, where only about 30% of them pass successfully.

The main body responsible for translation and written communication is DGT of EC, consisting of some 2000 professionals (more than 1.650 translators and 550 secretaries) who make possible multilingual communication, enabling translation and bringing the EU documentation closer to citizens. DGT is then divided into official EU language departments, which are then specialized in particular subjects, where each unit consists of 20 translators on average.

DGT provides also training to its own staff for translation from the new languages into current EU languages. According to data in press release, out of 245 translators from different language departments trained in January 2005, 89 of them have reached full proficiency and out of 98 translators who have started training in Bulgarian, Romanian, Croatian and Turkish, 7 of them have reached full proficiency. According data in EU News from the 21st January 2005, DGT has only 22 Maltese recruits, while in the next position are Hungary, Latvia and Slovenia with 33 recruits each, out of some 90 translators needed per member country.

The task force prepared for the enlargement in few years, recommending measures of Bulgaria and Romania, in order to provide the same level of service for those languages, would probably apply to the Croatian as well.

5. Computer Tools in DGT

Translating more than million pages per year and employing more than 2000 professionals, the DGT of the EC has been recommending the systematic use of MAT tools, according to Translation Tools and Workflow [2].

5.1. Administration and document tools

Administration and document tools include tools for electronic transmission of translation requests (Poetry software) and integration into electronic archiving system, software for the electronic management of translation requests (Suivi), interface for translation management incorporating reference documents, document comparisons and on-going translation (Dossier Manager), electronic archiving system for all incoming and outgoing documents for DGT (internal software Vista) and Eur-Lex which is public and free of charge, providing direct access to the Official Journal of the European Union containing treaties, secondary legislation and preparatory acts in all official EU language.

5.2. MAT Tools

DGT uses three main types of MAT Tools:

- Terminology tools
- Translation memory technology (at two levels: local and central)
- Machine translation

5.2.1. Terminology Tools

Eurodicautom (Euro - Dictionnaire Automatique) is the European Commission's central terminology database. It is also public, containing more than 6,500.000 terms and 300.000 abbreviations, operating in all EU languages and in Latin. It is in the process of migration to the new term database IATE (Interinstitutional Terminology Database).

Quest is an internet-based metasearch interface, developed by DGT, which translators can use to search several databases at the same time in order to speed up terminology searches.

Euramis is the huge central translation memory, developed as part of Euramis project, providing opportunity to share data between DGT staff, containing 88 million TUs in all official EU languages. It can be accessed through Translator's Workbench and/or Word.

Translator's Workbench (TWB) by Trados is an integrated local TM that holds previously translated pages, chosen because of multilingual capability, integration with word-processing system and performance. Since legislative and preparatory documents are based on existing ones and characterized by repetitiveness, it is used to create revised history of translation material. A set of attributes (translator, domain, number, year, and client) has been defined for individual labeling of segments into TM. It can be also saved into Euramis central TM.

5.2.2. Machine Translation

The principal software for machine translation is Systran (short for System Translation) which has been used since 1976. Based on dictionaries from different domains and linguistic programs, the official version of EC Systran translates between 18 operational language pairs plus a number of prototypes. The EC uses the version distinct from the commercial product, which translates among 36 language pairs in 20 different domains. The quality of the translated text depends greatly on the similarity of source and target languages, restricted domain, writing style and rich dictionaries. Raw outputs are then post processed and manually corrected.

According to A. Petrits [8] MT users of EC Systran could be divided into two main groups: administrators and translators where Systran has three main uses:

- Browsing texts in the language the user doesn't know with high speed of 2000 pages per hour
- Drafting in a language other than mother tongue
- Fast translation, which is the principal reason of use in DGT, of texts having standard structure and terminology (reports, minutes of meeting) and intended for internal use? The raw translations can be corrected.

5.3. Voice Recognition

The speech recognition software used by DGT is called Dragon NaturallySpeaking applied for dictation of text in a natural language, achieving accuracy up to 98% and speed up to 160 words/min. It is available for German, Spanish, English, French, Italian and Dutch. The program saves time for translators who do not wish to type the text.

6. Experiences from EU member states

Facing the same problems regarding EU integration process, new member states have undergone similar problems regarding creation of translation infrastructure, translation capacity, language learning, CAT literacy, terminology management and organizational and educational changes. Number of termbanks, glossaries, bilingual dictionaries as well as style guides for 20 languages have been created.

In Hungary it was necessary to create translation infrastructure for working languages (English, French, German) into Hungarian [5] and vice versa, although the most used was English (95% of legislation translated from English). Terminology was considered the major problem and the task was taken by several organizations and the Ministry of Justice. The next problem was shortage of foreign language speakers and communication at professional level. Besides, the translation workflow had to become more efficient and exploit new technologies. Therefore, the course on use of CAT tools at the Budapest ELTE University was introduced at the postgraduate level, but also as supplemental educational support at several institutions. The course is tailored according to needs of the Hungarian language with special regards to circumstances in EU.

One of close examples is the translation of Slovenian's legislation into mainly English using MAT tool Trados that applies terminology management and translation memory tools.

Translation of the EU legislation into Slovenian [6] was made greatly by assistance of terminology management tool and tool based on TM by Trados (MultiTerm, TWB, WinAlign). TUs are based on translation and glossaries provided by EC, document authors and terminology working groups, enabling database sharing among translators. The term base was intended to be available to the general public While in the first year not much use of TM has been made, the situation was rapidly changing, with intensive work invested in compiling aligned English-Slovenian translation segments when all documents had to be converted into TM so the future translations would be done directly in TWB.

The next stage in which the complete translation need to be revised by an expert, followed by language revision and legal check. The first two years concerned more on organizational and managing requirement, which are then redirected to machine-assisted translation using TWB . Translators were educated to use TWB and build bilingual parallel database. Two linguistic tools Evroterm and Evrocorpus have been developed [11] during translation of EU legislation in order to help manage the EU terminology. Evroterm, a collection of terms, and became useful tool in rendering EU legislation in Slovenian and frequently visited by general public.

Evrocorpus is a bilingual corpus (English-Slovenian) consisting of translation memories of legal acts of EU, enabling thus term comparisons. Both linguistic tools are accessible through SVEZ (Office of European Affairs) website. One of the first resources was Slovene-English parallel corpus of 1 million words which was used for building translation memory database and terminological base.

7. The present state in Croatia

Implementation of MAT Tools requires organizational, educational and professional changes. Since there is considerably augmented need for translation of Croatian, some steps have been undertaken in order to start with preparation activities for EU integration.

Regarding organizational changes where use of MAT tools is integrated into translation workflow, several private and state companies have adopted translation technology. After creation of positive climate explaining benefits, limitations and necessary conditions for use of translation technology, translators had possibility to attend seminars and workshops organized for the specific purpose.

Translation work in the case of Croatian is the most often assisted by some “traditional” tools, such as online dictionaries, glossaries, spelling checkers, thesauri, CD ROMs and text revision tools, although there are some isolated examples of interest to use TM tools more consistently for translation from/into Croatian, or to use machine translation for other language pairs.

Ministry of Foreign Affairs and European Integration is in charge of translation of “acquis communautaire” into Croatian and translation of Croatian regulations into English. As the consistent use of terminology is absolutely needed, the following manuals and glossaries have been created and used: Glossary on Stabilization and Association Agreement (Cro-Eng, Eng-Cro), Manual for translation of EU legal acts, Glossary of Banking, Insurance and other Financial Services (Eng-Cro), Euroterm – glossary of terms for translation of “acquis communautaire” and of Croatian legislative to English. Eurovoc thesaurus has been translated into Croatian by HIDRA, intended for indexing of documents on EU activities.

7.1 Integration into curriculum

Integration of MAT/MT Tools and education for future jobs within university setting has been one of the first steps. The courses are offered at undergraduate and postgraduate levels. The course «Machine Translation» is offered as obligatory and elective course at the 3rd year of the Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb for students of information sciences and other study group, mainly philological. The course «Machine-Assisted Translation Tools» has been introduced into doctorate level of the same department. The course «Translator and Computer» is offered as an elective course at the 4th year of the Department of Linguistics. Besides this, two postgraduate specialist studies for translation are offered at the Faculty of Humanities and Social Sciences. The courses have been adapted according the Bologna requirements with considerable number of European and American universities, regarding the theoretical approach, practical exercises, seminar work, literature and ECT system.

Cooperating with some professional and non-academic institutions the courses aim to practical

solutions aiming to incorporate into the curriculum practical use of MAT tools and educate students for a new type of jobs.

7.2. Questionnaire

In order to proceed with preparation activities regarding organizational, educational and professional changes it would be necessary to conduct a questionnaire among translators working in different organizations, institutes, companies, ministries or as freelancers regarding use of MAT tools. The main aim would be to obtain insight into the present state and readiness of translators to use the MAT tools. For this purpose, it would be necessary to determine number of translators and target groups, whether translation is (non)professional job, if they work individually or in team, number of translated pages (in general and last year), type of documents for translation (language, direction and field of translation), type and amount of documents translated using MAT tools, which MAT tools were used, awareness of difference between translation memory and machine translation, readiness for further education and assessment of usefulness.

8. Conclusion

Over the past few years Croatian companies, ministries and other institutions have been undergoing considerable changes, caused among other things, by growing need for quick translation, localization and multilingual contacts. In order to cope with organizational, educational and professional changes and to start with preparation activities for integration into EU, it is necessary to analyze and adopt EU standards regarding language resources, data exchange, translation tools, educational and professional demands.

As preparation for translation work in the EU is long-term, systematic, well-organized job, it is necessary to build-up language resources for Croatian, where examples of the neighboring countries could be very useful.

Creation of TMs with Croatian as a language pair, intended for future reuse, would augment considerably consistency, enable data-sharing, reduce costs, increase speed and integrate Croatian module into European environment.

Therefore existence of digital resources for Croatian, educational and organizational changes and adaptation of EU standards represent necessity for any further cooperation , a way to preserve cultural heritage and inclusion of Croatian modules into European environment.

9. References

[1] 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]

[2] 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]

[3] Hutchins J. 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]

[4] Hutchins J. 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]

[5] Kis B. Technology in the Translation Class: Introducing CAT Tools to Hungarian Translation Students http://www.uem.es/web/fil/invest/publicaciones/web/EN/autores/kis_art.htm [20/1/2006]

[6] Krstič A. Translation of EU legislation in Slovenian; 2000. http://www.eamt.org/archive/ljubljana/Krstic.htm [12/1/2006]

[7] Loffler-Laurian A-M. La traduction automatique. Villeneuve d'Ascq: Presse Universitaire du Septentrion; 1996.

[8] Petrits A. 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]

[9] Roxburgh A. Translating is EU's new boom industry. BBC News Online. http://news.bbc.co.uk/2/hi/europe/3604069.stm [20/1/2006]

[10] Systran: Language Translation Technologies http://www.systransoft.com/company/index.html [10/1/2006]

[11] Vintar Š. A Parallel Corpus as a Translation Aid: Exploring EU Terminology in the ELAN Slovene-English Parallel Corpus. http://www2.arnes.si/~svinta/germ.rtf [20/3/2006]