Seljan, Sanja; Gašpar, Angelina; Pavuna, Damir. *Sentence Alignment as the Basis For Translation Memory Database.* // INFuture2007-The Future of Information Sciences: Digital Information and Heritage / Seljan, Sanja; Stančić, Hrvoje (ur.). Zagreb: Odsjek za informacijske znanosti, Filozofski fakultet, 2007.

Sentence Alignment as the Basis For Translation Memory Database

Sanja Seljan, Ph.D.
Faculty of Humanities and Social Sciences – University of Zagreb
Department of Information Sciences
Ivana Lučića 3, 10 000 Zagreb, Croatia
E-mail: sanja.seljan@ffzg.hr

Angelina Gašpar
SOA Centre Split
Katalinića prilaz 2, 21000 Split
E-mail: ginasplit@yahoo.com

Damir Pavuna, M.Sc.
Integra d.o.o.
A. Stipančića 18, 10000 Zagreb
E-mail: damir.pavuna@integra.hr

Summary

Sentence alignment represents the basis for computer-assisted translation (CAT), terminology management, term extraction, word alignment and cross-linguistic information retrieval. Created out of the sentence alignment process, translation memory (TM) represents the basis for further research in translation equivalencies. Automatic sentence alignment, based on parallel texts, faces two types of problems: robustness and discrepancies between source and target texts in layout and omissions which have an influence on the accuracy of the alignment process.

The aim of the paper is to present research on the sentence alignment process carried out on the Croatian-English parallel texts (laws, regulations, acts and decisions) and implemented by the alignment tool WinAlign 7.5.0 by SDL Trados 2006 Professional.

The alignment process and its impact on the creation of translation memories is presented through comparison of translation memories that differ regarding the levels of expert intervention in the set up of the alignment program and preparation of the source text for the segmentation. Recommendations for further development using statistical analysis, automatic learning techniques and language knowledge are suggested.

Key words: sentence, alignment, translation memory, computer-assisted translation (CAT), tool, segmentation, set up

Introduction

The need for fast translation of a large number of pages in several languages, use of specialized and consistent terminology, sharing of common resources, time-saving, cooperation in larger translation projects and cost-saving, have caused a growing use of translation memories.

Sentence alignment represents the basis for computer-assisted translation (CAT), terminology management, term extraction, word alignment, cross-linguistic information retrieval, etc. Created out of sentence alignment process, translation memory (TM) represents the basis for further research in translation equivalencies.

Witnessing the importance of the sentence alignment process, different international projects have been undertaken in order to develop its evaluation metrics: the two-year project ARCADE (1995-96) aiming to produce a bilingual French-English corpus suited for the alignment task and its evaluation; MULTEXT-East Project (Multilingual Text Tools and Corpora for Eastern and Central European Languages) where each of the six translations of the novel 1984 by G. Orwell were sentence aligned with the English original, and the alignments hand validated; the Egypt Statistical Machine Translation Toolkit (1999) and the following GIZA++ for training statistical translation models.

In this paper, the sentence alignment problem is elaborated through several aspects. The main reasons are presented for the use and development of translation memories created out of the alignment process and their integration into translator's workbench.

As translation memories (TMs) work best on the voluminous and highly repetitive types of texts (e.g. new versions of software or products, regulations/laws, decisions, catalogues, manuals) the research was done on Croatian-English parallel legislative texts. Therefore, the research was done on a highly structured type of text (legislation) whose main characteristics are presented (structure, enacting formulas, specific terms and expressions) that can influence the alignment process.

The results of the alignment processes are presented through bitexts, created out of Croatian-English parallel texts, which are imported into translation memories. As various approaches in the alignment process are used, different types of TMs are elaborated and tested. In the conclusion, recommendations for further development are suggested.

The research presented here is an outcome of the research project "Information Technology in Computer-Assisted Translation of Croatian and in e-Language Learning" (130-1300646-0909) undertaken with the support of the Ministry of Science, Education and Sport of the Republic of Croatia.

When to use TMs?

Need for fast and consistent translations are obvious in the EU and for candidate member states when a large number of legislative documents are to be translated, but also in multilingual societies, multinational companies, in government institutions and agencies, or when translating simultaneously multilingual documentation for new versions of products and services in several languages. Translation memories are used to speed up the translation process, enable the sharing of resources, consistent terminology and cost-reduction.

In the process of creation of TMs, there are two possible ways:

- When the translator is giving input through the source text which has to be translated, the program scans the text trying to find matching pairs as full match or fuzzy matches, which are then subject for the translator's review. The new segments are then stored in the database, i.e. translation memory and can be used in future work.
- Another way is the building of TMs out of already translated material through the alignment process, as presented in this case study.

Translation memories are usually integrated with other CAT (Computer-Assisted Translation) tools (e.g. word processing programs, multilingual dictionaries, thesauri, terminology management system, machine translation software) into translator's workbench. Local translation memories can be integrated into the central TM stored on the central server, which is usually part of the global management system.

Translation memories are mostly used when a translator has a feeling of "having already translated something like this". This is where the TM has the best effect: it offers the same translation (100% match) or similar translations (fuzzy matches) using already translated units. Besides this, TMs are a valuable resource for concordance search to determine the appropriate use of the term or as a terminology management source providing specialized terminology.

Corpus used

Translation memories are mostly used for translation of highly structured and voluminous documents. The use of controlled language, specific forms and structures can augment results obtained using CAT tools.

As the research is done on a highly structured type of text (legislation) the main characteristics are presented regarding its structure, enacting formulas, specific terms and expressions that can influence the alignment process. The results of the alignment processes are presented through created bitexts of Croatian-English texts that are imported into translation memories.

The alignment is carried out on parallel texts, consisting of corresponding texts between Croatian and English. The analysis is based on Croatian legislative acts: laws, regulations, decisions, and ordinances (NN122/03, NN51/04, NN49/03, NN30/97, NN10/02, NN164/04, and NN 173/03) related to

competition acts, crafts act, trademarks, electronic signature, agreement of minor importance, of relevant market, on concentrations and bylaws. The bilingual corpus was examined regarding its structural and lexical levels since various acts have their own standard presentation and standard formulas. They are set out in the Style Guide for Croatian Legislation inspired by the English Style Guide from EC DGT (2005) and the Interinstitutional Style Guide (2005). The percentage ratio for word count in English translations is 33.15%, which is due to the fact that the English language is an analytic type of language, contrary to the Croatian language, which uses a highly flective system. This is presented in Table 1, which compares the number of words in Croatian and English texts.

Table 1: Size of parallel texts

	Words	**Characters** (with spaces)	**Pages** (1.800 char./page)
Croatian	**39,956.00**	274,198.00	**152.32**
English	**53,203.00**	330,513.00	**176.29**
Total	93,159.00	604,711.00	**328.61**

Standard structure and formulas

The drafting style takes account of the type of act for the sake of uniformity and terminological consistency. However, every act is specific for its own repetitive legal terms, phrases and sentences. For instance, the following main components of a regulation are: the title, preamble (citations, recitals, enacting formulas which can be obligations, permissions, admissions and statements), enacting terms, addressee, place, date and signature.

Acts with a simple structure comprise articles and subdivisions of articles. The arrangement of enacting terms in Croatian acts is the following: Part (Division, Title, Chapter, Section, Subsection, Heading, Subheading) and Article (Paragraph, Subparagraph, Point/Item, Indent, Annex, Appendix; Schedule). Also, the textual components of the enacting terms comply with relatively strict rules of presentation: the subject matter and scope, the definitions, the provisions conferring implementing power, provisions concerning penalties or legal remedies, transitional and final provisions. Each article contains a single provision or rule laid down in an act. Furthermore, the standard form prescribes the layout on the page, including spacing, paragraphing, punctuation and even typographic characteristics (capitalisation, typeface, boldface and italics).

Use of verbs in enacting terms

The enacting terms of binding Croatian legislation can be divided into imperative and declarative terms. While declarative terms refer to definitions or amendments, in order to express commands and prohibitions, the Croatian language often uses the present tense whereas 'shall' is used in English translations. For instance, /Sudionici koncentracije obvezni su podnijeti…/ is

translated into English as / The parties to the concentration shall be obliged to submit…/or /Zabranjeni su svi sporazumi…/ /There shall be prohibited all agreements/. Also, modals in Croatian legislation such as 'morati' or 'trebati' are also translated into English as 'shall' or 'must' meaning 'is required to'; for instance, /Informacijski sustav …treba biti oblikovan tako..'/ The information system shall be organized in such way...'/ or /'davatelj usluga certificiranja….mora čuvati svu dokumentaciju'/ /'Certification authorities…. must safeguard all documentation'/. Prohibition is expressed in Croatian legislation by terms 'ne može se' or 'nije dopušteno', which have an English equivalent in the term 'may not'. Also, permission in the Croatian language is expressed by the terms 'može se', 'smije se' or 'dopušteno je', which have the English equivalents 'may' or 'it is admissible'. The terms such as 'it is permitted' or 'it is allowed' are not used in English legislation. For instance, / 'Agencija može…odobriti produženje roka'/ The Agency may ...extend the period...'/ or /'Iznimno od odredbe…dopušteno je članstvo...'/ 'Without prejudice to the provision…, it is allowed to be a member...'/; /'Dopuštena je svaka gospodarska djelatnost...'/ 'Every economic activity ...is permitted'/.

Authorisation is expressed by two terms in Croatian: 'ovlašten je' or 'može se', whereas English legislation uses several terms such as :'may', 'is authorised to' , 'is empowered to' , 'has the power to' or 'shall', in case the authorisation is binding or comprises certain activity. For example, /'Predsjednik i članovi Vijeća mogu pisati...'/ /'The President and the members of the Council are authorised to write...'/.

As far as expressing the rights is concerned, Croatian legislation uses indicative expression 'imati pravo' which corresponds to the English expressions 'has the right to', 'is entitled to' or 'may'. Definitions of legal terms are binding and are expressed with the present tense in Croatian but still translated as 'shall' in the EU, although the Anglo-American legislative has recently started using present. /'Pojedini izrazi koji se rabe u ovom zakonu imaju sljedeće značenje:'/ /'Individual terms in this Act shall have the following definitions:'/

Verbs expressing descriptive functions or statements such as 'biti', 'postojati', 'nalaziti se' and 'imati' are used in the present tense in Croatian, which correspond to English 'There shall be' or rather 'There is hereby established'. For instance, /'Upisnik je knjiga koja sadrži podatke i isprave...'/ /'The Register is a book containing the data and documents...'/.

Capital letters

In Croatian language only the first noun in names of institutions, administrative bodies, laws, geographical names, agreements etc. is capitalised whereas in English all nouns and adjectives are written in capitals. For instance, /Ministarstvo obrazovanja i sporta/ Ministry of Education and Sports, / Zakon o zaštiti tržišnog natjecanja/ The Competition Act, / Madridski sporazum/ the

Madrid Agreement. All titles in both source texts and their translations are capitalised.

Hyphens and compound words

The hyphens used in English compound words, for instance, 'a five-year-term' / 'razdoblje od pet godina' or 'local self-government' / 'jedinica lokalne samouprave', are omitted in the source text. It is to stress that some compound words in the Croatian language do not have the same word order as in English translation, e.g. / 'jedinice lokalne i regionalne samouprave' / 'bodies of regional government and local self-government'/.

Punctuation

The punctuation is almost the same in both the source texts and the translations. A semicolon is mostly used instead of a linking conjunction and to separate intends. The comma is used to divide adjectives in series but also before 'and' and in parenthetic and introductory phrases. For instance, / 'Usmena rasprava je, u pravilu, javna.'/ 'The oral hearing is, as a rule, public.' /. Round brackets or parentheses are used when citing numbered paragraphs in English, for instance, /'Article 11 paragraph (1) item1 '/, but in source texts they are omitted, /'…članak 11.stavka1.točka 1.

Numbers

Numbers up to 10, but also larger numbers, are written either in words or in figures in both the source texts and the translations, for instance, /...'najmanje jednu milijardu kuna'/ corresponds to English /'1 billion Kuna/ or /'..najmanje 100.000.000,00 kuna'/ ...at least 100,000,000.00 Kuna. Although in plural and lowercased, the symbol for Croatian money 'kuna' is capitalised and written in singular, 'Kuna', in translations.

Dates

Dates use figures for days and words for months both in the source texts and the translations. However, the usage is different: / '21.srpnja 2003.'/ '21 July 2003' / or '7.21.03' in the American dating system, or '2003-07-21' in the international one.

Foreign words, expressions and synonyms

Some foreign Latin words and expressions can be found in both the source texts and the translations, and also English words and expressions in the source texts (e.g. *ex officio, know- how, joint venture, world-wide, franchising*). Synonyms used in the source texts are mostly avoided in the translations, for instance, /'pripajanjem ili spajanjem poduzetnika'/ 'merger association of undertakings'/ or /'Stjecanje dionica ili udjela'.../Acquisition of Shares'.../ /'grafički prikaz ili dijagram' / 'the graphic presentation (diagram)'/.

Both the source texts and the translations have consistent terminology. Defined terms are used in a uniform manner in order to facilitate comprehension and interpretation of legislative acts. Gender-neutral language is preferable. While the Croatian language uses the active voice more frequently, the English language makes more use of passive. Sentences in the active voice are generally, though not always, clearer and more concise than those in the passive voice because fewer words are required to express action. Unlike English, Croatian language has a very rich case system whose nouns, pronouns and adjectives are inflected by the case. It is due to this variety that the cases bear the main burden in marking the syntactic functions of a noun phrase and that word order is relatively free.

Research

The alignment was carried out on Croatian legislative acts: laws, regulations, decisions, and ordinances and their respective English translations (a total of 328.61 pages). After comparing the number of words and pages, automatic alignment was carried out with the total number of translation units. The comparison between different types of alignment models is presented together with their impact on the creation of translation memories. The results between four types of translation memories are elaborated distinguishing the level of expert intervention in the setting up of segmentation parameters and source document segmentation.

Tools used

In the research we used SDL Trados 2006 Professional, part of which is the alignment tool WinAlign 7.5.0. For document structure analysis we used a very common tool for such purposes, AnyCount 4.0 (version 405). Bitexts in the .txt form are exported out of the WinAlign tool and imported into SDL TRADOS Translator's Workbench 7.5.0. Translation memory is then saved in SDL Trados native .TMX format, which is standard and convertible to almost all recommended format.

Activities

In the field of language technologies, the term bitext is used, denoting a merged document consisting of the source and target texts, generated by the alignment tool. A collection of *bitexts* is called bitext database or bilingual corpus. The main difference between bitext and translation memory is that matched segments are stored in the way that is unrelated to the original, with lost sentence order, while the bitext holds up the original sentence order.

As translators often have at their disposal a considerable amount of translated material, it can be aligned and converted into a TM database, although certain preparatory activities should be taken:

- comparison of the source and target texts (whether all text is translated)

- defining set up of end and skip rules (delimiters, creating abbreviation user list)
- preparation of the source text for better segmentation (spelling, automatic bullets and numbering, deleting of soft returns, hyphens, certain punctuation, tables created with tabs and revision marks)
- modification of set up rules
- verification of the alignment (especially 1:2 and 2:1 pairs and commitment of pairs)
- creation of translation memory and verification.

Automatic alignment

WinAlign has language independent algorithms that count:

- the quality of translation units which can have tree levels (low, medium, high)
- translation units aligning 1:2 or 2:1 pairs
- unconnected target segments.

In the case study, nine legislative documents in parallel Croatian and English languages were aligned. Statistics from Table 2 represent the results of the automatic alignment using language independents algorithms.

Table 2. Automatic alignment

Target File Name:	1	2	3	4	5	6	7	8	9	**To-tal**	**Per-cent**
N. of source segments:	530	525	262	82	107	218	249	132	484	2589	104.73
N. of target segments:	504	512	290	89	107	224	246	129	482	2583	104.49
N. of aligned units:	503	504	255	77	104	211	230	124	464	**2472**	
N. of com-mitted units:	0	0	0	0	0	0	0	0	0	0	0.00
N. of high quality units:	259	193	71	25	34	51	57	43	226	959	**38.79**
No. of medi-um quality u.:	217	297	152	45	67	149	153	73	222	1375	**55.62**
N. of low quality units:	27	14	32	7	3	11	20	8	16	138	**5.58**
N. unconnec-ted source s.:	8	0	1	0	0	0	0	0	2	11	**0.44**
N. unconnec-ted target seg:	0	0	18	2	0	0	1	0	0	21	**0.85**
N. of 1:2 and 2:1 units:	20	29	23	15	6	20	34	13	36	196	**7.93**
										2472	**Align ed**

152.32 pages of the Croatian text were aligned with 176.29 pages of the English text, creating automatically all together 2,472.00 translation units (alignment performed 104.73%) and 11 units that were not aligned (0,44%). Out of the total number of aligned units, 38,79% (959) were marked as high quality units, 55,62% (1.375) as medium quality units, and 5,58% (138) as low quality units. 7,93% (196) were marked as 1:2 and 2:1 aligned units.

From the figures presented, it can be seen that every alignment should be verified, manually corrected and the whole process supervised. Part of the problems relate to different layout of texts, omissions, inversions, different structure orders and paragraph numbering. Therefore, expert intervention in the set up of the alignment program and pre-editing activities of the source text for better segmentation should be included. That way, improper segmentation would be reduced, since automatic marks would be hidden, alignment would be carried out relating to the text and the number of high quality translation units augmented.

Automatic and manual alignment

The significant difference between alignment processes made with and without expert interference is presented in Table 3. The first column shows automatic alignment without a language expert. As WinAlign uses language independent algorithms, it is estimated that the automatic alignment process found 5.58% of low quality units, 55.62% of medium quality units and 38.79% of high quality units.

Table 3. Alignment: automatic vs. manual

	Automatic alignment		**Diffe-rence**	**Manual alignment**	
Text 8 (Bylaws)	No.	%		No.	%
N. of source segments:	132	104.73	**?**	120	100.00
N. of target segments:	129	104.49	**?**	125	104.17
N. of aligned units:	124		**?**	120	
N. of committed units:	0	0.00	**High**	120	**100.00**
High quality units:	43	**38.79**	**High**	0	0.00
Medium quality units:	73	**55.62**	**High**	0	0.00
Low quality units:	8	**5.58**	**High**	0	0.00
Unconnected source s.:	0	0.44	**OK**	0	0.00
Unconnected target s.:	0	0.85	**OK**	0	0.00
N. of 1:2 and 2:1 units:	**13**	7.93	**High**	**5**	4.17
Aligned:				120	

After the manual alignment, all sentence pairs are marked as committed units (as presented in 'Manual alignment' columns), out of which the TM base can be created. Another problem are 1:2 and 2:1 units. As WinAlign does not have any language algorithm and does not care which are the source and target languages, in this study there are 13 cases (7.93%) marked as 1:2 or 2:1 pairs. Out of the 13

suggested cases suggested in automatic alignment, 8 were wrong, and the total number of such cases is 5, as stated in manual alignment.

Comparison of TMs

Table 4 presents four types of translation memories:

- TM created out of the automatic alignment using SDL Trados language independent engine with null expert intervention (Raw TM)
- TM created out of the alignment but without any expert intervention in the set up of the alignment program and without intervention on the text segmentation, stating only that the source text corresponds to translated target segment (Aligned TM)
- TM created out of the automatic alignment, with expert intervention in the set up of the alignment program (Aligned TM + Set up rules, e.g. segment and skip rules, abbreviation user list)
- TM created out of the manually confirmed alignment, including setting up of segmentation rules in the alignment program and expert intervention on the segmentation of the source text (e.g. changes of soft returns, check of colon segmentation)

The presented translation memories differ regarding expert intervention in the WinAlign set up and in the segmentation of the source text. When translating the same text, out of which the TM has been created, it is to be expected that this automatic translation would completely match the created TM, and that machine translation would match 100%. But this is not the case, especially when TMs are created automatically using a language independent engine and without expert intervention in the setting up of segmentation rules and revision of the segments in the source text.

For the purpose of this study the following changes were made in the set up of the alignment program: ":" is not considered as delimiter, "br." is not considered as delimiter but moved to the abbreviation user list, "I." is also not considered as delimiter but moved to the abbreviation user list.

In the process of segmentation of the source text, soft return was eliminated and ":" deleted.

Therefore, the presented evaluation was made on the automatic translation of the same text, out of which different types of TMs had already been made with the difference regarding expert intervention in the process of creation of TMs.

Results

The automatic alignment presented in the first column was carried out by the language independent engine and with null expert intervention (Raw TM). Although the results seem very good (the same text translated with 91.67%) thanks to highly structured texts, alignments are very imprecise and wrong, often without sense and linguistically incorrect. The TM created out of this alignment contains 61.2% of medium and low quality units (see Table 3),

suggesting that more than every other segment is not properly aligned. Therefore, the generated translation would be unclear and useful only for experimental purposes.

Table 4. Alignment: automatic, manual

	Raw TM	**Aligned TM**	**Aligned TM**	**Aligned TM**
			Set up rules	**Set up rules**
Segments/				**Segmented source docum.**
Context TM	0	0	0	0
Repetitions	0	0	0	0
100%	121	106	112	**120**
95% - 99%	0	0	0	0
85% - 94%	2	5	0	0
75% - 84%	2	2	1	0
50% - 74%	1	1	2	0
No Match	6	18	11	0
Total	132	132	126	120
Percent	**91.67%**	**80.30%**	**88.89%**	**100.00%**

Although the results using aligned TM without any expert intervention (Aligned TM) seem much worse (80.30%), all translated text is linguistically correct. The first column, in spite of results, can not be compared with other three columns since their translated segments are linguistically correct and correspond to each other, which is not the case with Raw TM presented in the first column.

With the setting up of segmentation parameters (e.g. segment end and skip rules, creating abbreviation user list), the result is much better (88.89% text translated).

Ultimately the 100% translated text could not be produced without the setting up of segmentation parameters and without the preparation of the source text for segmentation, as in the last case when an expert used the tool to see hidden characters and made final changes (in practice it is advisable to do it at the beginning of the whole alignment process) in the source text and according to the experience in the alignment process (e.g. the changing of soft returns, checking colon segmentation).

Conclusion

Sentence alignment is prerequisite for further corpus processing and research in the fields of computer-assisted translation (CAT), terminology management, term extraction, word alignment and cross-linguistic information retrieval. In the process of sentence alignment two main types of problems are considered: robustness, differences in layout (omissions, inversion, 1:2 or 2:1 alignments) between the source and target texts, and the segmentation of the source text in order to achieve better accuracy and to create a translation memory of good quality.

The standard and uniform manner of legislative texts, prescribing the layout (space, paragraphs, punctuation and capitalisation) and relatively strict rules should facilitate sentence alignment, although expert intervention is necessary, as presented in comparison of different translation memories.

The translation memories created in this study out of different types of the alignment processes give different results regarding the quality of the translated material. The results show necessary interventions of an expert when defining the set up rules, in preparation activities for the source text segmentation and in the verification of suggested translation units.

A good quality translation memory created out of such an alignment process then becomes a valuable source for further research in translation equivalencies, terminology extraction, terminology management, word alignment or cross-linguistic information retrieval. Although the results are augmented, they are static and could be improved with the integration of language knowledge, rule-based algorithms and further research in statistical alignment. Integrated with other translation tools, TMs can be a useful tool to increase the speed and augment terminology consistency in the translation process, but with a human as the main supervisor.

References

Arcade: Evaluation of parallel text alignment system. (http://www.up.univ-mrs.fr/veronis/arcade/arcade1/index-en.html, 01st September 2007)

Barzilay, Regina ; Elhadad, Noemie. Sentence alignment for Monolingual Corpora. Proceedings of the Conference on Empirical Methods in NLP - EMNLP, 2003, Japan.

Ceauşu, Alexandru; Ştefănescu, Dan; Tufiş, Dan. Acquis Communautaire Sentence Alignment using Support Vector Machines, LREC 2006, Italy.

Corpus Markup. COP Project 106 MULTEXT-East: Work Package WP2 - Task 2.3, 1997. (http://nl.ijs.si/ME/CD/docs/mte-d23f/node2.html, 25 August 2007.)

del Pino, Santiago. Using Translation Memory Software (TMS): An Organisational Checklist. Terminologie et Traduction, pp. 132-139.

DGT of the EC. Translation Tools and Workflow, 2005. http://europa.eu.int/comm/dgs/translation/bookshelf,/toolsandworkflowen.pdf

EC DGT. English Style Guide: A handbook for authors and translators in the European Commision, 2005 (http://ec.europa.eu/translation/writing/style_guides/english/style_guide_en.pdf)

Egypt Statistical Machine Translation Toolkit, 1999. (http://www.clsp.jhu.edu/ws99/projects/mt/toolkit/, 25 September 2007)

GIZA++: Training of Statistical Translation Models. http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html, 25 August 2007)

Krstič, Adriana. Translation of EU legislation in Slovenia. Proceedings of the 5th EAMT Workshop, 2000, Slovenia.

Knauf, Ansgar. TR16: Development, Use and Profitability of Transaltion memory Systems. Translation Issues, 4-99. December 1999.

LISA: 2004 Translation Memory Survey (http://www.lisa.org/products/surveys/tm04survey.html)

LISA: The Localization Industry Standards Association. (http://www.lisa.org)

Marrafa, Palmira ; Ribiero, António. Quantitative Evaluation of Machine Translation Systems : Sentence Level. Workshop on Example-Based Machine Translation, Proceedings of the MT Summit VIII conference, 2001, Santiago de Compostela.

Och, Franz Josef; Ney, Hermann. Statistical Machine Translation. Proceedings of the 5th EAMT Workshop, Ljubljana, 2000.

Office for Official Publications of the European Communities:Interinstitutional Style Guide, 2004 (http://eur-lex.europa.eu/en/techleg/pdf/en.pdf)

Planas, Emmanuel. Extending Translation Memories. Proceedings of the 5th EAMT Workshop, Ljubljana, 2000.

Priručnik za prevođenje pravnih propisa RH na engleski jezik, MVPEI, Zagreb, 2006. (http://www.mvpei.hr/ei/download/2007/02/26/prirucnik_za_prevodenje_pravnih_propisa_RH.pdf)

Seljan, Sanja; Pavuna, Damir. Translation Memory Database in the Translation Process // Proceedings of the 17th International Conference on Information and Intelligent Systems IIS 2006, Varaždin : FOI.

Simard, Michel; Plamondon, Pierre. Bilingual Sentence Alignment: Balancing Robustness and Accuracy, 1998.

Steinberger, Ralf ; Pouliquen, Bruno ; Widiger, Anna ; Ignat, Camelia ; Erjavec, Tomaž; Tufiş, Dan; Varga, Dániel. The JRC-Acquis : A Multilingual Aligned Parallel Corpus with 20+ Languages. Proceedings of the 5th International Conference on Language Resources and Evaluation LREC, Genoa, Italy, 24-26 May 2006.

Wiki http://en.wikipedia.org/wiki/Translation_memory (2nd September 2007)

Zetzsche, Jost. Translation Memories: The Discovery of Assets: Recognizing opportunities and overcoming obstacles to TM sharing. MultiLingual Computing & Technology 72, vol. 16 (4), 2005.