To post-edit or not to post-edit?

Estimating the benefits of MT post-editing for a European organization

Alexandros Poulis
Intrasoft Intl
European Parliament
alexandros.poulis@ext.europa
rl.europa.eu

David Kolovratnik
Charles Oakes & Co
European Parliament
david.kolovratnik@ext.europa
rl.europa.eu

Abstract

In the last few years the European Parliament has witnessed a significant increase in translation demand. Although Translation Memory (TM) tools, terminology databases and bilingual concordancers have provided significant leverage in terms of quality and productivity the European Parliament is in need for advanced language technology to keep facing successfully the challenge of multilingualism. This paper describes an ongoing large-scale machine translation post-editing evaluation campaign the purpose of which is to estimate the business benefits from the use of machine translation for the European Parliament. This paper focuses mainly on the design, the methodology and the tools used by the evaluators but it also presents some preliminary results for the following language pairs: Polish-English, Danish-English, Lithuanian-English, English-German and English-French.

1 Introduction

The European Parliament (EP) has witnessed a significant increase in translation requests in the last few years. For instance the total amount of source pages translated by the Directorate General for Translation (DGTRAD) in the first quarter of 2010 was 43,963. In the first quarter of 2012 this number increased to 60,275 while the number of translators has remained rather stable. This situation requires a significant productivity increase in the most cost-efficient way so that DGTRAD can keep accomplishing its mission:

making available in all official languages of the European Union (EU) all documents relating to EP's role as co-legislator and enabling the EP to permit all EU citizens to communicate with the EU institutions in their own language as efficiently and effectively as possible.

So far all this has been possible thanks to the extensive use of various translation technologies such as Translation Memory systems, terminology databases, bilingual concordancers and other reference tools which have been seamlessly integrated in the translation workflow in the last 6 years. Nevertheless, current demand requires new technologies to be tested and Machine Translation is probably the most important one.

To examine what can be expected and evaluate the most obvious deficiencies we organized a large-scale evaluation of a general-purpose MT system developed by the European Commission (Eisele et al. 2011). The tests will be conducted by 62 translators in 24 language pairs.

1.1 Use-case

The vast majority of EP documents are written in English, with French and German following in the second and third place. On that basis we decided to start testing the following language pairs: English to all official EU languages (Table 1), German to English and French to English. Each evaluator works always from one source language into her mother tongue.

For the current round of tests we have selected documents which do not contain highly repetitive text and therefore their segments are rarely found in our translation memories. Some of these document types are parliamentary questions,

petitions, notes from various bodies of the EP and draft resolutions[1]. With translation memories not providing much input for those document types we see a strong case where MT could be of some help to translators.

MT can and most probably will be used for other purposes such as communication and gisting but this study focuses only on its use as a translation aid.

EU Languages	
Bulgarian	Italian
Czech	Latvian
Danish	Lithuanian
Dutch	Maltese
English	Polish
Estonian	Portuguese
Finnish	Romanian
French	Slovak
German	Slovene
Greek	Spanish
Hungarian	Swedish
Irish	

Table 1: EU official languages.

2 Translation technologies in the current workflow

The current translation workflow relies largely on Translation Memory (TM) technology which is the main component of the so called Translation Environment Tools (TEnT). TMs are large databases that contain pairs of segments (usually sentences) in the source and target language. Each such pair of segments is called a translation unit. Translation memories can be bilingual (one source and one target language) or multilingual (one source and multiple target languages). In the EP the available TEnTs support only bilingual memories although this will change in the near future. As one source segment may have more than one translation equivalents within the same TM, each translation unit contains also some meta-data that provide information about its origin, creator, requestor and its creation date. These meta-data can help the translator assess the reliability of each available translation option for a given segment and select the most appropriate translation in a

---

[1] For more information about and access to European Parliament's documents please visit http://www.europarl.europa.eu/RegistreWeb/search/typedoc.htm?language=EN

given context.

While translating a document each source segment is compared to the TM content and translations of matching segments are proposed by the system. Matching segments can be either identical to the source segment (100% or full match) or similar to it (fuzzy match). Fuzzy matches are usually between 65% and 99%. Full matches are usually accepted without changes but fuzzy matches need to be post-edited.

Besides TMs our translators have access to large terminology databases which are constantly enriched with the support of a dedicated terminology service which makes sure that the terminological entries are inserted in time for new translation projects and that they are complete including all of our working languages and references following expert translators' or terminologists' quality approval.

Bilingual concordancers enable searches of terms, phrases or any strings within their context. Depending on the input format of the tool that context can be a whole document or just a translation memory segment.

An interinstitutional search engine called Quest2 brings many databases under a common user interface and offers almost 4,500 translators access to various reliable terminology, document and TM resources.

All these tools have helped the EP cope with the increasing workload so far. Nevertheless, it is clear that some additional leverage is needed and MT seems to be the way to go. In section 2 we mentioned that TMs can provide 100% matches and fuzzy matches. If no matching segment is found in the database or if the match value is lower than 65% the source segment needs to be translated from scratch. It is primarily –but not exclusively- in those cases that MT can be of use if it is of sufficient quality to allow for faster post-editing than translating the whole segment. When fuzzy matches are available the MT output will be offered to the user too. Previous research such as (Simard et al. 2009) has shown that MT performs better when there is also a good fuzzy match and its usability may even outperform that of the corresponding fuzzy match. In-house experience has shown that MT output can help translators edit the fuzzy matches faster. Taking this into consideration we are currently investigating the possibility of automatically enhancing the fuzzy

matches with MT implementing the algorithm proposed by (Koehn et al. 2010). The introduction of MT to the workflow will have a great impact on the role of translators. They will now mainly be asked to post-edit TM and MT output rather than translate free text at least as far as certain document types and language pairs are concerned.

3 The project MT@EP

Following the promising results of the Exodus experiment which was presented in (Jellinghaus et al. 2010) the EP DGTRAD decided to launch an IT project with the objective of estimating the benefits of MT and ensuring its efficient implementation in the translation workflow but also potentially as a communication tool between staff members or between the citizens and the EP. This paper focuses only on the first use case - MT as a translation aid. MT is expected to bring certain benefits to the EP; therefore MT post-editing is being carefully evaluated taking into account various parameters which are presented in section 4.

3.1 Expected business benefits

DGTRAD expects that MT will help increase its translation productivity - measured in number of standard pages[2] per period of time - at least for certain document categories/domains and language pairs. MT is expected to offer more added value to domains with higher availability of internal documents that can be used in the training corpora as well as to language pairs with higher data availability and similarity between source and target. First experiments confirm this view showing that reaching usable MT quality levels when translating into Finish, Hungarian or other morphologically rich languages is much more challenging than most other language pairs. This does not come as a surprise as it has been repeatedly observed in the MT literature as for example in (Koehn 2005). To what extent can DGTRAD expect MT to increase its translation productivity and how can we estimate that? This is the main question that we will try to answer in the next sections of this paper using a MT post-editing and some other MT evaluation tasks.

At the same time it is expected that MT will

help maintain a high level of translation quality by helping translators cope with their workload in the available amounts of time. The continuous increase of translation requests could, in theory, have an impact on the quality of translated documents. Nevertheless, this cannot be allowed for legislative documents as it will most certainly affect the whole legislative procedure.

Furthermore, MT may contribute to a better value for money of translations particularly by reducing the cost of translation outsourcing per outsourced page. The overall expenses for external translations may not decrease but possibly lower charges for machine translated segments may provide an opportunity for more documents to be outsourced.

Unlike TM, Machine Translation does not include references to the source of translations. TM meta-data indicate which document a proposed translation comes from, which legislative procedure it is linked to, when it was produced etc. The lack of this information in our current MT implementation will have an impact on post-editing time even if the MT output is linguistically perfect in particular in the case of legislative documents. This is mainly due to the fact that our translators are obliged to re-use the exact same translations that have been produced in other documents which are being referred to in the current source document. If the source of a machine translated string is unknown then the translators will have to spend some time controlling the origin of certain translation suggestions and this is a risk with a direct impact on the above mentioned expected benefits.

3.2 Project deliverables

The main deliverable of this project is an MT solution for more than 700 in-house translators and 506 language pairs - from and into all EU official languages. The quality of the MT is expected to be good enough to reduce translation time in all language combinations while there is also a use case for raw MT for gisting purposes (without or with minimal revision). In this case the MT output is expected to be of understandable but not necessarily of human quality.

Synchronous (real-time) MT services are currently out of this project's scope. Machine translated segments will be incorporated in the translation memories and offered as part of a pre-

---

[2] One "standard page" consists of 1500 characters

translation package. Pre-translation packages are prepared and provided to translators before the beginning of a translation task and nowadays they usually include translation memory segments relevant to their working document. Real-time MT would require a significantly higher investment on hardware resources to achieve much faster decoding times.

3.3 Data

European Institutions have established a close collaboration framework in the area of translation technologies. The first step was taken with Euramis (Blatt 1998), a huge translation memory with almost 300 million segments available to different EU institutions. Thanks to Euramis the Council, the Court of Auditors, the Court of Justice, the Committee of the Regions, the European Economic and Social Committee, the Parliament and the Translation Centre for the Bodies of the European Union can contribute to each others work by adding their own translation segments.

Along with translation memories the EP has also important amounts of documents in its archives as well as on its web-site many of which are not included in the translation memories. These resources are being collected and parallelised to be used for MT purposes. In the future external corpora that have not been produced in-house should also be incorporated.

4 MT evaluation

To estimate the expected benefits described in section 3.1 the EP is conducting a large scale MT evaluation for the first time in its history relying on the contribution of 62 in-house translators. The main conclusions we expect to draw concern MT quality, MT comprehensibility and MT editing time compared to translation time. The test users work with a web-based evaluation tool which was initially used for ACL's WMT workshop and described in (Callison-Burch et al. 2009) and configured in-house to meet our own specifications. The MT solution that is being tested at this stage is the one developed by the European Commission which is described in (Eisele et al. 2011). This solution has been chosen in the context of interinstitutional collaboration which started in 2009 in the MT field and it is a statistical MT

system based on Moses (Koehn et al. 2007).

4.1 Methodology

For the selection of the evaluation methodology the MT@EP project team has collaborated with a user group that has been created for this purpose. The participants of the user group are mainly representatives of the business (translators), one business analyst and one computational linguist with many years of experience in MT.

First of all the document types were carefully selected as MT seems to be more appropriate for some than for the others. Legislative documents were left out of this process because lacking the source documents of MT-translated strings translators would not be able to evaluate or post-edit the MT output as required by the testing specifications. Therefore documents with more free text, less quotes and of diverse domains and language registers were chosen.

Translation demand was another parameter that was taken into consideration when the test corpus was selected. Therefore, document types more frequently translated than others have been selected.

4.1.1 Categorization and Error Detection

To evaluate MT quality the test users are provided with segments in the source language and their machine translated equivalents and they are asked to mark them as Excellent, Good, Medium or Poor. Test instructions provide precise definitions of those marks to make sure that the test users take common criteria into consideration to the extent that this is possible. Here we used the categories used by (Roturier 2009). More precisely the test users were provided with the following definitions:

Excellent MT Output: Your understanding is not improved by the reading of the source because it is syntactically correct; it uses proper terminology; the translation conveys information accurately.

Effect: No post-editing required.

Good MT Output: Your understanding is not improved by the reading of the source even though the MT segment contains minor errors affecting any of these: grammatical (article, preposition), syntax (word order), punctuation, word formation

(verb endings, number agreement), unacceptable style. An end-user who does not have access to the source text could possibly understand the MT segment.

Effect: Only minor post-editing required in terms of actual changes or time spent post-editing.

Medium MT Output: Your understanding is improved by the reading of the source, due to significant errors in the MT segment (textual coherence / textual pragmatics / word formation / morphology). You would have to re-read the source text a few times to correct these errors in the MT segment. An end-user who does not have access to the source text could only get the gist of the MT segment.

Effect: Severe post-editing is required or maybe just minor post-editing after spending too much time trying to understand the intended meaning and where the errors are.

Poor MT Output: Your understanding only derives from the reading of the source text, as you could not understand the MT segment. It contained serious errors in any of the categories listed above, including wrong Parts Of Speech. You could only produce a translation by dismissing most of the MT segment and/or re-translating from scratch. An end-user who does not have access to the source text would not be able to understand the MT segment at all.

Effect: It would be better to manually retranslate from scratch (post-editing is not worthwhile). Moreover the participants have the option of selecting among some basic types of errors in the MT: syntax, wrong lexical choice or idioms, incorrect form/grammar, wrong punctuation, wrong spelling/typo/numbers, style/register. We didn't provide a more detailed error classification because at this stage we prefer receiving more evaluation data than feedback on specific error types.

There can be cases where the MT output is fluent but it is not clear to the translator if it conveys the message of the original text for the simple reason that often the original text may be incomprehensible (badly formulated or out of context). Therefore, the test-users are able to mark a bad original as such.

To evaluate the comprehensibility of MT and its appropriateness for gisting purposes a next task offers the test users a paragraph in the source language with its MT target. In this task test users only need to state if the translation conveys the meaning of the original text or not. They are also given the option to select "bad original". To make sure that users would not abuse the latter in order to proceed to the next segment they still have to state if the MT output conveys the meaning of the original text instead of proceeding directly to the next one. In the opposite case test-users may feel tempted to skip the most complicated cases or paragraphs containing long sentences. This is the only task where paragraphs are provided instead of segments because context is often necessary to understand the information contained in a single sentence.

Figure 1: Categorization and error detection task

Assess whether the meaning is conveyed

Source (document type: *RE*): having regard to its resolution of 30 November 2006 on Implementing the Community Lisbon Programme: small and medium-sized enterprises SMEs policy for growth and employment,

Un-edited MT Output:
Vu sa résolution du 30 novembre 2006 sur la mise en œuvre du programme communautaire de Lisbonne: Les petites et moyennes entreprises, notamment la politique en faveur des PME pour la croissance et l'emploi,

Does the translation convey the meaning of the original paragraph?
- ○ Yes
- ○ No
- ☐ Bad original

Annotator:	**Task:** EN-FR Editing	Annotate

Answer **yes** or **no** with the buttons above to the question whether the translation conveys the meaning of the original paragraph?

0.00 out of 346.23 available pages have been processed for your language pair of which you have processed 0.00. Minimum number of pages to be processed for this task is 20.00 per language pair.

Figure 2: Paragraph assessment for gisting purposes

4.1.2 Post-Editing and Translation

Approximately 80% of the paragraphs displayed in the previous task are machine-translated. The purpose of the post-editing task is to edit the MT output until it's considered to be of publishable quality. If the MT output is already of publishable quality users select "Perfect Translation, no editing needed". The post-editing time is measured from the moment that the page is loaded until the end of the last action taken (editing, selection of a radio-button etc.).

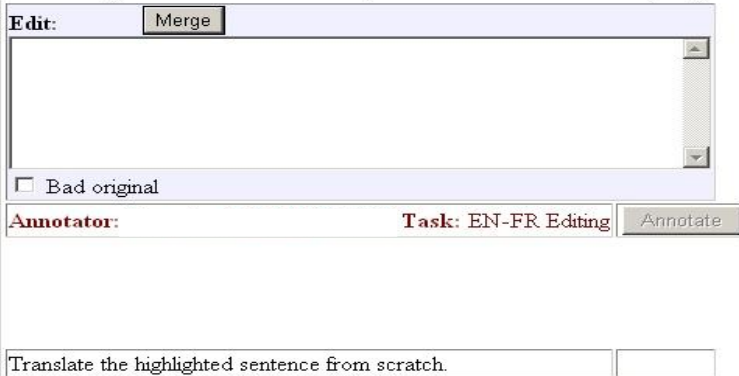## Evaluation, post-editing and translation

Source: (document type: *RE*) having regard to its resolution of 30 November 2006 on Implementing the Community Lisbon Programme: small and medium-sized enterprises SMEs policy for growth and employment,

Un-edited MT Output:
Vu sa résolution du 30 novembre 2006 sur la mise en œuvre du programme communautaire de Lisbonne: Les petites et moyennes entreprises, notamment la politique en faveur des PME pour la croissance et l'emploi,

Edit: [Merge]
Vu sa résolution du 30 novembre 2006 sur la mise en œuvre du programme communautaire de Lisbonne:

Reset Edited Sentence
- ○ Perfect Translation, no editing needed
- ○ Edited
- ○ Not editable, translating from scratch would be faster
- ☐ Bad original

Figure 3: Post-editing and translation task

Translate the sentence

Source: (document type: *DV*) Whereas, for some committees, the first half of 2011 was characterised by culminating efforts to finish "preparatory works" ahead of the release of major legislative files, other committees were already in the full process of "testing" Parliament's newly acquired prerogatives. In the second half of 2011, the release of a number of proposals for major policy reforms in the leading areas kicked in the start of the true legislative work in several committees (MFF, cohesion policy reform, CAP reform, CFP reform, Rules governing the trans-European Networks reform, etc.). In the legislative field, several important results were achieved, such as the economic governance package ("six-pack"), the Directive on combating sexual abuse, sexual exploitation of children and child pornography, the single application procedure for residence and work, new legislation on road safety, on clear food labelling, on consumer rights and on cross-border healthcare, to name but a few. The SURE and CRIS special committees wound up their activities in June and July respectively.

Edit: [Merge]

☐ Bad original

Annotator: **Task:** EN-FR Editing   [Annotate]

Translate the highlighted sentence from scratch.

Figure 4: Translation task

Measuring post-editing time is certainly not enough to estimate the possible benefits of MT. 20% of the paragraphs displayed are not followed by MT output. The segments of these paragraphs have to be translated segment by segment from scratch to obtain reference values for each participant. Subsequently the translation throughput (words per hour) of one translator will be compared to her post-editing throughput. By "translation from scratch" we mean that no MT output is provided. Nevertheless, translators are able to use all the tools they usually have access to in their normal workflow. For obvious reasons the only resources they are not allowed to access are translation memories or documents that can provide them with complete translations of the segments displayed in the test application. To avoid possible bias towards post-editing or translation from scratch, in both cases translators are given access to the same tools and references. These tools are briefly presented in section 2.

4.2 What will be measured

The results of each test will be analysed separately for each language pair. The data collected from the categorization task will help us measure the quality of the tested MT solution at segment level. For this purpose the number of Excellent, Good, Medium and Poor segments will be reported whereby different segment lengths (short, medium and long) will be taken into consideration. To make sure that the results are consistent, intra- and inter-annotator agreement will be taken into consideration. This is possible thanks to the regular re-appearance of segments within the evaluation application. Intra-annotator agreement will be measured using the Kappa coefficient (Callison-Burch et al. 2012) and inter-annotator agreement will be estimated using the Fleiss kappa as presented in (Fleiss 1971).

In the post-editing task the time needed to post-edit a segment is the most important variable. This will be measured from the moment that a new segment is loaded until the last action on the page is taken. This action (editing or selection of radio button etc.) is not defined a priori because the test-users might select any sequence every time. Translation time is measured in the same way at the translation. Segments that appear in the post-editing task may not re-appear in the translation task. If a test-user encounters a sentence at the post-editing task and then is asked to translate it from scratch in the translation task there is no doubt that she will remember it and therefore

translate it faster. Average post-editing and translation times per character may also be compared.

It is expected that users will adapt to the application as well as to the post-editing task itself. Therefore we also intend to measure individual change of post-editing speed over time taking into consideration each user's familiarity with the task.

To evaluate the current MT solution as a tool for gisting purposes we will compare the number of machine translated paragraphs that convey the meaning of the original text compared to those that do not.

4.3 Evaluation Data

Translation demand was the main criterion for the selection of the language pairs that are currently evaluated. As the vast majority of source documents are written in English test users were provided with data that have been machine translated from English to all official EU languages. The English translators have been provided with data translated from French to English and from German to English. French and German are the two other of the so called "pivot" languages. Although most translation units in the EP have translators that cover a very big number of languages (some of them master 6 languages or some times even more), there are certain language combinations that are very rare. For example when a document is drafted in Maltese and it has to be translated in Lithuanian it is not very likely to find an in-house translator who is able of translating between these two languages. The same is the case for other target languages of course. Therefore, many documents are translated into the three pivot languages first which are mastered by the majority of translators and subsequently into all official EU languages.

To gather a sufficient amount of data without increasing too much the translators' workload at the same time a total amount of 40 pages will be processed per language pair. Two or three translators have been made available for each language pair and they have two and a half months to accomplish the task.

4.4 Preliminary Results

At the time when this paper was written two translators had accomplished their categorization

task and another 6 had reached at least 50%. The current results are summarized by language pair in Table 2.

Language Pair	Poor	Medium	Good	Excellent
EN-PL	25 %	30 %	34 %	11 %
EN-DA	4 %	17 %	51 %	29 %
FR-EN	34 %	14 %	16 %	36 %
EN-LT	50 %	30 %	10 %	10 %
DE-EN	48 %	13 %	17 %	21 %

Table 2: Preliminary results of segment categorization by language pair

With maximum two users for each language pair having completed in most cases roughly 60% of their categorization task these results can merely show a certain trend: at least 50% of all segments evaluated for each language pair are of medium quality and thus post-editable with this percentage reaching up to 96% for English to Danish. At this stage the used MT system seems to provide less usable results for EN-LT while according to direct feedback from the English evaluators DE-EN is rather problematic too with many results being of very poor quality. It should be added here that the two English evaluators that worked on DE-EN and FR-EN have accomplished their categorization task.

As expected these results are not consistent for all document types. For example 92% of segments coming from QO documents (oral questions) were judged as poor while other document types had much fewer or some times no segments at all judged as poor. Two possible reasons for the high number of poorly translated QO segments are data scarcity (not many QO documents in the training data) as well as the style and register used in these documents which is totally different from any other document type. So far most evaluators have shown a high intra-annotator agreement.

5 Conclusions

In this paper we described the evaluation methodology and some preliminary results of a state of the art statistical MT system at the European Parliament. With the use of post-editing and other MT evaluation tasks fine-tuned to our business needs we will use the collected data to

estimate the benefits that DG TRAD may have from the implementation of MT technology in the current translation workflow as a complementary tool to Translation Memories, terminology databases, bilingual concordancers and other reference tools.

6 Future work

After the end of the current evaluation exercise we will try to use the collected data to estimate the expected business benefits.

The conclusions that will be drawn from this evaluation procedure will be used in the future as a baseline to avoid re-running similar exercises too often as they require the involvement of many human resources. Future evaluations will most probably ask the users to compare the output of the future MT engines to that of the current ones. A more detailed manual error-analysis will also be conducted to identify key areas of MT improvement. One such example could be specific grammar errors in morphologically reach languages which may be solved with language-specific rules.

The analysis of the annotation data will also help us understand our needs for post-editing training and come up with more precise specifications.

In the future we expect to integrate MT in the translation workflow in such a way that similar conclusions will be drawn in the real translation environment without creating extra work for translators. Creating this translation-feedback loop we expect to get more reliable results as our current method is similar but not identical to real translation conditions.

Acknowledgments

References

Achim Blatt. 1998. EURAMIS: Added value by integration. In T&T Terminologie et Traduction, 1.1998, pages 59–73

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In Proceedings of the Fourth Workshop on Statistical Machine Translation, pages 1–28, Athens, Greece, March. Association for Computational Linguistics

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut and Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation Proceedings of the Seventh Workshop on Statistical Machine Translation

Andreas Eisele and Caroline Lavecchia. 2011. Using statistical Machine Translation for Computer-Aided Translation at the European Commission. JEC 2011

Michael Jellinghaus, Alexandros Poulis and David Kolovratnik. 2010. Proceedings of the Fifth Workshop on Statistical Machine Translation

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin, vol. 76, No 5, Pages 378-382

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. MT-Summit 2005

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In Proc. of ACL Demo and Poster Sessions, pages 177–180

Philipp Koehn and Jean Senellart 2010. Convergence of Translation Memory and Statistical Machine Translation. AMTA Workshop on MT Research and the Translation Industry

Johann Roturier. 2009. Deploying novel MT technology to raise the bar for quality: A review of key advantages and challenges. In The twelfth Machine Translation Summit, Ottawa, Canada, August. International Association for Machine Translation

Michel Simard & Pierre Isabelle. 2009. Phrase-based machine translation in a computer-assisted translation environment. MT Summit XII: proceedings of the twelfth Machine Translation Summit, August 26-30, 2009, Ottawa, Ontario, Canada; pp.120-127

D. Theologitis. 1997. EURAMIS, the platform of the EC translator. In EAMT Workshop, pages 17–32