Exploring Translation Memory for Extensibility Across Genres: Implications for Usage and Metrics[1]

Carol Van Ess-Dykema	Susan P. Converse	Dennis Perzanowski	John S. White
U.S. National Virtual Translation Center	U.S. National Virtual Translation Center	U.S. Naval Research Laboratory	The MITRE Corp. U.S. National Virtual Translation Center
carol.j.vaness-dykema@ugov.gov	susan.p.converse@ugov.gov	dennis.perzanowski@nrl.navy.mil	john.s.white@ugov.gov

Abstract

Translation memory (TM) has successful applications in certain genres, such as system documentation, which contain repetition within or between versions of a document. The U.S. National Virtual Translation Center (NVTC), with a charter to provide translations of a very wide variety of genres, has undertaken a pilot investigation toward finding metrics to determine whether TM is useful for translators in multi-genre enterprises, and also whether less repetitive genres can derive benefit from using TM.

Introduction

Translation Memory (TM) is a popular computer-aided translation application for many translation enterprises. Available at a variety of levels of capabilities, flexibility, and cost, TM can help both translator and quality control professional develop translations with improved consistency, shared knowledge, and faster turnaround.

TM has shown its best performance in translating document sets with large amounts of recurring text, either across the breadth of a document (e.g., bibliographic footers on journal articles) or across a document's history (e.g., unchanged passages in an updated system manual). Less is known about the benefit of TM for translating other genres, such as transcripts of audio interviews or email traffic. Passages in such genres will likely be less recurrent than in system manuals, yet there may be significant potential for TM in these genres nonetheless. Thus there is a need among translation stakeholders to determine the attributes of texts that make them amenable to the potential benefit of TM.

Apart from the nature of the texts being translated, there is a more general question about TM, namely whether the technology constitutes a significant step forward in actual operational use, over more traditional resources and processes. On the one hand, translators may benefit from embedded databases, search, and character handling tools in a TM system, even when the matches themselves are not useful (Lagoudaki 2006; Gow 2003). On the other hand, TM systems may fail to take

[1] The research from which this paper derives was sponsored by the Foreign Language Program Office of the U.S. Office of the Director of National Intelligence. The authors gratefully acknowledge the contributions of Tucker Maney and Rachael Richardson to the research and to this paper.

advantage of a fuller measure of the translator's knowledge (Macklovitch and Russell 2000). Thus there is a need to gauge the plusses and minuses of TM usage, aside from the presence of domain/genre-relevant TM banks per se.

From these two perspectives, it is evident that metrics are called for which can be used consistently to gauge the usefulness of TM in the U.S. National Virtual Translation Center (NVTC) workflow, and more specifically, its performance in important genres other than those known to be amenable to TM treatment. These metrics must be discrete, repeatable, and above all relevant to capturing the observed behavior of aspects of the TM tools and processes.

NVTC, along with the U.S. Naval Research Laboratory (NRL), has conducted the first in a series of experiments on the usability and effectiveness of TM for non-assessed genres. A variety of different measures have been captured, perhaps most importantly the assessment by professional translators, and indirectly by quality control professionals, of the net benefit of translation memory for the translation process.

This paper is a report of the methodology, the execution, and analysis of this pilot experiment. Capturing the actions and assessments of professional users, the experiment measured the potential contribution of a TM system to a series of controlled translation and quality control activities in Russian, Arabic, and Chinese into English. We describe the experimental processes and results, and what they tell us about the application of new metrics for experiments already in planning. We will also discuss the next set of objectives, as well as collaborative partners who have formed a team with NVTC to help determine a future vision for the role of TM in a range of translation workflow scenarios.

Background

NVTC produces high-quality translations for the U.S. Department of Defense and the Intelligence Community. The Center has translated in more than 70 languages, in many critical subject domains, over more than 80 genres. The NVTC mission has resulted in a rapid growth in the quantity and range of coverage. This requires a commitment to the right configuration of the best translation technology at the relevant points in the translation process, to have a positive impact on enhancing translator effectiveness.

NVTC translators work both onsite and offsite, connected virtually to a translation management system, and eventually to a shared tool space. Its ability to surge and shift to meet emerging needs in new domains and languages makes NVTC unique in U.S. government translating, but also lends a number of challenges to technology and process. In particular, a partial dependence on external translators results in idiosyncratic, personal translation tools and processes, which, though flexible and rapidly responsive, result in a loss in consistency and a failure to gain domain and genre knowledge enterprise-wide.

The NVTC's emerging model architecture addresses these issues while keeping the flexibility and responsiveness of the current tasking configuration (Van Ess-Dykema

et al., 2008). In large measure this will be accomplished through automatic and semi-automatic feedback loops for augmenting the machine and human knowledge and performance.

NVTC Envisioned Workflow

The NVTC model under development incorporates the optimum configuration of translation technologies, approaches, and systems at the front end of the workflow. The Center receives source language materials in any of a variety of modalities (online text, audio, document images, etc.). In Figure 1, one or more of several sequences of integrated special lexicons, translation memory, and machine translation will produce a translated artifact, which is vetted through the human role of the "Paralinguist." (Van Ess-Dykema et al. 2008, Van Ess-Dykema et al. 2009; Barabe 2009)). Based on the output and the human judgment, the appropriate action is chosen for the document (re-translation or post-editing), along with the appropriate translator/post-editor. Translation support will include contemporary collaboration tools and wikis, as well as asynchronous interaction with the quality control professional.

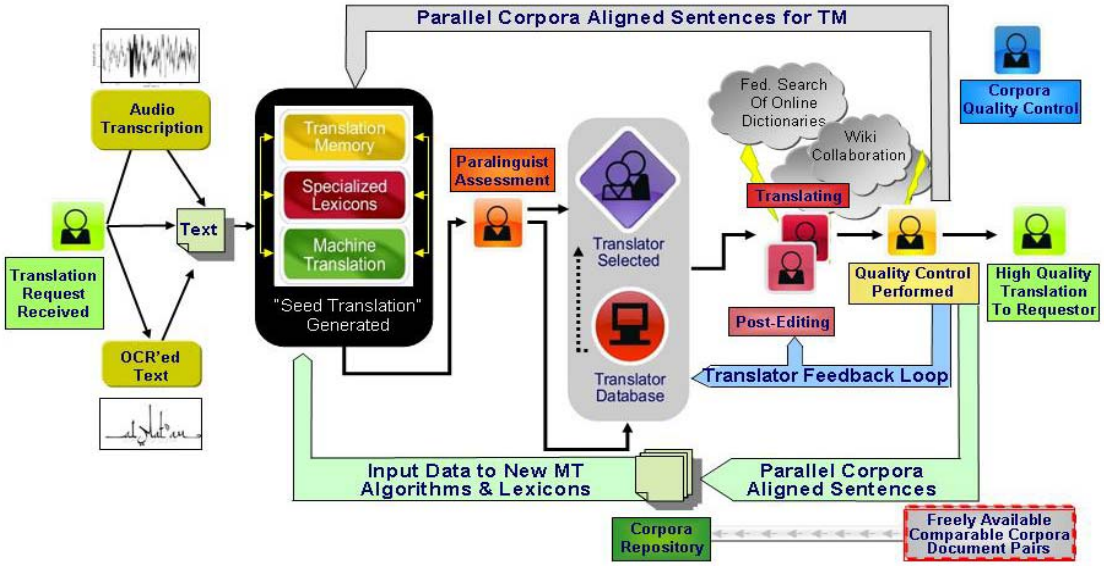


Figure 1. Envisioned translation workflow mediates front-end automatic translation processes with appropriate human roles, while feeding back translation outputs to continuously optimize the automated components.

The feedback subsystems are key to the success of the model. Each of these is designed to improve the functionality of a major component of the workflow, both immediately at translation time, and incrementally as bodies of translation products and judgments are captured and fed back into training models for adaptive algorithms. Of particular note to this discussion is the feedback to the translation memory; the accuracy and quality of the automated subsystems, especially TM, will be enhanced by

the continuous flow of aligned pairs of source and target documents translated at the Center.

The result of the envisioned end-to-end process is a rapid capability for high-quality full translations of a large variety of languages, domains, and genres. The capability ensures consistency of terminology, style, and format while maintaining the flexibility of the distributed translation concept. The operation of the translation workflow itself will ensure improvement in consistency, accuracy, and throughput via the feedback mechanism; this shared property of translation memory is the driver for the NVTC Pilot Study.

Pilot Study Procedure

The team developed a range of measures for gauging the conduct and performance of both translators (who used the TM tool) and quality control experts (who did not). Qualitative measures included opinion parameters embedded into carefully constructed questionnaires for a manual translation exercise, the TM exercises, and the quality control (QC) exercise. A series of quantitative measures focused on translator time (e.g., keystrokes/second), number and size of TM matches, and the number and nature of QC corrections.

For the purposes of the pilot experiment, we proceeded with these hypotheses:

- ***Hypothesis:*** *that translations aided by a TM will be better, in terms of quantitative and qualitative measures, than translations using only a dictionary;*
- ***Hypothesis****: that translations performed with TMs that have been optimized for a specific genre will be better than those translated with a generic TM.*

Translator interaction with a TM system

The pilot experiment began in the spring of 2009, at a site hosting a representative commercial TM system (Van Ess-Dykema et al., 2009). Each of 6 translators in the experimental sessions performed a manual translation of a document, using his/her own professional processes, but with a control for access to resources: ordinary desktop applications were available, but dictionary lookup was restricted to a specific hardcopy bilingual dictionary and (where available) a hardcopy of a specialized domain dictionary. After translating, the translator completed a questionnaire regarding the manual translation and test conditions. This translation and questionnaire served as controls for the experiment.

Each translator was then given training in the TM system, and subsequently used the system to translate three documents of comparable size, one on each of three days. After the first and third translations using the system, the translator answered additional questionnaires, each eliciting reactions to the test conditions and the interaction with the TM system. Each translator also had the opportunity to provide comments of a general or specific nature pertaining to aspects of their reaction to carrying out the translation tasks using the TM system.

During these tasks, a number of quantitative usability metrics were captured, including mouse clicks, tool use, and updates to the translation memory bank.

Quality control (QC) interaction with TM-assisted translations

Six QC professionals were recruited to edit the translations produced in the translation portion of the experiment. Each QC professional worked from his/her ordinary workplace, apart from the TM experiment site, and was not told which translation was produced with TM and which was not. After editing each document (using a common word processing application with automatic edit tracking turned on), the QC professional completed a questionnaire regarding the task.

Observations from Pilot Study

As a pilot study, we used a range of qualitative and quantitative measures that appear to be relevant to the usefulness and efficiency of the TM system. The initial objective, given the relatively small sample (24 translations) and subject size (6 translators and 6 quality control experts), was to note which of these measures appear to provide salient metrics, sufficiently germane to the hypotheses and sensitive enough to indicate some effects even with a small sample. We have identified such measures that meet those characteristics, and subsequent experiments will rely heavily on these for determining the applicability and effectiveness of TM. Table 1 gives an overview of the data, measures, and metrics used in the evaluation of the pilot study data.

Table 1. Measures and metrics showing promise in pilot study.

Measurement method	Data source	Subjects	Type of Data	Potential metric
Translation memory matches	TM system logs	Translators	Quantitative	Match length per language per TM Bank
Time translating	Keystrokes, off-line time	Translators	Quantitative	Keystrokes per second
TM bank updates	Private TM banks created during translation sessions	Translators	Quantitative	Match length per language
QC corrections	QC mark-up of translation	QC experts	Quantitative	# Corrections per error type
Questionnaires	Question responses	Translators	Qualitative	Value change in Perceived Ease of Use
Questionnaires	Question responses	QC experts	Qualitative	Value change in Quality Control Accuracy
Written comment	Comments on aspects of process and system	Translators	Qualitative	Observations by topic

Quantitative data

As noted, translators were timed for total time spent translating each document, time spent actually using the system, and break time. Their mouse clicks were recorded by running key-logging on the TM session, and the matches of source strings to TM bank strings were captured from internal system logs.

We measured QC professionals by the amount of time it took them to correct a translation, the number of corrections they made per document, and the correction type (e.g., word error, syntax error, etc.).

Translators' Quantitative Results

TM interactions

The number and extent of TM matches occurring during a translation session are important metrics to capture as a direct measurement of suitability and also as a means of baselining other measurement such as translation time and QC corrections. Such factors as length of the matched unit (a long phrase versus a single word, for example) indicate the suitability of the TM bank for the translation task at hand. Thus measuring the difference between translation sessions with only a system-provided TM bank (Russian) and the optimized TM banks (Chinese and Arabic) directly speak to the hypothesis that genre-relevant population of TM banks improve their effectiveness. Also, the expected trend toward improvement of the TM-assisted translations as the sessions progressed should be evidence of the increasing value of the personal translation memory (that automatically enters the source-target segment pairs in the TM system as the translation proceeds).

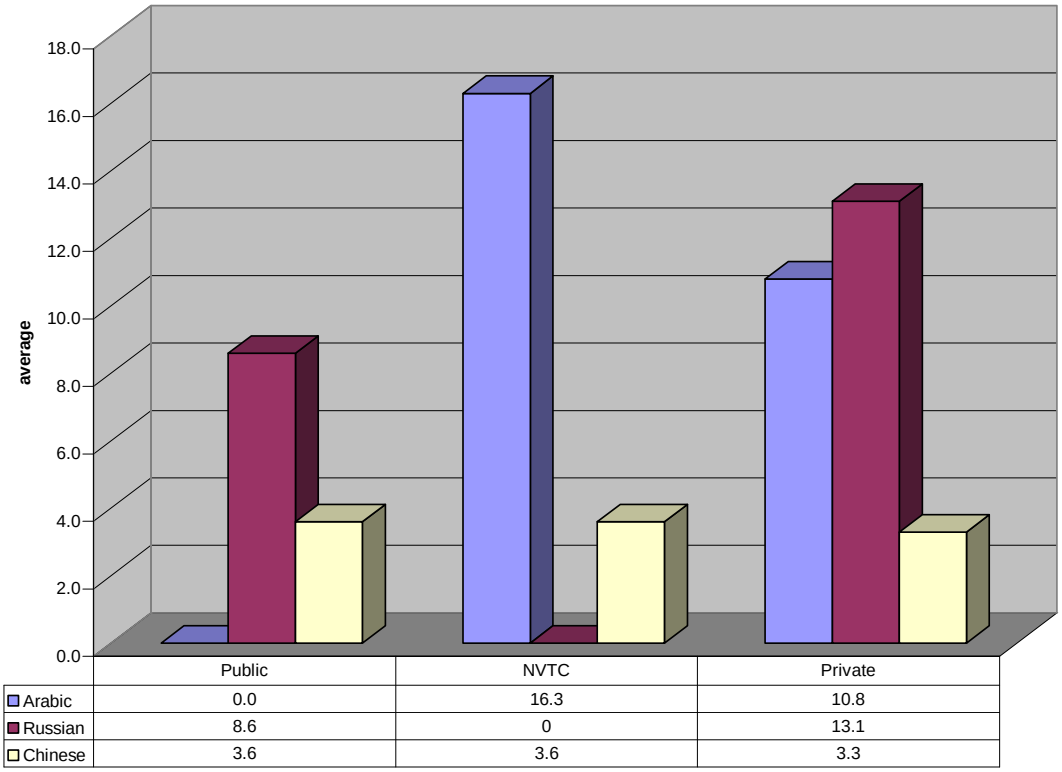
Different TM banks were simultaneously available during the TM session, which matches typical operational practice. The "public" TM bank was provided with the system itself, and thus was of a general distribution of domains and genres. The "NVTC" bank was created from open source, aligned translations of broadcast transcripts, and was intended to optimize the TM behavior by domain and/or genre; the Chinese and Arabic sessions had this bank, but the Russian did not. The final, "Private" TM bank was created in real time during the translation process with the translator-finished passages of text. Table 2 provides brief definitions of each of the TM banks employed.

Table 2. Different TM banks used in the experiment.

TM Bank Type	**Definition**
Public	The TM bank provided with the product; generic in content and small in size
NVTC	The TM bank built from aligned bilingual corpus data, corresponding in domain and genre with the texts to be translated
Private	The TM bank generated from the translation session itself, used to improve the matches in current and subsequent translation tasks

The Arabic and Chinese TM banks, overall, matched a greater percentage of source text than did the Russian, which accords with the expectation for optimized vs. non-optimized systems. Thus, the maximum length match among all the hits returned for a segment, as a percentage of source segment length, was on average about 20% for both Arabic and Chinese, but only 7% for Russian. The average maximum length was about 9% from the translators' private memory banks, vs. 5% from the vendor-provided memory bank.

Figure 2 shows the average length for TM matches from each of the three memory banks.



	Public	NVTC	Private
Arabic	0.0	16.3	10.8
Russian	8.6	0	13.1
Chinese	3.6	3.6	3.3

Figure 2. Average length of TM matches by language and TM bank. "Public" refers to the TM bank supplied with the product; "NVTC" is the TM bank built for the experiment; and "Private" is the TM bank created in the course of the translation task.

These observations indicate that populating the TM banks with domain and genre segments pertinent to the translation task will result in more numerous and useful matches.

Translation rate

Among the several approaches to identifying a salient quantitative metric for translator performance with and without the TM tool, one that bears further exploration is keystrokes/second. The measure appears to be sensitive to optimized vs. generic memory banks; in particular, to the critical question of whether a TM bank optimized for the genre and/or domain of the documents to be translated will perform better than one which only uses general-purpose memory banks.

One such measurement across the three languages appears to be promising, but with a substantial caveat. A comparison of the average keystrokes/second of the Chinese and Arabic (both having optimized TM banks) shows them to be consistent with each other, and higher than the average score for the Russian texts (that only had the default generic TM bank).

Figure 3 shows this comparison, and also a competing hypothesis for the difference in average keystrokes/second: one Russian translator (a target-language native) scored significantly higher than the other (source native), and indeed the target-native scored similarly to the average Chinese and Arabic scores. The competing hypothesis, then, is that individual translator differences may be the source of the experimental variance, and not the optimization of the TM banks.

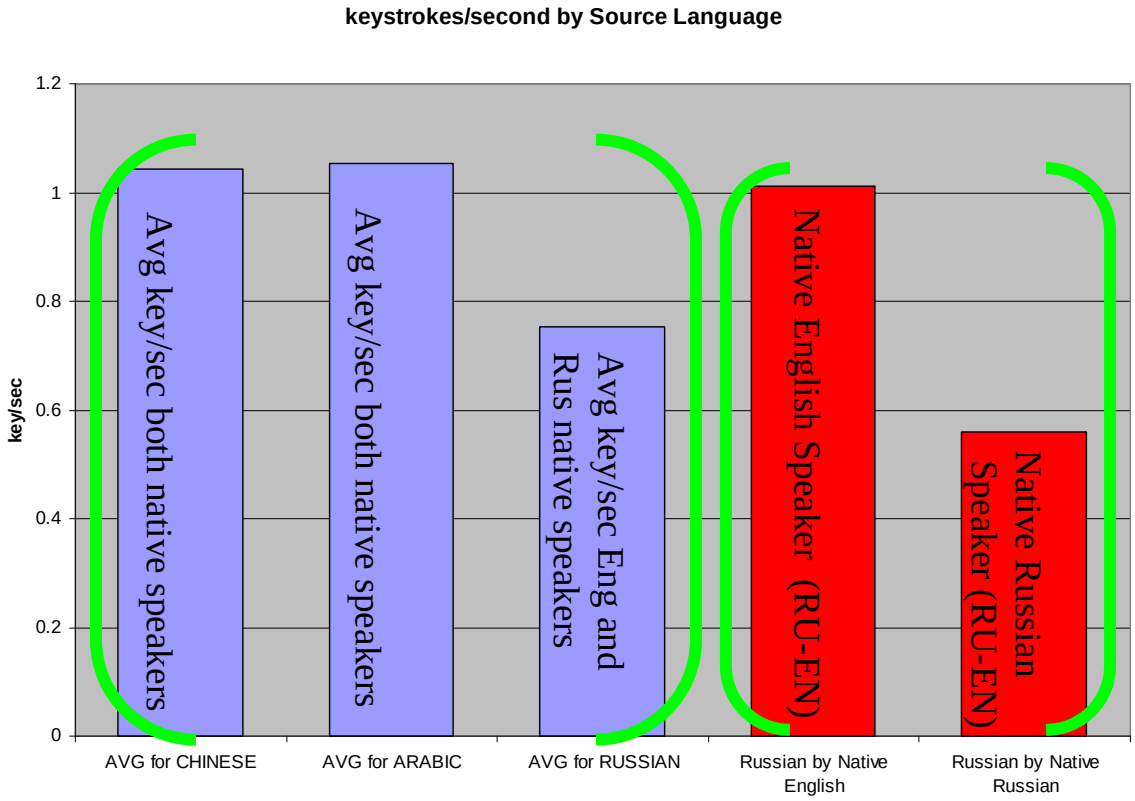


Figure 3. Average keystroke/second scores (Blue columns) seem to show a difference between using optimized TMs (Chinese and Arabic) and generic TM (Russian). But target native/non-native differences may be the real source of variance.

These results demonstrate the need to control for the fluency of the translator in the target language, so that the actual effects of using an optimized TM bank can be discerned from the data.

In subsequent experiments, every effort will be made to standardize the target-fluency of the translators (for instance, that they are all target-native), while continuing to use the key-logger as a quantitative measure.

Quality Control Quantitative Results

While human factors obscure variations in the quality of the translations that can be attributed to TM, a preliminary analysis of the types of corrections that the quality control professionals made to the TM-assisted translations suggests focal points for future investigation.

The edits that the quality control professionals made were viewable in the Track Changes function of a common word processing application.

The kinds of edits that the quality control professionals made were categorized by edit type using a quality coding method inspired by the Society of Automotive Engineers (SAE 2001).

The categories used for the corrections are as follows:

- lexical (**lex**): changes to single words or phrases
- proper name (**pname**): changes to words or phrases comprising proper names
- syntax (**synt**): changes to word order; or splitting/combining sentences
- omission or addition (**add**): correcting omissions
- morphology (**morph**): correcting inflectional, derivational, and agreement errors
- determiners (**det**): correcting, adding, or deleting determiners
- misspelling (**spell**): changes to spelling
- punctuation (**punct**): changes to punctuation.

All errors were categorized as either minor (style or fluency) or serious (meaning-changing).

Using the Arabic-English data as representative, a comparison of the types of corrections made to the TM-assisted translations on the last day of TM use to the types of corrections made to the control "hand" translations done the first day shows a similar distribution of edit types, with word choice (**lex**) corrections predominating (35%-40%).

Figures 4 and 5 below show, respectively, the corrections for the Arabic control translation (without TM) and the translation of the last session (with all available TM banks).

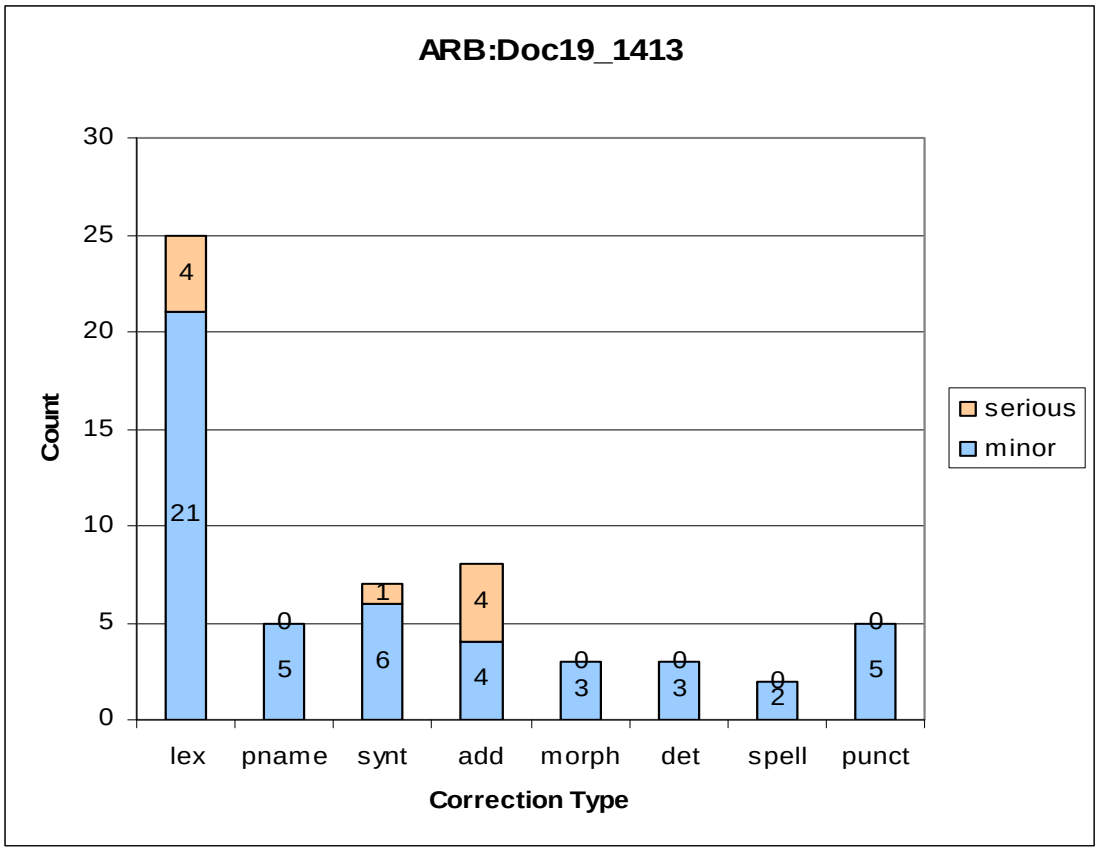


Figure 4. Correction results for Arabic translation without TM

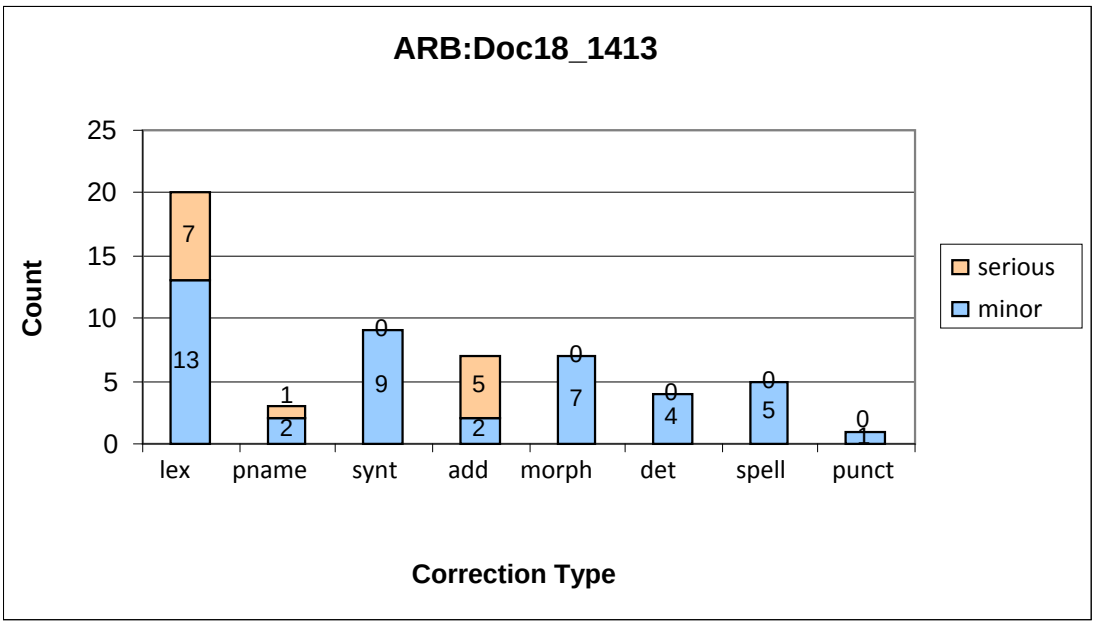


Figure 5. Correction results for Arabic, last translation with all TM tools

The results suggest that the total number of correctable errors declines with the increase in TM usage, for lexical phenomena (**lex** and **pname**). This measure will be

of value going forward, as long as certain other sources of variance (individual stylistics, intrinsic differences in the documents, etc.) are controlled.

Qualitative data

Translator Qualitative Results

For this study, we investigated six variables to evaluate Translator's performance and satisfaction with the TM tool. These variables are identified in Table 3. Each item in the questionnaire was intended to contribute to one of the variables used in the analysis of the results data, either directly or through its inverse value when stated in the negative for control purposes. The validity of these variables was verified through the Pearson Bivariate two-tailed test.

Table 3. Translators' variables used in the study

Variable	Questionnaire	Questionnaire items
Perceived Translator's Performance (PTP)	Session Evaluation Questionnaire	1. I completed this translation quickly. 2. I completed this translation easily. 3. I completed this translation efficiently.
Perceived Translators Quality (PTQ)	Session Evaluation Questionnaire	4. I had the resources I needed to produce a translation accurate in meaning. 5. I had the resources I needed to produce a translation with proper grammar. 6. I had the resources I needed to produce a translation employing the proper style. 7. I had the resources I needed to produce a translation of good quality.
Perceived Ease of Use (PEU)	Software Evaluation Questionnaire	2. I felt in control while using the system. 3. I found using the system inconvenient. 4. I felt confused while using the system.
Perceived Information Retrieval Performance (PIR)	Software Evaluation Questionnaire	5. The translations provided by the system were relevant. 6. The translations provided by the system were easy to use. 7. I did not find the translations provided by the system useful.
Perceived System Utility (PSU)	Software Evaluation Questionnaire	8. I did not find the translations provided by the system useful. 9. I would like to use this type of system in the future. 10. I would recommend this system to others. 11. Using this software is a waste of time.
Perceived System Performance (PSP)	System Comparison Survey (Items 1, 2, 4, and 6)	1. I preferred translating with only a dictionary to using the Translation Memory system. 2. The Translation Memory system improved the speed of my translation work. 4. I felt more efficient while using the Translation Memory system than I did working with a dictionary alone. 6. I felt more confident in my work using the Translation Memory system than using the dictionary alone.

Table 3. Translator questionnaire items relevant to qualitative variables for measuring translator experiences with the TM system.

We summarize the results for one of the variables here, Perceived Ease of Use (PEU), as indicative of a qualitative measure that holds promise for future experiments.

PEU was computed from the questions on the Software Evaluation questionnaire:

Item 2. I felt in control while using the system.
Item 3. I found using the system inconvenient.
Item 4. I felt confused while using the system.

The variable derived from these three items addresses Translators' comfort with TM tool usage. Analysis of the data shows that translators' opinions of the tool were generally more positive than neutral, indicating the potential sensitivity of this measure for future TM experiments. Figure 6 is a box-and-whisker chart showing the change in PEU values from the 2nd and 4th translation sessions (the first and last TM usage, respectively).

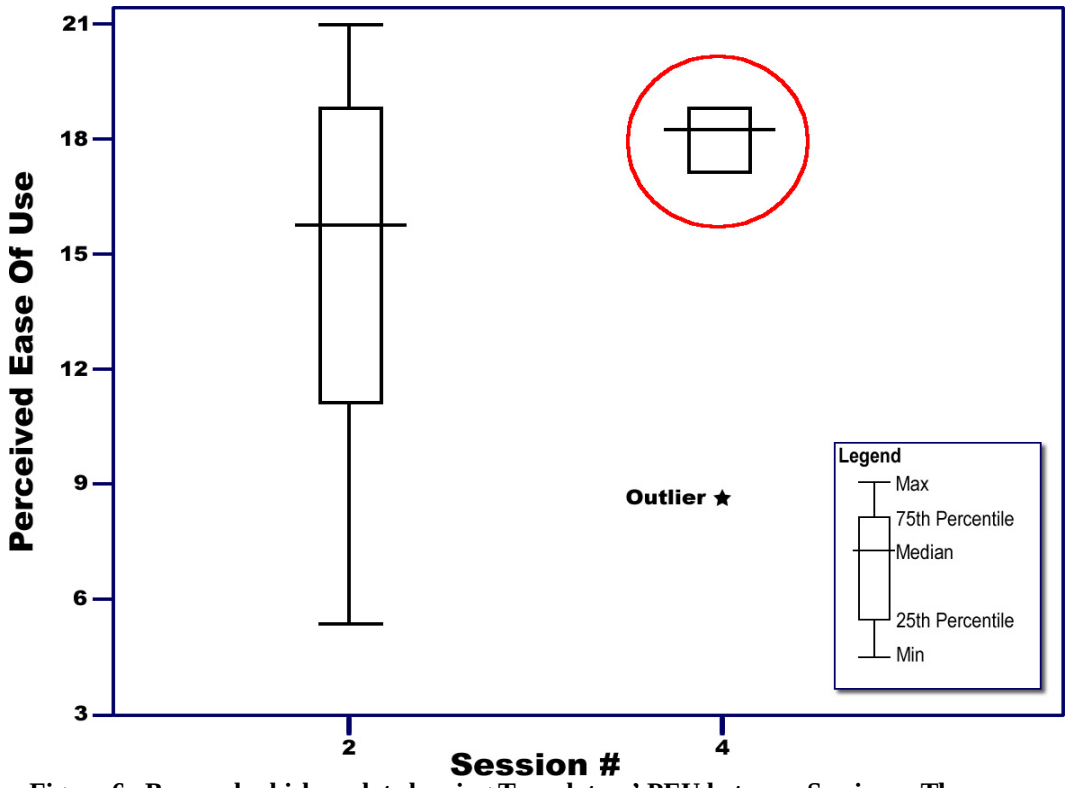


Figure 6. Box-and-whisker plot showing Translators' PEU between Sessions. The responses for the PEU questionnaire given after the 4th translation session show an increase in perceived ease of use.

In Figure 6, the data at the value "2" on the x-axis represents the 2nd overall session for all translators (the first session using the TM). There is a considerable breadth of responses over that session, where the box represents the range within the mean and the footed lines ("whiskers") represent the maxima and minima. The improvement reveals itself in the more consistent, less-variable, and higher-valued configuration of the data at the value "4" on the x-axis (the last TM session).

QC Qualitative Results

Like the translators, the QC professionals received questionnaires concerning their experiences with correcting the translations. The objective was to determine whether variables captured by these questionnaire items would result in useful measures for future experimentation.

The questions from the Quality Control Professional's questionnaires were distilled into four variables: Quality Control for Accuracy; Quality Control for Grammar, Quality Control for Style, and Quality Control for Fluency. Of these, Quality Control for Accuracy (QCA) appears most sensitive to improvement in translations received by the QC'ers, thereby indicating a positive influence by the TM system.

The QCA variable comprises these questionnaire items from the QC-specific questionnaire:

Item 1. The translated text uses appropriate and accurate vocabulary.
Item 3. The translated text includes new information not found in the source document.
Item 4. The meanings of the source document and translated text are identical.
Item 5. Overall, the translated text is accurate.

These four items in the variable address a QCers' evaluation of the accuracy of any given translation (i.e., its fidelity to the meaning of the original).

Figure 7 is a box-and-whisker chart showing the QC distribution across the corrections from the translation from all sessions (i.e., with and without the TM system). It is evident that there is a positive shift between the perceived accuracy of the unaided translations and the first round of TM-assisted ones (session 2). The QC experts saw the translations in a different order than the translators did, and thus the distribution of QC accuracy judgments in sessions 3 and 4 appear to reflect translator behavior. This fact indicates the value of reusing this measure in the planned future experiments.

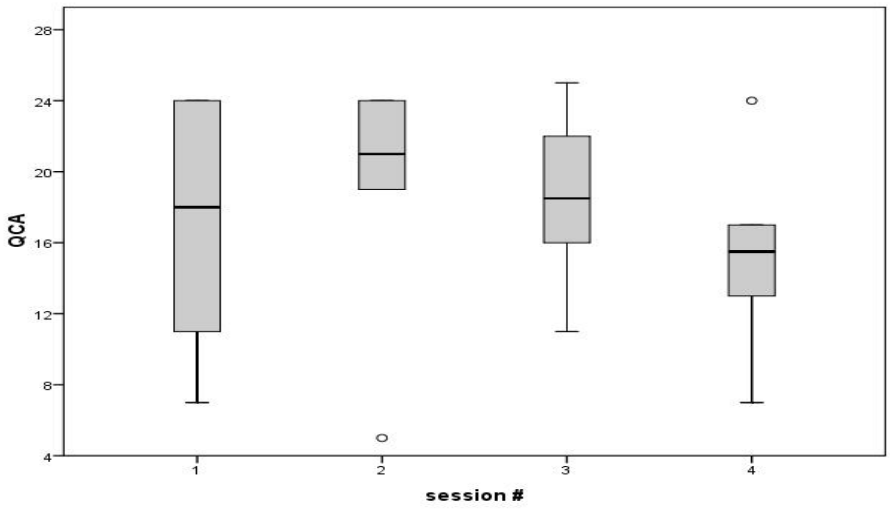


Figure 7. Box-and-whisker plot showing QCers' Accuracy (QCA): Session 1 without TM; sessions 2, 3, and 4 with TM. The trend implies an improvement in accuracy followed by a leveling, perhaps owing to the TM tool use by the Translator.

Analysis of Written Translator Comments

Several of the participants wrote valuable comments reflecting on their understanding, their approach to the tasks and their responses to the TM system, the test environment, and to the test itself.

TM work environment

Translators stressed that their normal work environment allows them to constantly consult a wide variety of sources, including Internet sites, fellow translators, and different types of dictionaries or specialized glossaries. The protocol of the experiment confined them to consulting the TM tool and a single paper dictionary. Every subject noted the difficulty of this restriction. The experimental design intended to control for external bias by the restriction, but it may have made unduly artificial the test environment. This is an experimental design issue to be addressed in subsequent experiments.

Acclimation to the TM tool

We observed that translators' reaction to the TM system improved after a few days of use. They generally agreed the tool could be useful, but stressed that the current TM banks were insufficiently populated. By the conclusion of the experiment, their reactions to the tool were optimistic, if the TM system were to be used as an adjunct to existing translator processes and tools.

TM banks (resources)

The Russian translators had more negative reactions to the TM system. As noted previously, the Russian TM banks were much smaller than the Arabic and Chinese, and not optimized for the domains / genres of the test translations. These facts may have caused the negative responses of the Russian translators.

Current Efforts

Work has continued in enhancing the experimental design necessary to determine the optimum benefit of TM to translation processes. These efforts include understanding the empirical nature of text genres. We, therefore, are expanding our collaboration and incorporating the National Institute of Standards and Technology (NIST) for the purposes of introducing new approaches to future experimentation on TM technology.

Genre Analyses

An experiment currently underway examines the potential for using empirically-derived text properties as evidence of both genre type and suitability for TM. This experiment runs a variety of automatic metrics against corpora of texts in known genres (meeting minutes, reports, and speech transcripts). These metrics, including readability, complexity, and repetition, may be able to help automatically identify the genre of a document as well as estimate its potential suitability for handling with TM. Key to this investigation will be a partnership with the Canadian National Research Council, who will aid in the development of advanced empirical determinants of TM suitability.

NIST-led Technical Workshops

In early December 2009, NIST and NVTC will bring together TM users, developers and interested researchers from U.S. industry, academia and government to discuss current TM technology issues and solutions in a technical workshop. Through panel-led discussions on issues such as: user approval of TM; the next generation of TM technology; and the best evaluation measures and metrics, the workshop will springboard new approaches to evaluation of individual TM systems, similar to the highly effective NIST Machine Translation evaluation series (Przybocki et al., 2008). A second workshop will solidify a roadmap with implications for new trends in TM design, implementation, and usage in the context of informative and relevant evaluation strategies.

Follow-on Work

The pilot study used data such as broadcast news transcripts that are not commonly associated with successful use of TM. However, NVTC translation requirements are far broader and include different genres, such as audio files, email messages, images, television broadcasts, web pages, etc. The follow-on experiments will build on the pilot assessment to obtain more robust results that may generalize across languages, and take into account additional genres of relevance to NVTC.

Language Considerations

Subsequent experiments will include Swahili, an African language with a large speaker base but a relatively low online/parallel resource inventory. Its inclusion will additionally indicate how readily TM technology can adapt to the addition of a new language.

Partitioning Memory Banks

Subsequent experiments will rely on effective means of partitioning TM banks into domain and/or genre sub-repositories, to determine the effect of TM on a translation using only the genre-relevant TM bank versus using the entire TM bank for the same translation.

Updating Memory Banks

The current and planned investigations into the usefulness of TM in the variety of translation tasks in the NVTC mission will lead to several improvements to translation processes. One such improvement relates to the concept of the TM feedback loop discussed earlier (cf. Figure 1 and the discussion). Many of the translations coming out of the translation work of NVTC will funnel into the feedback loop to update and improve the existing TM banks for the associated language pairs, domains, and genres. A future analysis, therefore, will gauge the impact of such a feedback design.

Conclusion

Translation memory (TM) has the potential to empower the translator through ready access to standard usage, exact coverage of recurrent passages, and consistency

across documents within a genre. The promise potentially extends to those genres whose documents are not as repetitive as, say, user manuals, but which can benefit in standardization and consistency. At the same time, the quality control professional may derive significant benefit from TM, as a means of providing a basis for terminological and constituent standards to enforce throughout the document – the parts translated entirely manually as well as those that are not.

The pilot study reported here, as well as the current work, establishes a framework for the next iteration of experiments that will enable rigorous designs and protocols for the evaluation of TM in the U.S. government user context.

References

Barabe, D. 2009. Perspective of the Canadian Federal Translation Bureau. Participant address to the Panel, "Translation in Government." *MT-Summit XII*. Ottawa, Canada.

Davidov, D. and Rappoport, A. 2009. Translation and Extension of Concepts Across Languages. *Proceedings of the 12th Conference of the European Association for Computational Linguistics.* Athens, April 2009.

Gow, F. 2003. *Metrics for Evaluating Machine Translation Software.* Master's Thesis, University of Ottawa. Ottawa: F. Gow.

Lagoudaki, E. 2006. *Translation Memory Systems: Enlightening Users' Perspectives*. Imperial College, London. November 2006.

Macklovitch, E., and Russell, G. 2000. "What's been forgotten in translation memory." In White, John S. (ed.)., *Envisioning Machine Translation in the Information Future: Proceedings of the 4th Conference of the Association for Machine Translation in the Americas (AMTA 2000).* Cuernavaca, Mexico.

Przybocki, M., Peterson, K., and Bronsart, S. 2008. *Official results of the NIST 2008 "Metrics for MAchine TRanslation" Challenge (MetricsMATR08)*, http://nist.gov/speech/tests/metricsmatr/2008/results/.

Van Ess-Dykema, C., Gigley, H., Lewis, S., and Bannister, E. 2008. Embedding translation at the front end of a human translation workflow: An NVTC vision. *Proceedings of the 2008 Association for Machine Translation in the Americas (AMTA-2008).* Waikiki, Hawaii.

Van Ess-Dykema, C., Perzanowski, D., Converse, S., Richardson, R., Maney, T., and White, J.S. 2009. Translation memory technology assessment. *Proceedings of MT Summit XII.* Ottawa, Canada: International Association for Machine Translation.