

# Applying Total Leveraging to TDA Shared TMs

Ilia Kaufman, KCSL Inc.

October 1, 2009

On August 11, 2009 the TAUS Data Association (TDA) released results of a pilot project summarizing the benefits of [sharing translation memories](#).

Although the reported benefits are very significant, it has been suggested that they are not completely meaningful without additional information about the following:

- Quality of translations.
- Reductions in times required to perform post-editing.
- Actual machine translations that were generated from the available data.

Since KCSL contributed to the TDA pilot and agrees with the concerns raised, additional information as it pertains to its contribution is provided here. This information has been extracted from a separate and more detailed study that was recently conducted by KCSL on the same TDA pilot data.

## Study Objectives

- To examine technology that uses “Total” rather than “Legacy” leveraging, which at the segment level exploits only single best (100% or fuzzy ) matches.
- To measure translation/localization time and production cost savings that are realized by applying Total leveraging technology to good quality translation memories (TMs).
- To determine if the time and production cost savings are achieved without reducing translation quality.

## Technology Used

KCSL’s NoBabel Enhancer was used since it was seen to be ideally suited to achieve the study objectives. NoBabel Enhancer can process small TMs with only hundreds of translation units (TUs) as well as very large TMs with millions of TUs. For a given source sentence (segment) it leverages simultaneously multiple relevant TUs within the TM. NoBabel Enhancer’s leveraging is total. It produces translations for all, or virtually all, segments and the quality of these Total leveraging translations can be scored in a similar fashion to that of scoring translations produced by Statistical Machine Translation (SMT) systems.

## Data Used

As our sample source document we selected a chapter entitled “[Questions and Answers](#)” from a document related to computer software supplied by a TDA member company. Our selected document contains 589 words. Selecting such a relatively small source document allows for a thorough analysis of the data while minimizing the cost of human translations. Multiple

human translations are required to assess the quality of both machine and human translations. (See **Time and Quality Measurements** below.)

The parallel data used in this study consisted of merged translation memories from five TDA member companies in one language pair (EN-FR), in the same domain (computer software), and of a similar content type (software strings, documentation):

- TDA Member Company A: 1,517,937 words
- TDA Member Company B: 5,564,996 words
- TDA Member Company C: 2,224,301 words
- TDA Member Company D: 2,850,334 words
- TDA Member Company E: 2,142,196 words

Since we also wanted to investigate the impact of using more data to get better translations, the TMs were grouped as follows:

- TM1 – TM from TDA Member Company A, which also contributed the document with our selected “Questions and Answers” chapter. For leveraging analysis with TM1 see the Legacy column in [Analysis of NoBabel Productivity Gain Using TM1](#).
- TM2 – TM1 plus TMs from TDA Member Companies B, C, D and E. For leveraging analysis with TM2 see the Legacy column in [Analysis of NoBabel Productivity Gain Using TM2](#).

## Our Approach

The following five scenarios were analyzed:

- **Scenario 1** – The translator in this scenario was given the “[Questions and Answers](#)” document and was asked to translate it from English into French without any additional information from us. The translator was also asked to report the time spent on performing the translation task.
- **Scenario 2** – Same as Scenario 1 but in this scenario the translator was also given [TM1](#) with all the 50% to 100% matches and asked to apply Legacy leveraging.
- **Scenario 3** – Same as Scenario 1 but in this scenario the translator was also given [TM1 Enhanced by NoBabel](#), that is, a TM with TM1 TUs as well as TUs generated by NoBabel for all required segments and asked to apply Total leveraging. Each NoBabel TU includes an identifier ALIGN! or MT! to indicate the quality of that TU. ALIGN! indicates that the translation is very good and may require only minor correction. MT! indicates that a more thorough revision of this TU may be required. In this scenario the translator’s task was post-editing.
- **Scenario 4** – Same as Scenario 1 but in this scenario the translator was also given [TM2](#) with all the 50% to 100% matches and asked to apply Legacy leveraging.

- **Scenario 5** – Same as Scenario 1 but in this scenario the translator was also given [TM2 Enhanced by NoBabel](#), that is, a TM with TM2 TUs as well as TUs generated by NoBabel for all required segments and asked to apply Total leveraging. Each NoBabel TU includes an identifier ALIGN! or MT! to indicate the quality of that TU. ALIGN! indicates that the translation is very good and may require only minor correction. MT! indicates that a more thorough revision of this TU may be required. In this scenario the translator’s task was post-editing.

In summary, the translators used no leveraging in Scenario 1, Legacy leveraging in Scenario 2 and Scenario 4, and Total leveraging in Scenario 3 and Scenario 5; the translator’s task in Scenario 3 and Scenario 5 was post-editing.

### Time and Quality Measurements

To reduce the impact of variations among the individual translators in reporting their time to perform the translations as well as the quality of their translations/post-editing we assigned each of the 5 scenarios to 3 different translators. In total 15 translators, both freelancers and LSP employees, participated in this study.

Since we had collected 15 different human translations for the same document we felt that [BLEU](#) scores would be well suited to measure the quality of the various translations. Our BLEU scores were obtained as follows:

- Each human translation was scored using the remaining 14 human translations as reference translations.
- For Scenarios 3 and 5 we also scored the NoBabel Enhancer translations using all 15 human translations as reference translations.

Both the BLEU scores and the reported times were averaged for each of the 5 scenarios.

### Time and Quality Results

	Scenario 1	Scenario 2	Scenario 3		Scenario 4	Scenario 5	
	Human Translation	Human Translation with TM1	Machine Translation NoBabel(3)	Post-Editing of NoBabel(3)	Human Translation with TM2	Machine Translation NoBabel(5)	Post-Editing of NoBabel(5)
<b>BLEU Score Human Ref.</b>	0.6328	0.9092	0.7281	0.8893	0.7597	0.7727	0.8654
<b>Average Time in Minutes</b>	122	75	N/A	68	73	N/A	61

The highest BLEU score was 0.9394, which was obtained for a translator who reported a time of 64 minutes for Scenario 5. This translation is in [Questions – réponses](#).

## Other Machine Translation Systems Scored and Reviewed

It is interesting to compare the results for NoBabel Enhancer (Scenario 5) using TDA shared data to the translations of other MT systems, such as [Google translate](#) and [Microsoft Translator](#). Both are well known in the industry, use very large parallel corpora and lend themselves to post-editing applications.

In addition to obtaining the BLEU scores for the 3 MT systems (NoBabel, Google, and Microsoft) using the 15 human translations as reference translations, we also scored each of the 3 systems using each of the other MT translations as a reference translation. This allowed for a “rough” comparison of the 3 systems.

Reference Translations ▶	15 Human Translations			
MT Systems ▼		NoBabel(5)	Google	Microsoft
NoBabel(5)	0.7727	1.0000	0.4508	0.5468
Google	0.7542	0.4508	1.0000	0.6125
Microsoft	0.8392	0.5468	0.6125	1.0000

The following points are worth noting:

- All 3 MT systems, Google, Microsoft and NoBabel, produced similar scores.
- All 3 systems scored relatively high against the human translations. The BLEU scores for all 3 MT systems were even higher than the BLEU scores for human translations in Scenario 1 in which the translators were given only the original document.
- The highest BLEU score of the 3 systems, 0.8392, was achieved by the Microsoft system. This high score is partly attributable to the fact that the content of the sample test document is closely related to Microsoft software for which Microsoft obviously has a lot of good parallel data.
- Google’s system is likely trained on even more data than Microsoft’s system, but the data used by Google isn’t specifically targeted to the subject matter of the sample test document. Its BLEU score, 0.7542, was the lowest of the 3 systems.
- NoBabel Enhancer was in the middle with its 0.7727 BLEU score. It used TDA shared data selected for computer software. This data was relevant to the sample test document; however, the TDA data in the computer software domain is currently neither as vast as that of Microsoft or Google nor as focused on Microsoft software as that of Microsoft.
- The similarity between the translations of the Microsoft and Google systems is evident from the relatively high BLEU score of 0.6125. NoBabel’s BLEU scores compared with those of Google and Microsoft were lower at 0.4508 and 0.5468 respectively. These scores reflect, in part, the fact that the algorithms used by the Google and Microsoft systems are in many ways very similar while the algorithms used by the

NoBabel system are different. Google and Microsoft use fundamentally the same SMT technology while NoBabel uses a combination of Natural Language Processing (NLP) and Statistics technologies. Also, as mentioned earlier, Google and Microsoft both use very large parallel corpora while in this study NoBabel used the TDA database that is currently smaller than the databases used by Google and Microsoft.

## Study Limitations

The conclusions that follow from this study are subject to variations resulting from several factors that had an impact on the quantities shown in the tables. The following limitations are to be noted:

- The size of the sample document was small, making it difficult to obtain accurate timings from the translators, especially when some translators rounded their times to full hour(s) rendering the measurements of differences in times difficult.
- Our instructions for the translators were lacking in clarity with regard to the use of our TMs, proofreading terminology research and project management, resulting in somewhat inconsistent reporting of the times spent on the work.
- There was no allowance for the experience level of translators and/or their familiarity with the subject matter. In our study of the 5 scenarios, we could have addressed these issues (but at a higher cost) by using 5 different sample documents of similar length and 5 different translators. With a proper assignment schedule we could have ensured that with a total of 25 translations:
  - Each translator translated all 5 documents.
  - No translator translated the same document more than once.
  - No translator used the same scenario more than once.
- We cannot be certain that the translators always did as instructed because we did not ask them to return the translations along with the corresponding TMs. This would have ensured that the translators used appropriate software and also used our additional original and enhanced TMs to perform their tasks.

## Study Conclusions

Notwithstanding the above limitations, there are important conclusions that can be drawn from our study. Some of these conclusions are already well known.

### Legacy Leveraging Compared to Human Translation

- **Using Legacy leveraging** (Scenario 2 and Scenario 4) **saves time and thus reduces translation production costs as compared to using human translations alone.**
- **When using Legacy leveraging** both the reported and the estimated savings are similar and they indicate that **there is little to gain in terms of time savings from additional**

**TMs.** The savings from the reported times for Scenario 2 and Scenario 4 are 38.5% and 40% respectively. The estimated savings for Scenario 2 and Scenario 4 (from [TM1 Analysis](#) and [TM2 Analysis](#)) are 31.07% and 31.41% respectively.

- **Using Legacy leveraging (Scenario 2 and Scenario 4) results in higher quality translations as compared to using human translations alone.**
- While the BLEU scores for both Scenario 2 and Scenario 4 are considerably higher than those of Scenario 1 (that did not make use of any TMs), the results indicate that **with Legacy leveraging using additional TMs (Scenario 4) does not contribute to better quality translations as compared to using a single highly relevant TM (Scenario 2) and may contribute to a slight degradation in translation quality/consistency.**

### **Total Leveraging Compared to Human Translation and to Legacy Leveraging**

- **Using Total leveraging followed by post-editing (Scenario 3 and Scenario 5) saves time and thus reduces translation production costs as compared to using human translations alone.**
- **Using Total leveraging followed by post-editing (Scenario 3 and Scenario 5) results in higher quality translations as compared to using human translations alone.**
- **Using Total leveraging followed by post-editing (Scenario 3 and Scenario 5) results in time and production cost savings exceeding those achieved with Legacy leveraging (Scenarios 2 and 4).**
- **Using additional TMs with Total leveraging followed by post-editing (Scenario 5) further increases time and production cost savings (as compared to Scenario 3).** This is not the case for Legacy leveraging (Scenario 4 as compared to Scenario 2). The savings from the reported times for Scenario 3 and Scenario 5 are 44.3% and 50% respectively. The estimated production cost savings for Scenario 3 and Scenario 5 (from [TM1 Analysis](#) and [TM2 Analysis](#)) are 55.35% and 55.69% respectively. Although the estimated savings with the original TM and the additional TMs are almost the same, the reported savings in Scenario 5 are in fact greater. This is explained by the better quality of the NoBabel translation in Scenario 5 that was given to translators for post-editing. The BLEU score in Scenario 3 with one TM increased from 0.7281 to 0.7727 in Scenario 5 which used the additional TMs.

### **Study Summary**

- Legacy leveraging and Total leveraging result in higher quality translations as compared to using human translations alone.
- Legacy leveraging and Total leveraging offer time and production cost savings over human translation alone.

- Total leveraging offers time and production cost savings over Legacy leveraging.
- Using additional TMs with Total leveraging increases time and production cost savings more than with Legacy leveraging.
- In this study with multiple (15) human translations available, the BLEU scores are consistent and offer an accurate measure of the quality of both human and machine translations.
- The reported results and conclusions are substantiated by human translations and post-editing work.
- The results and conclusions are in line with those reported in the TDA pilot project on the benefits of [sharing translation memories](#).

## **Acknowledgements**

The author thanks Ariane Duddey of Okapi Localization & Project Management both for assigning the translations and post-editing work and for ensuring that the work was done in a professional and timely manner. Also the author thanks Ziad Bhunnoo, Felix Liao, Adrian Mak and Robert J. van Noggeren of KCSL for their contributions to generating NoBabel Enhanced TMs, various analyses and obtaining the relevant BLEU scores.