Kokiri Translations

Pilot Project Proposal

For The Legend of Zelda: Breath of the Wild

Linka Wade, Kyle Chow, Nicole Henry, Xingyue (Silver) Zhang

Pilot Project Results

The purpose of this pilot project was to estimate the resources required to train a neural machine translation (NMT) engine to translate Zelda games from Japanese to English to demonstrate the feasibility of using post-editing of machine translation (PEMT) for future Zelda titles. A total of 8 training rounds were completed on the data, using two different NMT engines; Microsoft Translator and Systran. The first round was done in Microsoft Translator and post-edited, then the best rounds from both Systran and Microsoft Translator were also post-edited. The results of the pilot project are as follows:

	Original Objectives	Results
Efficiency:	PEMT 30% faster than human translation	PEMT 84% faster than human translation
Cost:	PEMT 30% savings over human translation	PEMT 82% savings over human translation
Quality:	PEMT quality deemed acceptable through player evaluation of factors such as story and consistency in character speech style.	PEMT quality deemed acceptable through player evaluation of factors such as story and consistency in character speech style.

Cost

Costs were calculated based on the assumptions that a human translator would translate 300 words/hour and would be paid $0.12 per word. The per-word rate assumption was based on the current market average for Japanese translations.

	Hypothetical Human Translation		PEMT ($40/hr)		Savings	Proposed Objectives
Best Round (Microsoft)	Time	289 min	Time	46.5 min	84%	30%
	Cost	$173.52	Cost	$31.00	82%	30%
Best Round (Systran)	Time	289 min	Time	37 min	87%	30%
	Cost	$173.52	Cost	$24.67	86%	30%

Efficiency

Two team members were responsible for PE – one who was familiar with the Zelda franchise and Breath of the Wild and one who was not. The PE process was flawed in that these team members were given the same segments to post-edit over the course of the training rounds. There was enough of a gap between the first training round and the best round in Microsoft Translator that the post-editors' familiarity with the material wasn't an issue, but as there was not a similar gap for the post-editing of the best Systran round, that efficiency rate is not as valid a data point. It is worth noting, however, that regardless, all efficiencies far surpassed the original objective of 30% faster than human translation.

Quality

Although the original pilot project proposal stated that the Dynamic Quality Framework (DQF) QA Metric model would be used to assign numerical values to quality standards, this has been removed from the project. Instead, segments were divided into four text types: Dialogue, Instructions/Tips, Narrative, and Item Descriptions. These text types were then rated on three criteria: Character/Style Consistency, Terminology Consistency, and Understandability. Each team member assigned a number from 1-10 for each criterion in each text type, 10 being a translation of launch-ready quality, and 1 being completely incomprehensible text. The scores were then averaged across all four team members' evaluations.

Each criterion was also assigned a benchmark score that if surpassed, would be deemed a translation of acceptable quality. Those benchmark scores are outlined below and were weighted based on importance to overall player experience.

Text Type	Dialogue	Instructions	Narrative	Item Description
Character/Style Consistency	9	8	9	8
Terminology Consistency	8	9	8	9
Understandability	8	9	8	9

The average scores based on team member human evaluation are outlined below. Unfortunately, none of the post-edited rounds were able to successfully surpass the quality benchmarks and are thus not considered acceptable. The NMT needs further training to meet the expected quality standards.

1st Round (Microsoft)	Text Type	Dialogue	Instructions	Story	Item Description
	Character/Style Consistency	3.0	6.3	3.3	5.3
	Terminology Consistency	4.0	5.7	5.7	6.7
	Understandability	4.8	8.0	6.0	7.5
Best Round (Microsoft)	Text Type	Dialogue	Instructions	Story	Item Description
	Character/Style Consistency	5.0	7.3	5.0	7.7
	Terminology Consistency	7.7	7.7	7.7	7.7
	Understandability	6.0	8.5	6.5	8.0

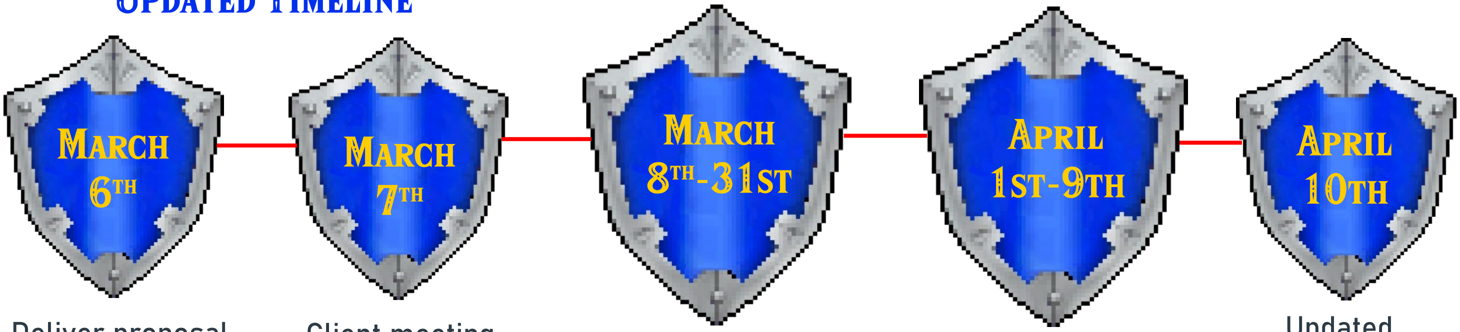
Best Round (Systran)	Text Type	Dialogue	Instructions	Story	Item Description
	Character/Style Consistency	5.3	7.7	5.3	6.7
	Terminology Consistency	6.3	6.7	6.0	6.7
	Understandability	6.8	8.0	6.5	7.8

Updated Project Objectives

The project's objectives have been revised to reflect the higher expected standards of efficiency and cost, as well as the updated quality metric model.

- Efficiency: PEMT 80% faster than human translation
- Cost: PEMT 80% savings over human translation
- Quality: PEMT quality deemed acceptable through player evaluation of factors such as story and consistency in character speech style, as seen in the benchmark scores table in the Quality section.

Updated Timeline



Deliver proposal

Client meeting

Data prep/alignment
8 rounds of training

Post-editing, human evaluation, savings, and improvement calculation

Updated proposal

Updated Project Costs

Project costs at the conclusion of the pilot project were $123 above the original anticipated cost.

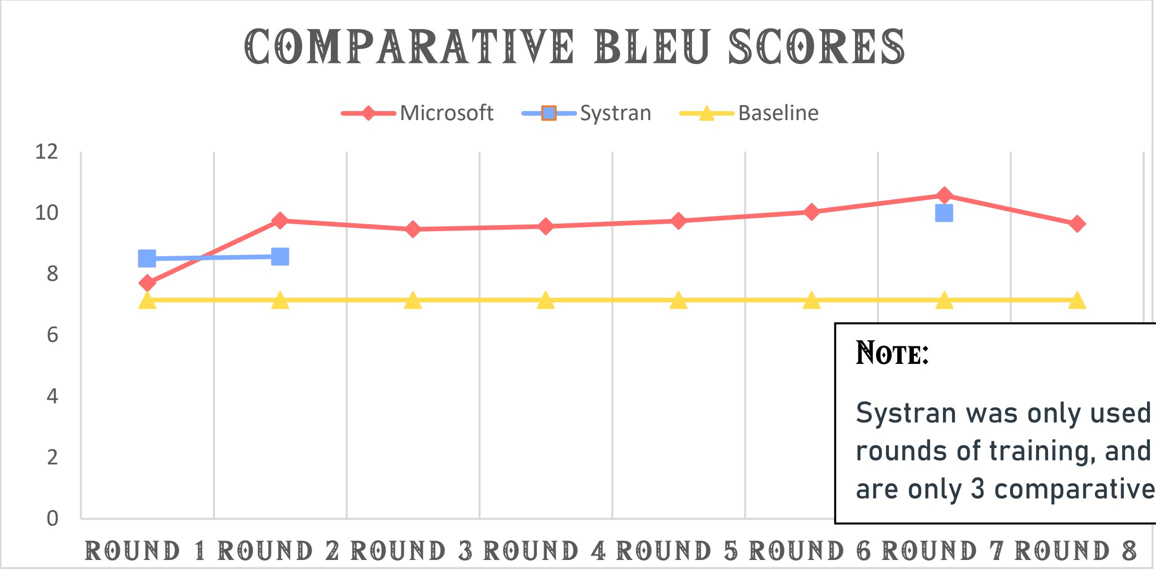
Item	Estimated Hours	Actual Hours	Rate/Hour	Estimated Cost	Actual Cost
Source/Target Text Collection	10	8.75	$40	$400	$350
Segmentation/Alignment of Documents	20	26.25	$40	$800	$1050
Glossary Creation	3	1.5	$40	$120	$60
Data Cleaning	3	3	$40	$120	$120
MT Training/Troubleshooting	8	3.3	$40	$320	$132

Post-Editing & Quality Measurement	4	8	$40	$160	$320
Subtotal	48	50.8	-	$1920	$2032
Project Management Fee			10% of total	$192	$203
TOTAL				$2112	$2235

Training Rounds

Round	Time Spent	BLEU Score	Baseline Score	Dataset	Changes
Round 1	1 hour(s) 36 minute(s)	7.7(↑0.5)	7.15	Training:9827 Tuning: 501 Testing: 501	-
(Systran Round 1)	2:39:11	8.5		Training:9827 Tuning: 501 Testing: 501	
Round 2	1 hour(s) 56 minute(s)	9.74(↑2.59)	7.15	Training:9,827 Tuning:501 Testing:501 Dictionary:298 (↑298)	Added dictionary of terms scraped from AntConc
(Systran Round2)	2:33:16	8.57		Training:9,827 Tuning:501 Testing:501 Dictionary:298 (↑298)	
Round 3	1 hour(s) 33 minute(s)	9.46 (↑2.31 /↓0.28)	7.15	Training:11,652 (↑1825) Tuning:501 Testing:501 Dictionary:298	Added Twilight Princess (Zelda Series) in training data
Round 4	1 hour(s) 35 minute(s)	9.56 (↑2.41 / ↑0.1)	7.15	Training:11,652 Tuning:679 （↑178） Testing:501 Dictionary:436 （↑138）	Added dictionary and tuning data (from Breath of the Wild)
Round 5	1 hour(s) 35 minute(s)	9.73(↑2.58/↑0.17)	7.15	Training:11,861 （↑209） Tuning:679 Testing:501 Dictionary:436	Added Final Fantasy IX (outside of Zelda Series) in training data
Round 6	1 hour(s) 29 minute(s)	10.03(↑2.88/↑0.3)	7.15	Training:12,225 (↑364) Tuning:679 Testing:501 Dictionary:436	Added another part of Final Fantasy IX in training data

Round 7	1 hour(s) 29 minute(s)	10.57(↑3.42/↑0.54)	7.15	Training:12,225 Tuning:882（↑203） Testing:501 Dictionary:437（↑1）	Added one dictionary term and more tuning data (from Breath of the Wild)
(Systran Round3)	2:35:30	9.99		Training:12,225 Tuning:882（↑203） Testing:501 Dictionary:437（↑1）	
Round 8	1 hour(s) 12 minute(s)	9.64(↑2.49/↓0.93)	7.15	Training:12,225 Tuning:882 Testing:501 Dictionary:436(↓1)	Removed a dictionary containing a single specific game term



Note:

Systran was only used for 3 rounds of training, and thus there are only 3 comparative data points.

Recommended Additional Training

1. **Expansion of Glossary**
   The clearest way to improve the NMT engine further is to expand the glossary with more terms. The addition or omission of glossary terms led to by far the biggest differences in BLEU scores between rounds (see Round 3 and Round 8), and it is especially notable that removing only a single glossary term between Rounds 7 and 8 caused a 9% drop in BLEU score. This demonstrates just how important it is in a video game context to standardize terminology, so spending further efforts towards developing a comprehensive glossary for the NMT engine would have great benefits.

2. **Expand the source of dataset**
   From Round 1 to Round 4, the new data we added were all restricted to the series of Zelda as decided in our proposal. However, we noticed we had hit a bottleneck in BLEU scores not improving, so we needed a breakthrough by adding something different. The new game text dump introduced in Round 5 brought about a slight increase in BLEU scores, but due to the time limit, we did not have the chance to do more rounds with data outside the Zelda series to investigate this trend further. Therefore, for additional training, we suggest data

collection from other games or novels with adventure themes to see if they help improve BLEU scores. We may have been tunnel-visioned in our previous rounds of training.

Recommended Translation Settings

1. **Add Termbase/Dictionary for Post-Editing and Automated Quality Check**
   Unlike MT, human post-editors cannot remember the translation of every single term at one time. Therefore, it is necessary to add dictionaries used in the training process to the CAT Tool so post-editors can always refer to it while doing the PE. In addition, QA checks could be run one more time to ensure terminology consistency of translation.

2. **Add Punctuation Check in QA**
   We noticed sometimes that the counterparts of Japanese quotation marks (「」 『』) were missing in the target text, so regular expressions could be set up to detect all the segments without English quotation marks in the target but Japanese quotation marks in the source.

3. **Add Segment Length Check in QA**
   Text will generally expand 20% to 60% when translating from Japanese into English, so it is recommended to make sure that the length of the English segments is not too long after human post-editing, which is important for the presentation of strings in video games. Therefore, we recommend a setting such as “Check for target segments which are longer by 40%” in QA checks so that post-editors can shorten any long translations accordingly.

Calculations for a Full Project

As efficiency and cost objectives were very easily met with the pilot project, quality would be the main focus of further efforts in a full project. It is difficult to give anything more than rough figures for time and cost of such a project because of the subjective nature of quality, but an educated guess can be made based on the improvement seen in quality scores. Below is a table showing the percent progress made towards outlined quality score goals from the first round to the best round in Microsoft Translator:

Progress Towards Quality Goal from First to Best Round (Microsoft)	Text Type	Dialogue	Instructions	Story	Item Description
	Character/Style Consistency	33%	60%	29%	88%
	Terminology Consistency	92%	60%	86%	43%
	Understandability	38%	50%	25%	33%

The goal would be to achieve 100% in all criteria, so the limiting factor would be the cell with the lowest amount of improvement so far, Understandability for Story (25%). Making the assumption that progress is linear, it can be extrapolated four times the effort thus far would need to be expended for this criterion to reach 100%. Following this logic, the additional costs would be roughly four times the pilot project cost, or around $8000. However, there are many additional factors to consider for what goes into the accuracy of this estimate. For example, additional effort generally has diminishing returns, so progress is not entirely linear, making $8000 an underestimate. However, on the flipside, the standardization of our processes throughout the pilot project will make any further work more efficient, making $8000 an overestimate. Thus, overall, several factors in either direction may roughly balance out, making $8000 an acceptable baseline number to work from.

Recommended Choice of Engine

Based on our results from both MT engines, we recommend using Microsoft Translator for further training. From the table showing average scores for human evaluation, it can be seen that Systran showed slightly higher scores overall in subjective text types like Dialogue and Story, and Microsoft Translator resulted in better overall quality scores for more objective text types like Instructions and Item Descriptions.

Although this may initially suggest that either engine could be used for further training, it is worth noting that the lack of a glossary in Systran limits its potential for further improvement. Glossaries are extremely important particularly for the objective text types in terms of consistency, so it is unlikely Systran will reach the quality levels stated in our updated goals, while Microsoft Translator can continue improving across all text types (as shown by the increase in quality scores from the first to best rounds).